A Machine Learning Approach to Evaluate Ransomware and Data Breach in Higher Education Institutions

by Chelsie C. Thompson

B.S. in Commerce & Business Administration, December 2016, University of Alabama
M.S. in Management Information Systems, April 2019, University of Alabama at Birmingham

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 5, 2024

Praxis directed by

Muhammad F. Islam
Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Chelsie Cara Thompson has passed the Final Examination for the degree of Doctor of Engineering as of August 15, 2023. This is the final and approved form of the Praxis.

A Machine Learning Approach to Evaluate Ransomware and Data Breach in Higher Education Institutions

Chelsie Cara Thompson

Praxis Research Committee:

Muhammad F. Islam, Professorial Lecturer in Engineering and Applied Science, Praxis Director

John Sahlin, Professorial Lecturer in Engineering and Applied Science, Committee Member

Raymond Chow, Professorial Lecturer in Engineering and Applied Science, Committee Member

© Copyright 2023 by Chelsie C. Thompson
All rights reserved

Dedication

To my loving grandmother, Carol Hill, and my parents, Dr. Karen, and Anthony Thompson – who have worked tirelessly and made countless sacrifices to support and nurture my educational pursuits from my earliest days. Your unwavering faith and steadfast support form the foundation upon which this achievement stands.

To Dr. Brittany Williams, my rock and anchor, who encouraged me to scale new heights even when the journey was fraught with challenges. Your steadfast belief in me has been a wellspring of courage and determination. Your unwavering support kept me calm amidst the storms this degree sometimes brought.

To my siblings, cousins, aunts, uncles, friends, and bonus parents - Paul, Stephanie, Jackie, and Leslie - who have each participated in this journey in their own unique ways. Your love encouragement, patience, flexibility, and inspiration have enriched this experience beyond measure. Lastly, to C.C., my loyal four-legged companion, whose comforting presence at my feet during hours of research served as a source of joy.

A special acknowledgment to Ms. Marchie. We remember you with love and gratitude, and you are dearly missed.

This praxis is lovingly dedicated to each one of you as a testament to your enduring love and unwavering support.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my Praxis Director, Professor Muhammad F. Islam. Your steadfast support, unparalleled expertise, and enduring patience significantly shaped my academic journey. I deeply appreciate your guidance and commitment to my research. My sincere appreciation extends to the faculty and staff at George Washington University. Their dedicated commitment to excellence in education and fostering a stimulating intellectual environment has been a source of inspiration throughout my studies.

I would also like to express my heartfelt thanks to my colleagues and especially my manager within Microsoft's Education Sector. Their solid support, flexibility, and belief in my abilities made it possible for me to maintain a balance between my professional commitments and the rigorous demands of this degree. Their provision of an innovative and conducive work environment, as well as their technical insights, significantly influenced the execution and outcomes of my research.

In terms of personal support, I owe a special word of thanks to my partner, family, and friends. The constant encouragement, the faith in my abilities, and the relentless optimism they displayed, particularly during periods of immense pressure, have been my pillars of strength throughout this journey. They instilled in me the resilience and determination required to strive for and achieve academic excellence.

Abstract

A Machine Learning Approach to Evaluate Ransomware and Data Breach in Higher Education Institutions

The academic sector, marked by its vast and diverse sensitive data, has emerged as a significant target for cybercriminals in the digital era. This trend is underscored by a 44% escalation in cyber-attacks on academic institutions from 2021 to 2022, culminating in a distressing weekly average of 2,297 incidents in 2022. The swift shift to digital learning, catalyzed by the COVID-19 pandemic, has amplified this concern.

In response to the escalating threat of digital attacks, this research applies machine learning methodologies to identify and quantify the primary factors associated with two prevalent types of cyber threats: ransomware and data breaches. The study consolidated demographic data and historical records of reported attacks from 2,297 U.S.-based academic institutions, leading to a comprehensive analysis. Given the disparity in the dataset—with 636 confirmed incidents contrasted against the total number of institutions—advanced data analytics and machine learning models were employed.

Four distinct machine learning models—Logistic Regression, Voting Ensemble, Random Forest, and Light GBM—were evaluated, validated, and trained. Their efficacious performance under such conditions is assessed through rigorous cross-validation and evaluation using several metrics, including weighted accuracy, precision, recall, and F1-score. Notably, the Voting Ensemble model outperforms the others, achieving an outstanding weighted accuracy score of 0.92290.

Insights from the robust Voting Ensemble model reveal key factors

distinguishing institutions with a history of cyber-attacks from those without any recorded incidents. The size of an institution emerges as the most influential factor, followed by the presence of a medical school, the number of degrees offered, research status, and academic ranking.

By employing advanced data analytics, this study delivers an in-depth analysis of characteristics that increase the risk of specific types of cyber-attacks in academic institutions, thereby laying a robust foundation for further research. Future research directions include augmenting the dataset with underrepresented classes to enhance classification accuracy and broadening the scope to include K-12 schools and similar entities in different sectors. Such expansion would provide more comprehensive insights and uncover additional key factors associated with ransomware and data breach attacks. The findings of this research significantly contribute to the existing body of knowledge and have crucial implications for the security of academic institutions in the digital age.

Contents

Dedication ..... **iv**

Acknowledgments ..... **v**

Abstract ..... **vi**

List of Figures ..... **xiii**

List of Tables ..... **xiv**

List of Symbols ..... **xv**

List of Acronyms ..... **xvi**

Glossary of Terms ..... **xvii**

Chapter 1 – Introduction 1

1.1 Background 1

1.2 Research Motivation 3

1.3 Problem Statement 5

1.4 Thesis Statement 6

1.5 Research Objectives 7

1.6 Research Questions and Hypotheses 8

1.7 Scope of Research 10

1.8 Research Limitations 11

1.9 Organization of Praxis 13

Chapter 2 – Literature Review 14

2.1 Overview 14

2.2 Overview of Cybersecurity Challenges in Higher Education 14

2.2.1 The Nature and Impact of Data Breaches in Education 15

2.2.2 Characteristics of Ransomware Attacks in the Education Sector 15

2.2.3 Understanding the Cyber Threat Landscape in Educational Institutions 16

2.3 Existing Research on Cyber-Attacks in Educational Institutions 19

2.3.1 Causes of Cyber-Attacks in Educational Institutions 21

2.3.2 Consequences of Cyber-Attacks in Educational Institutions 22

2.3.3 Cyber-Attack Varieties and Their Impact on Educational Institutions 24

2.3.4 Mitigation Strategies for Cyber Threats in Educational Institutions 28

2.4 Predictive Methods for Preventing Cyber-Attacks 29

2.4.1 General Predictive Techniques 30

2.4.2 Advanced Predictive Models: Feature Engineering 31

2.5 Appraisal of Current Predictive Techniques: Strengths and Limitations 32

2.6 Leveraging Predictive Methods for Cybersecurity in Educational Institutions 33

2.7 Risk Measures for Predicting Cyber-Attacks 34

2.8 Application of Machine Learning Models for Cyber Threat Prediction 35

2.8.1 Overview of Machine Learning Models: Data-Driven Classification 36

2.8.2 Machine Learning Models: Case Studies in Educational Institutions 38

2.9 Conclusion 40

Chapter 3 – Methodology 42

3.1 Introduction 42

3.2 Data Preparation 44

3.2.1 Data Collection and Sources 44

3.2.2 Data Integration and Cleaning 46

3.3 Data Preprocessing 50

3.3.1 Data Normalization 50

3.3.2 Data Transformation 51

3.3.3 Handling Imbalanced Data 52

3.4 Exploratory Data Analysis (EDA) 54

3.4.1 Refinements and Profiling of Data 54

3.4.2 Correlation Analysis 55

3.5 Feature Engineering 55

3.5.1 Feature Creation 56

3.5.2 Feature Selections 56

3.6 Model Development and Validation 56

3.6.1 Machine Learning Model Selection and Training 57

3.6.1.1 LightGBM 58

3.6.1.2 Logistic Regression 59

3.6.1.3 Random Forests 61

3.6.1.4 Voting Ensemble 63

3.6.2 Model Optimization: Hyperparameter Tuning 64

3.7 Model Evaluation and Metrics 67

3.7.1 Cross-Validation 67

3.7.2 Metric Comparisons 68

3.7.3 Confusion Matrix 70

3.7.4 Feature Importance Analysis 70

Chapter 4 – Results 73

4.1 Introduction 73

4.2 Data Preprocessing 73

4.3 Feature Engineering 75

4.3.1 Feature Creation 75

4.3.2 Feature Selection 76

4.4 Exploratory Data Analysis 77

4.4.1 Correlation Analysis and Variable Selection 77

4.4.2 Factors Contributing to Data Breach and Ransomware Risk 81

4.5 Model Performance Evaluation Comparison Results 86

4.5.1 AUC-ROC Comparative Analysis 87

4.5.2 F1 Score Comparative Analysis 89

4.5.3 Precisions Score & Average Score Comparative Analysis 90

4.5.4 Recall Score Comparative Analysis 91

4.5.5 Matthew's Correlation Coefficient Comparative Analysis 92

4.5.6 Accuracy and Weighted Accuracy Comparative Analysis 93

4.5.7 Comprehensive Evaluation of Model Performance 94

4.5.8 Model Validation and Performance Evaluation 96

4.6 Model Development Results 98

4.6.1 Model Feature Importance 98

4.6.2 Hyperparameter Tuning Results 102

4.7 Hypothesis and Research Question Results 104

4.8 Conclusion 108

Chapter 5 – Discussion and Conclusions 110

5.1 Discussion 110

5.2 Conclusions 111

5.3 Limitations 112

5.4 Contributions to Body of Knowledge 112

5.5 Recommendations for Future Research 113

References 116

Appendix A. Data Variables ..... **Error! Bookmark not defined.**

Appendix B . Criteria U.S. News Colleges Ranking Criteria 129

Appendix C. Carnegie Classification Ranking System 130

Appendix D. Inclusion and Exclusion Criteria for Higher Education Institutions 131

Appendix E. Machine Learning Results 132

Appendix F. Multivariate Logistic Regression Odds Ratio Results 133

Appendix G. Univariate Logistic Regression Odds Ratio Results 134

List of Figures

Figure 1 *Industry Sector Targeted by Nation Station Actors* 18

Figure 2 *Enterprise Malware Encounters Reported in the Last 30 Days* 26

Figure 3 *Distribution of Cyber-attacks across Industries in 2022* 28

Figure 4 *Overview of the Research Methodology Framework* 44

Figure 5 *Suggested Data Transforms in Power Query* 47

Figure 6 *Data merge based on School Name* 49

Figure 7 *Illustration of Synthetic Minority Over-Sampling Technique (SMOTE)* 54

Figure 8 *Initial Heat Map Displaying Variable Correlation* 78

Figure 9 *Correlation Matrix Post-Variable Removal* 79

Figure 10 *Portion of Cyber-attacks by Region* 82

Figure 11 *Data Breaches by Research Status* 82

Figure 12 *Portion of Cyber-attack by school Rank* 83

Figure 13 *Portion of Digital Attack by Urbanization* 84

Figure 14 *School Size* 84

Figure 15 *School Affiliation* 85

Figure 16 *Odds Ratios of Cyber-attack Factors* 86

Figure 18 *Initial Feature Importance Map* 100

Figure 19 *Feature Importance Map After Feature Reduction* 101

List of Tables

Table 1 *Data Dictionary Describing the Variables in the Research Data Sources* 45

Table 2 *Functions of Various Data Pre-processing Tools Used in the Study* 46

Table 3 *Summary of Machine Learning Models Utilized in the Study* 57

Table 4 *List of Hyperparameters for the Evaluated Machine Learning Models* 66

Table 5 *Definitions and Formulas of Machine Learning Performance Metrics* 68

Table 6 *Confusion Matrix* 70

Table 7 *AUC-ROC Comparative Analysis* 88

Table 8 *F1 Score Comparative Results* 89

Table 9 *Precision Score Comparative Results* 90

Table 10 *Recall Score Comparative Analysis* 91

Table 11 *Comparative Analysis of Matthews Correlation Coefficient* 92

Table 13 *Analysis of Model Accuracy and Weighted Accuracy* 94

Table 14 *Voting Ensemble: Overall vs. Test Performance* 97

Table 15 *Hyperparameters of the Final Ensemble Model* 103

List of Symbols

1. α Significance level
2. β Coefficient of the independent variable
3. γ Intercept of the linear regression model
4. Δ Change in a variable
5. σ Population standard deviation
6. θ Parameter of the population distribution
7. n Sample size
8. x Independent variable
9. y Dependent variable
10. p Probability
11. H_0 Null hypothesis
12. μ Mean
13. $\sum$ Sigma

List of Acronyms

1. AUC Significance level Area Under the Curve
2. AI Artificial Intelligence
3. EDA Exploratory Data Analysis
4. HE Higher Education
5. FN False Negative
6. FP False Positive
7. kNN K-Nearest Neighbors
8. LR Logistic Regression
9. ROC Receiver Operating Characteristic
10. SMOTE Synthetic Minority Oversampling Technique
11. TN True Negative
12. TP True Positive
13. U.S. United States

Glossary of Terms

Azure: Microsoft's comprehensive suite of cloud computing services, providing a wide array of solutions for computing power, analytics, storage, and networking.

Accuracy: A metric that quantifies the proportion of correct predictions made by the model within the test results.

Area Under the Receiver Operating Characteristic Curve: Probability curve, and the AUC quantifies the model's ability to distinguish between classes.

Boosting: A machine learning technique that involves training a sequence of tree-based models, with each subsequent model learning from the mistakes of its predecessor to improve overall performance

Ensemble Model: A technique in which multiple diverse models, referred to as base learners, are employed to collectively predict an outcome

Exploratory Data Analysis: An approach to analyzing data sets to summarize their main characteristics, often with visual methods.

F Distribution: A probability distribution that is primarily used in ANOVA (Analysis of Variance) for testing the equality of several variances.

F1 Score: A measure calculated as the weighted average of Precision and Recall, offering a balanced assessment of these two metrics.

Feature Importance: A machine learning techniques that assigns a score to input features based on how useful they are in predicting a target variable.

Function: A relation that maps each element of a set to one or more elements of another set.

Hypothesis Testing: A statistical method that uses experimental data to accept or reject a hypothesis.

Mean ("μ"): A measure of central tendency representing the average of a set of values.

Microsoft Fabric: A comprehensive analytics platform that integrates data movement, data engineering, real-time analytics, and data science for seamless data management and interpretation.

Precision: A measure of the "correctness" of a model classification, calculated as the ratio of True Positive to the sum of True Positive and False Positive.

Probability: The likelihood of an event occurring, expressed as a number between 0 and 1.

Regression Analysis: A statistical process for estimating the relationships among variables.

Scikit-Learn: A machine learning library for Python that provides efficient tools for data mining and analysis.

Sigma (summation): A mathematical operator that indicates the sum of a set of values.

Sigma (sample standard deviation): A measure of the variation or dispersion of a set of values.

Chapter 1 – Introduction

1.1 Background

The landscape of cyber threats has significantly evolved in recent years, with educational institutions emerging as prime targets for cyber-attacks (Ncube & Garrison, 2010). These institutions have been confronted with an alarming surge in high-impact cyber incidents and given their roles as custodians of sensitive personal and academic data, there is a significant need for a more comprehensive and robust cybersecurity measures (Ncube & Garrison, 2010).

A study by Check Point reported a 44% rise in cyber-attacks against educational institutions in 2022 compared to 2021, averaging 2,297 weekly attacks worldwide (Check Point, 2022). This surge positioned the education sector as the most targeted industry in 2021 and in 2022, with ransomware and data breaches being the primary threats (Iwendi et al., 2022).

Data breaches within educational institutions have raised significant concerns. Since 2005, schools and colleges in the United States have suffered 2,691 data breaches, impacting at least 32 million records (Cook, 2023). In 2021 alone, 771 institutions were affected, compromising 2.6 million records. The Illuminate Education data breach alone impacted at least 605 institutions (Cook, 2023). This trend continued into 2022, with 96 breaches exposing almost 1.4 million records, and as of 2023, there have already been 11 breaches affecting over 3,500 records (Cook, 2023).

These statistics underscore the escalating threat of cyber-attacks in the education sector, which can be attributed to increased reliance on technology, the sensitivity of managed data, and inadequate security measures. Recent research has emphasized the

importance of the need for the development of cost-effective cybersecurity management systems tailored for academic institutions (Alharbi, 2022). Additionally, research reports provide valuable insights into vulnerabilities and areas for improvement in organizational security practices. The need for robust cybersecurity risk mitigation techniques in higher education organizations is also emphasized (Ulven & Wangen, 2021).

This praxis seeks to leverage machine learning models to evaluate ransomware and data breaches in higher education institutions, with the objective of developing a predictive model to assist in proactive threat preparation and mitigation.

Despite the increasing frequency of cyber-attacks in the education industry, research on this sector remains limited. Historically, the education industry was not considered a high-risk sector for cyber-attacks, resulting in minimal research attention (Ulven & Wangen, 2021). However, the recent recognition of cyber-attacks as a significant concern in the education sector has prompted a sense of urgency, highlighting the existing research gap (Ulven & Wangen, 2021).

The limited understanding of the nature and scope of cyber-attacks in the higher education sector is another factor contributing to the need for more research. Until recently, there has been minimal research attention on cyber-attacks as a substantial threat in this sector. However, the growing awareness of cyber-attack risks and the increasing prominence of privacy concerns in digital environments have shifted the research focus (Ulven & Wangen, 2021). As the education sector recognizes cyber-attacks as a significant threat, research efforts are expected to increase accordingly.

The inherent complexity of the education sector further compounds the need for more research on cyber-attacks. The multifaceted nature of this sector presents challenges

when studying cyber-attacks within this context. With diverse groups and unique perspectives, conducting comprehensive research that accurately represents the entire sector is challenging (Ulven & Wangen, 2021). Moreover, the decentralized higher education sector makes it difficult to establish a unified research agenda applicable across all institutions (Ulven & Wangen, 2021).

Additionally, the lack of data collection and sharing among educational institutions poses a significant barrier to research. Many institutions need help collecting, analyzing, and sharing data on cybersecurity incidents, limiting their understanding and response to threats (Ulven & Wangen, 2021). This lack of standardized data collection and reporting methods hinders research and comparison across institutions.

In conclusion, the scarcity of research on cyber-attacks in the education industry can be attributed to several factors. The recent recognition of cyber-attacks as a significant concern within the sector, coupled with the complexity of the education system and the challenges associated with data collection and sharing, contribute to the limited research in this area. As awareness of the risks posed by cyber-attacks continues to grow, research efforts in this field are expected to increase (Ulven & Wangen, 2021).

1.2 Research Motivation

The education sector, especially higher education institutions, has increasingly become a prime target for cyber-attacks. Both internal and external actors exploit vulnerabilities to gain unauthorized access to sensitive information, emphasizing the critical need for comprehensive security measures. Despite the escalating frequency and severity of these attacks, research specifically focused on ransomware and data breach attacks within higher education remain noticeably sparse. This research gap underscores

the urgent need for an in-depth understanding of the unique challenges these institutions face in combating cyber threats.

Cyber-attacks on educational establishments pose risks not only to student data but also to the overall integrity and reputation of the institutions. Given the continuously evolving nature of threats, universities bear the responsibility to invest continually in the necessary infrastructure and expertise to address genuine cybersecurity concerns (Campbell, 2017).

Most existing research on cyber-attacks primarily targets other industries, leaving educational institutions without a nuanced understanding of the nature, scope, and impact of cyber-attacks within their unique context. A comprehensive database incorporating all academic ransomware and data breach attacks is essential for practical analysis and responses to these incidents.

This research aims to bridge this knowledge gap through an exhaustive analysis and synthesis of previously collected data, shedding light on the challenges besetting educational institutions. The study proposes to develop a data-driven classification model using historical data to discern specific characteristics that increase susceptibility to data breaches and ransomware attacks within these institutions. Machine learning methods will be employed to enhance the model's efficacy, with cross-validation conducted to derive reliable estimates of the model's performance on unseen data. Evaluation metrics, including accuracy, precision, recall, and F-Measure, will be used to assess the effectiveness of the classification model in identifying influential factors. The evaluation of the factors' coefficients will determine the influence that various factors have on an institution's likelihood of experiencing an attack.

The findings of this research could significantly enhance our understanding of cyber threats within the higher education sector, offering invaluable insights for schools and researchers working to comprehend and mitigate attacks within the educational sphere. By identifying specific characteristics that increase susceptibility to data breaches and ransomware attacks, educational institutions can formulate targeted strategies to strengthen their security measures and safeguard sensitive information. More broadly, this research seeks to fill the existing research gap and contribute to the knowledge base on cyber-attacks in the education sector by offering a practical method to evaluate attacks within the educational institutions.

1.3 Problem Statement

The education sector's sparse research and insufficient data collection obstruct efficacious mitigation of cyber-attacks. Consequently, there is a compelling need for a predictive model that could quantify the determinants of vulnerability and bolster cybersecurity within higher education institutions (Ulven & Wangen, 2021).

The scarcity of comprehensive research leaves educational institutions with a limited understanding of the nature and scope of the cyber-attacks they face. This knowledge gap hinders their ability to devise effective strategies to prevent and respond to these threats. Furthermore, the absence of a centralized and comprehensive database of academic ransomware and data breach attacks restricts the ability to systematically analyze and learn from past incidents.

This multifaceted problem is compounded by the need for standardized data collection and sharing amongst educational institutions. Without a comprehensive

understanding of the factors and patterns associated with these attacks, institutions are ill-equipped to anticipate and respond to potential threats.

To address this problem, this research proposes the development of a data-driven classification model. This model will leverage historical data to identify key characteristics of educational institutions that have been subjected to specific cyber-attacks. The model will utilize confirmed reported cyber-attack data from various universities in the United States. By analyzing patterns and trends in the data, the classification model will systematically examine the factors contributing to the likelihood of experiencing a cyber-attack.

The knowledge gleaned from this model is anticipated to equip educational institutions with invaluable data, thus empowering them to craft proactive approaches and judiciously distribute resources to alleviate cybersecurity vulnerabilities. Fundamentally, this investigation endeavours to reconcile the prevailing research void by creating a predictive model that quantifies vulnerability determinants, consequently augmenting cybersecurity within higher education institutions.

1.4 Thesis Statement

A data-driven classification model utilizing machine learning on historical cyber-attack data will identify factors to enhance cybersecurity in higher education institutions

This research focuses on the analysis of historical data related to ransomware and data breach incidents within the context of higher education. The primary objective is to create a robust, data-driven classification model that can proficiently identify and assess the factors associated with cybersecurity threats.

The insights derived from this model are intended to equip higher education institutions and researchers with the necessary tools to proactively address and mitigate cybersecurity risks. This proactive approach is projected to foster a safer and more secure educational environment, thereby enhancing the overall integrity of these institutions.

By contributing to the advancement of cybersecurity practices within the education sector, this research aims not only to enhance the security of individual institutions but also contribute to broader societal security. In essence, this study seeks to leverage advanced data analytics and machine learning algorithms to bolster cybersecurity measures within higher education institutions

1.5 Research Objectives

The primary objective of this research is to conduct an exhaustive evaluation of the factors associated with ransomware and data breach incidents within higher education institutions. This will be accomplished by employing advanced data analytics and applying machine learning models. The study will analyze historical data on cyber-attacks to discern patterns and correlations, with the aim of developing a robust data-driven classification model.

The specific research objectives are as follows:

- To identify factors associated with ransomware and data-breach attacks in higher education institutions, such as the size of the institution, whether the institution is a medical school, the number of degrees offered, research status, and academic ranking.
- To conduct a statistical analysis to ascertain the significance and impact of the identified factors.

- To develop and assess the performance of machine learning models, including Logistic Regression, Voting Ensemble, Random Forest, and Light GBM, using the identified factors.
- To create a unified dataset using publicly available data, data analytics, and machine learning algorithms to identify patterns and relationships between different variables related to cyber-attacks.
- To evaluate the temporal trends of ransomware and data breach incidents in higher education institutions, identifying any patterns or cycles in the frequency of such incidents over time.
- To compare the susceptibility of different types of higher education institutions (such as public vs private, medical schools vs other institutions, large vs small institutions, etc.) to ransomware and data breaches.

The anticipated outcome of this research is the development of a comprehensive and reliable data-driven classification model that accurately identifies the factors contributing to ransomware and data breach attacks. This research aims to provide valuable solutions to address this escalating problem, thereby enhancing the overall cybersecurity posture of higher education institutions.

1.6 Research Questions and Hypotheses

To focus this study on attacks within U.S. colleges and universities, the following research questions and hypotheses have been formulated:

Research Question 1: What are the key factors that significantly influence the classification of a U.S higher education institution as being at risk of a data breach or ransomware attack?

Research Question 2: Can a data-driven classification model be systematically developed and refined to identify and quantify the factors contributing to the susceptibility of a higher education institution to a data breach or ransomware attack?

Research Question 3: Can Machine Learning algorithms applied to publicly available data expose identifiable patterns and trends in cyber-attacks on educational institutions, with respect to the types, frequency of attacks, attack targets, and potential correlations between attack characteristics and institutional attributes?

Research Question 4: How do specific characteristics of educational institutions alter the risk of a data breach or ransomware attack in terms of direction and magnitude?

Hypothesis 1: There is a significant relationship between particular school characteristics and an increased risk of a data breach or ransomware attack in a higher education institution.

Hypothesis 2: The data-driven machine learning model developed will achieve a minimum accuracy of 85% in identifying higher education institutions with elevated factors for experiencing specific types of cyber-attacks.

Hypothesis 3: The Logistic Regression model will exhibit superior performance and accuracy compared to other machine learning techniques, namely Light GBM, Random Forest, and Ensemble, in identifying factors.

These research questions and hypotheses will guide the investigation and analysis conducted in this study. They aim to provide valuable insights into the relationships between specific school characteristics, factors, and the effectiveness of the data-driven classification model in identifying factors associated with cyber-attacks in U.S. colleges and universities. The outcomes of this study will contribute to the existing body of knowledge and offer practical recommendations to enhance cybersecurity preparedness in higher education institutions.

1.7 Scope of Research

This research focuses on exploring the factors contributing to ransomware and data breach attacks in higher education institutions. The study will utilize historical data on cyber-attacks in the education sector to develop a data-driven classification model. The specific objectives of this study are:

1. **Data Collection and Preprocessing:** Gather, preprocess, and integrate data from heterogeneous sources to construct a comprehensive dataset suitable for subsequent analyses.
2. **Model Development:** The research will design, train, validate, and assess a machine learning model that identifies and analyzes patterns, trends, and characteristics of ransomware and data breach attacks in higher education institutions, leveraging data from U.S. based academic institutions.

3. **Data Classification:** The research will identify critical factors that significantly influence the probability of a higher education institution experiencing a ransomware or data breach attack. The influence of these factors will be quantified, and their significance tested.
4. **Data Analysis:** The study will employ advanced analytics to extract insights from the historical data on cyber-attacks, focusing on identifying recurring patterns, emerging trends, and distinct characteristics associated with ransomware and data breach attacks. The data encompasses a wide range of institutional attributes, including the size of the institution, whether the institution is a medical school, the number of degrees offered, research status, and academic ranking.
5. **Model Evaluation**: The study will evaluate the performance of several machine learning models, including Logistic Regression, Light GBM, Random Forest, and Voting Ensemble, in identifying factors within higher education institutions. The models will be evaluated based on various metrics, including weighted accuracy, precision, recall, and F1-score.

The goal of this research is to conduct a data-driven analysis of cyber-attacks in the education sector, with the aim of understanding key factors and devising effective mitigation strategies. Although the immediate scope is limited to U.S. colleges and universities, the research methodology and insights have broader applicability across the education sector, as well as other sectors where similar data on cyber breaches and institutional characteristics are collected. By achieving these goals, this research will significantly contribute to enhancing cybersecurity within higher education institutions.

1.8 Research Limitations

This research, which explores the use of machine learning to evaluate factors associated with cyber-attacks in educational institutions, is subject to certain limitations:

- **Data Availability:** The study relies on historical data on ransomware and data breach attacks in educational institutions. The availability of comprehensive and reliable data within the education sector may be limited. This limitation could potentially introduce biases and impact the validity of the results. To mitigate this, the study will use the most reliable and comprehensive data sources available.
- **Model Validity:** The accuracy and reliability of the data-driven model developed in this research are contingent upon the quality of the data used and the assumptions made during the modeling process. To ensure the model's validity, the study will use rigorous data cleaning processes and make conservative assumptions where necessary.
- **Time Frame:** The study is based on historical data, which may not accurately reflect the current cybersecurity landscape due to the rapidly evolving nature of cyber threats. While the study acknowledges this limitation, the historical data still provides valuable insights into patterns and trends.
- **Methodological Limitations:** The use of statistical analysis techniques and machine learning models introduces potential methodological limitations. These are influenced by the chosen methods, models, and assumptions made during the analysis process. The study will clearly outline these methods and assumptions to provide transparency.
- **Class Imbalance**: The dataset might present a class imbalance issue, with fewer instances of institutions that have experienced a cyber-attack compared to those

that have not. Class imbalance can bias the machine learning models towards the majority class and affect their performance. Different strategies such as assigning class weights to increase the importance of the minority class, or using performance metrics designed for imbalanced data, will be employed to mitigate this issue.

- **Scope:** The research focuses on U.S. universities. Further studies would be required to investigate the model's applicability to other educational settings or organizations outside the U.S. While this limits the generalizability of the results, the study still provides valuable insights into the U.S. context.

1.9 Organization of Praxis

The praxis is organized into five interconnected chapters. Chapter 1, the Introduction, sets the stage for the research by providing an overview of the background, objectives, and research questions, and establishing the significance and scope of the research. Chapter 2, the Literature Review, synthesizes relevant information on logistic regression, ransomware, and data breach attacks in the education sector, providing a comprehensive overview of the current state of knowledge. Chapter 3, the Methodology, outlines the research design and approach, detailing the data preparation techniques, machine learning algorithms, and performance metrics used in the study. Chapter 4, the Results and Discussion, presents and evaluates the results of the data analysis, highlighting key findings and their implications in the context of the research questions and objectives. The final chapter, Chapter 5, the Conclusion, discusses the results in the broader context of the research field, summarizes the study's contributions to knowledge, discusses the limitations of the study, and provides recommendations for future research.

This structure ensures a logical progression of information, guiding the reader from the introduction to the conclusion, and offering insights into future research directions and practical applications.

Chapter 2 – Literature Review

2.1 Overview

This literature review provides a comprehensive analysis of existing research on cyber-attacks in educational institutions, with a particular emphasis on the methodologies employed to predict and mitigate such attacks. The objective is to inform the development of a data-driven classification model that can enhance cyber resilience in educational institutions. This review draws upon a wide range of sources, including academic journals, scholarly publications, and peer-reviewed articles, to guide this innovative research approach.

The literature review is structured into several sections, each focusing on a distinct aspect of the topic. The introduction outlines the purpose and scope of the study. The subsequent sections delve into various facets of the research on cyber-attacks in educational institutions, including the methods used to predict attacks, the application of these methods in different organizational contexts, the use of quantitative risk measures, and the employment of statistical machine learning models for outcome prediction.

2.2 Overview of Cybersecurity Challenges in Higher Education

Educational institutions represent attractive targets for cybercriminals due to the extensive amount of personally identifiable information (PII) they maintain. This data includes sensitive details such as student social security numbers, birth dates, and financial information (Beaudin, 2015; Liluashvili, 2021). However, these institutions

often lack the necessary resources and capabilities to adequately prepare for and respond to cyber-attacks (Rahman et al., 2020; Schneider, 2013). This section explores the primary obstacles that must be addressed to effectively tackle cybersecurity challenges within educational institutions.

2.2.1 The Nature and Impact of Data Breaches in Education

The U.S. records a higher number of data breaches than other countries, with the education sector being particularly vulnerable due to the sensitive information it handles (O'Driscoll, 2023). Educational institutions, ranging from schools to universities, collect, store, and utilize a wide array of data, including medical, financial, personal, and educational information pertaining to students, teachers, and other faculty members.

The potential benefits of accessing and exploiting this PII makes educational institutions enticing targets for threat actors. These cybercriminals can leverage and monetize the stolen information while causing significant disruption, aligning with their typical modus operandi. Understanding the specific nature and impact of data breaches in the education sector is crucial for developing effective cybersecurity strategies

2.2.2 Characteristics of Ransomware Attacks in the Education Sector

Ransomware, a type of malware that encrypts files on a device, rendering the files and the systems that depend on them inoperable, poses a significant threat to educational institutions (Cybersecurity and Infrastructure Security Agency, n.d.). From 2018 to mid-May 2022, approximately 954 separate educational institutions in the United States were potentially impacted by ransomware attacks, with the estimated cost of these attacks reaching around $3.56 billion (Bischoff, 2022).

According to a report by Verizon (2022), one of the primary reasons for the vulnerability of school districts is the lack of advanced defense mechanisms and resources compared to financial institutions or larger businesses. Educational institutions often operate on outdated IT systems that are more prone to cyber intrusions. Additionally, the limited time available for staff training increases the susceptibility of employees to phishing emails. The situation is further complicated by the fact that hundreds of students are using these computer systems (Rahman, 2020).

The attractiveness of higher education institutions as targets for cybercriminals has been on the rise; however, the pressure from cyberattacks and data security issues is intensifying for schools at all levels (Cook, 2023). This is a significant concern as schools, when attacked, often have limited data recovery options, making them more likely to give in to ransom demands. Furthermore, the need for schools to remain operational and the political pressure to quickly resolve the issue exacerbates the problem (Moore, 2023).

2.2.3 Understanding the Cyber Threat Landscape in Educational Institutions

Understanding the cybersecurity landscape and the sectors frequently targeted by cybercriminals is a crucial step in assessing the risks that educational institutions face. Over the past year, cybercriminals have successfully infiltrated and caused disruptions in educational institutions of various sizes and reputations (Cybersecurity & Infrastructure Security Agency, 2021). A recent study revealed that over 400 cyberattacks were launched against public schools in the U.S., with notable institutions like Howard University having to temporarily shut down due to a ransomware attack (Powell, 2021).

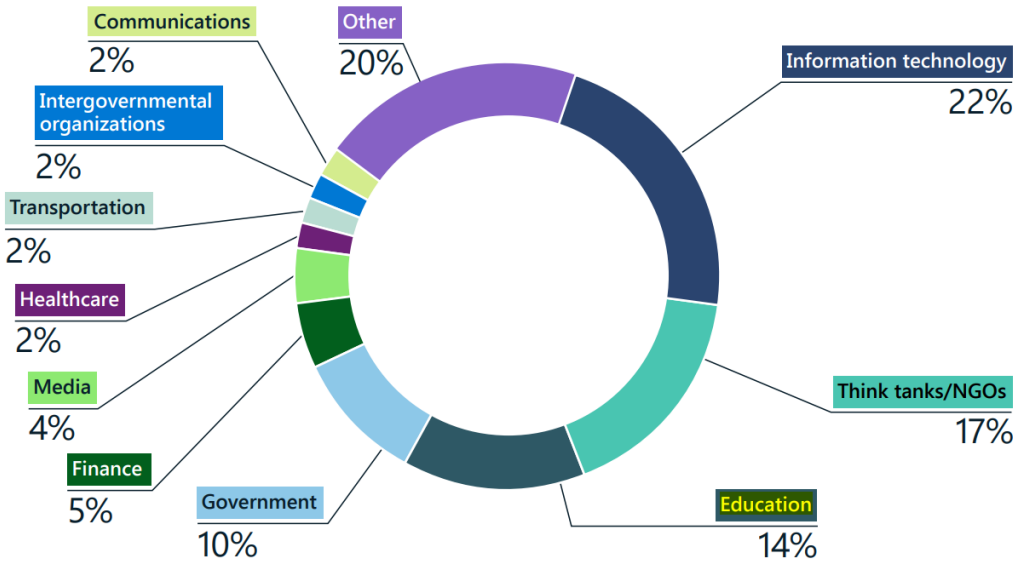
A report by Tenable, a company specializing in cyber exposure, disclosed that there were 1,825 data breach incidents between November 2020 and October 2021, leading to the exposure of more than 40 billion records globally (2021 Threat Landscape Retrospective, 2022). The report indicated that nearly 13% of these breaches were associated with the education sector, affecting students, educators, and parents due to class cancellations and inaccessible learning platforms (Tenable, 2022a). Moreover, a study by the Center for Strategic and International Studies (2022) underscored the increasing vulnerability of educational institutions to cyber threats, highlighting the need for robust cybersecurity measures to protect sensitive data and ensure uninterrupted learning.

In 2018, the U.S. Department of Education strengthened its requirements for data breaches in colleges and universities. These institutions must report any violation regardless of the number of records lost, but only if they are Title IV institutions, which accept federal funding through the federal student aid program. This covers most schools, with the exception of a small minority of religious institutions (Cook, 2023).

Recent research has revealed that certain industry sectors are frequently targeted by nation-state actors. These actors, motivated by a range of objectives, seek to illicitly access sensitive information, and destabilize critical infrastructure. As shown in Figure 1, educational institutions, accounting for 14% of attacks, must acknowledge their potential exposure to cyber threats. These institutions hold valuable intellectual property, research data, and personal information, making them appealing targets for nation-state actors seeking to gain a strategic advantage or disrupt operations (Beaudin, 2015; Liluashvili, 2021; Schneider, 2013).

Figure 1

Industry Sector Targeted by Nation Station Actors



Note. The chart offers a visual representation of the sectors most vulnerable to cyber-attacks from nation-state actors. It emphasizes the susceptibility of several industries, including, but not limited to, education, government agencies, financial institutions, healthcare providers, energy companies, and technology firms (Microsoft, 2022).

The susceptibility of higher education institutions to cyber-attacks is particularly alarming. The current study, therefore, seeks to apply machine learning algorithms to evaluate the factors associated with ransomware and data breaches in these institutions. The goal is to develop and implement more robust and effective protective measures, thereby reducing the risk and potential impact of such cyber-attacks.

By understanding the targeted sectors and the motivations driving these attacks, educational institutions can enhance their cybersecurity strategies and develop proactive measures to safeguard their networks, systems, and sensitive data. Investing in robust security measures and staying updated on the evolving threat landscape is crucial to mitigate the risks posed by nation-state actors.

2.3 Existing Research on Cyber-Attacks in Educational Institutions

The frequency and sophistication of cyber-attacks on higher education institutions have been on the rise. This is largely due to the institutions' possession of sensitive personal and financial information and their increasing reliance on technology for teaching and research, making them particularly vulnerable to cyber-attacks. The repercussions of these attacks can be severe, encompassing financial losses, disruptions to educational processes, and damage to an institution's reputation (Adam, 2021). A survey by Adam (2021) found that 74% of ransomware attacks on higher education institutions resulted in successful data encryption, a higher rate than other sectors surveyed.

Moreover, 60% of institutions reported experiencing such attacks in 2021, an increase from 44% in 2020 (Adam, 2021). The education sector had the highest rate of data encryption at 73%, compared to 65% in other sectors, and had the longest recovery time, with 7% taking at least three months to recover, almost double the average time for different sectors at 4% (Adam, 2021). High-profile ransomware attacks have been reported at universities such as the University of California – San Francisco (2019) and the University of Alberta (2018), underscoring the disruptive nature of these attacks and the difficulty in determining the most common or disruptive types of cyber-attacks for

universities. The impact of an attack depends on factors such as the university's size, location, security measures, and system vulnerabilities (Maranga & Nelson, 2020).

In the K-12 education sector, Kamaludeen et al. (2020) provides an overview of the challenges and risks facing schools and educational institutions regarding cybersecurity in their work, "A Framework for Cyber Protection (FCP) in K-12 Education Sector." The authors note that the increased reliance on technology in the K-12 sector has exposed schools to various cyber threats, including malware, ransomware, phishing attacks, and data breaches. To address these risks, the authors propose the Framework for Cyber Protection (FCP), a risk management approach designed to meet each educational institution's specific needs and resources. The FCP includes best practices and guidelines for implementing adequate cybersecurity measures to protect against cyber threats.

Cyber-attacks targeting educational institutions have become a growing concern in recent years (Bongiovanni, 2019; Cheng & Wang, 2022; Maranga & Nelson, 2019; Rahman et al., 2020). These attacks pose significant threats to the security and integrity of sensitive data, disrupt academic operations, and can have severe consequences for students, faculty, and staff (Maranga & Nelson, 2019). Understanding the causes and motivations behind these attacks is crucial for developing effective strategies to prevent and mitigate such incidents (Maranga & Nelson, 2019; Shen et al., 2017). This subsection provides a comprehensive literature review of existing studies that have explored the causes and motivations behind cyber-attacks in educational institutions.

2.3.1 Causes of Cyber-Attacks in Educational Institutions

Several factors contribute to the prevalence of cyber-attacks in educational institutions. These include financial gain, data theft, ideological motivations, and the desire to disrupt operations.

Financial Gain: A significant motivation behind cyber-attacks on educational institutions is financial gain. Cybercriminals target these institutions in search of valuable information, including personal data, financial records, and research data (Beaudin, 2015; Liu et al., 2016; Maranga & Nelson, 2019; Shen et al., 2017). This data can be sold illegally or used for identity theft and fraud. Financially motivated attackers exploit vulnerabilities in the institution's systems and networks to gain unauthorized access, compromise security, and extract valuable information for monetary purposes.

Data Theft: Educational institutions store a vast amount of valuable data, including research findings, intellectual property, student records, and financial information (Liu et al., 2016; Maranga & Nelson, 2019; Shen et al., 2017). Data theft is a common objective of cyber-attacks on educational institutions. Malicious actors seek to exploit vulnerabilities in the institution's systems to gain unauthorized access and exfiltrate sensitive data. This stolen data can be used for various purposes, such as gaining a competitive advantage, committing identity theft, or launching targeted attacks on individuals associated with the institution.

Ideological Motivations: Some cyber-attacks on educational institutions are driven by ideological motivations (Maranga & Nelson, 2019; Shen et al., 2017). Hacktivist groups or individuals with specific ideological agendas may target educational institutions to promote their beliefs or protest against certain policies or practices. These

attacks aim to disrupt operations, deface websites, leak sensitive information, or spread propaganda. Educational institutions are seen as symbolic targets due to their societal influence and role in disseminating knowledge.

Disruption of Operations: Another motive behind cyber-attacks on educational institutions is the desire to disrupt their operations (Liluashvili, 2021; Maranga & Nelson, 2019). Attackers may target the institution's network infrastructure, databases, or critical systems to cause downtime, disrupt classes, or hinder administrative processes. Disruptions can be motivated by personal grudges, revenge, or simply the desire to cause chaos and inconvenience. Such attacks can significantly impact the institution's reputation, financial stability, and learning environment.

2.3.2 Consequences of Cyber-Attacks in Educational Institutions

The repercussions of cyber-attacks on educational institutions are multifaceted and extend beyond the immediate operational disruptions. This subsection delves into the diverse impacts of such attacks, illuminating the potential harm inflicted on institutions, students, faculty, and staff. Comprehending these ramifications is pivotal for educational institutions to prioritize cybersecurity measures and allocate resources judiciously.

Financial Losses: Cyber-attacks precipitate substantial financial losses for educational institutions (Liluashvili, 2021; Maranga & Nelson, 2019; Shen et al., 2017; Smith et al., 2019). These attacks engender direct costs, encompassing expenses related to incident response, remediation, and legal proceedings. Indirect costs may emerge from reputational damage, decreased enrolment rates, loss of research funding, and potential lawsuits from effected individuals. The financial burden associated with recovering from a cyber-attack can significantly impact an institution's long-term financial stability.

Data Breaches: A common outcome of cyber-attacks is data breaches, leading to the compromise of sensitive information held by educational institutions (Beaudin, 2015; Javed et al., 2017; Liluashvili, 2021; Liu et al., 2016; Maranga & Nelson, 2019; Shen et al., 2017). These breaches can expose personal and financial data of students, faculty, staff, and alumni, leading to potential identity theft, fraudulent activities, and privacy violations. The ensuing loss of trust and confidence in the institution's ability to safeguard personal information can have enduring consequences, potentially eroding the institution's reputation.

Reputational Damage: Educational institutions' reputations, once compromised by cyber-attacks, can suffer long-lasting damage. News of a successful attack and data breach can propagate rapidly, tarnishing the institution's image and eroding public trust (Beaudin, 2015). The ripple effects of reputational damage can be far-reaching, impacting student recruitment, institutional partnerships, research collaborations, and alumni relations. Rebuilding a tarnished reputation can be a daunting and protracted process.

Operational Disruptions: Cyber-attacks can cause significant disruptions to the routine operations of educational institutions, leading to service outages, system downtime, and delays in critical processes (Shen et al., 2017). Inaccessibility of learning management systems, online platforms, and administrative databases can impede teaching, learning, and administrative activities. Such disruptions can engender frustration among students, faculty, and staff, adversely effecting the overall learning environment and academic calendar.

Potential Harm to Individuals: Beyond the financial and operational consequences, cyber-attacks can inflict harm on individuals associated with educational

institutions (Beaudin, 2015). Compromised personal information can lead to identity theft, financial fraud, and other forms of exploitation targeting students and staff. Moreover, the psychological impact of being a victim of a cyber-attack, which can lead to anxiety, stress, and a loss of trust in the institution's ability to protect its community, should not be underestimated.

Recognizing the consequences of cyber-attacks in educational institutions underscores the urgent need for robust cybersecurity measures and proactive risk management. By acknowledging the potential harm to financial stability, data security, reputation, service continuity, and individuals' well-being, educational institutions should judiciously allocate resources and implement strategies to effectively prevent, detect, and respond to cyber threats.

2.3.3 Cyber-Attack Varieties and Their Impact on Educational Institutions

Educational institutions face a diverse array of cyber-attacks, each with its unique characteristics and potential impacts. This section will delve into the various types of cyber-attacks, including phishing attacks, ransomware attacks, distributed denial-of-service (DDoS) attacks, insider threats, and social engineering attacks.

Phishing Attacks: Phishing attacks are a prevalent form of cyber-attacks in educational institutions (Liluashvili, 2021). These attacks involve the sending of deceptive communications, often via email, instant messaging, or text messages, that masquerade as legitimate entities or individuals. The objective is to deceive recipients into divulging sensitive information, such as login credentials, financial details, or personal data. The sheer volume of potential victims and the valuable information they possess make educational institutions prime targets for phishing attacks.

Ransomware Attacks: Ransomware attacks have emerged as a significant threat to educational institutions (CISA, n.d.). In these attacks, malicious actors deploy malware that encrypts the institution's data, rendering it inaccessible. The attackers then demand a ransom in exchange for the decryption key. These attacks can disrupt academic operations, compromise research data, and result in financial losses if institutions opt to pay the ransom or incur significant costs to recover the encrypted data.

Distributed Denial-of-Service (DDoS) Attacks: DDoS attacks aim to overwhelm an institution's network infrastructure or online services by inundating them with excessive requests or traffic (Liluashvili, 2021). These attacks can disrupt services, rendering systems inaccessible to legitimate users. The potential to disrupt online learning platforms, registration systems, or websites makes educational institutions attractive targets for DDoS attacks.

Insider Threats: Insider threats refer to attacks initiated by individuals within the educational institution who have authorized access to the network or systems (CISA, n.d.). These individuals may exploit their privileged access to steal sensitive data, sabotage systems, or engage in unauthorized activities. Insider threats can be accidental or intentional, and they pose a significant challenge for educational institutions due to their difficulty to detect and mitigate.

Social Engineering Attacks: Social engineering attacks manipulate individuals into divulging sensitive information or performing actions that compromise security (Liluashvili, 2021). These attacks exploit human psychology and trust, and they can target staff, faculty, and students. Educational institutions must educate their community about social engineering techniques to minimize the risk of successful attacks.

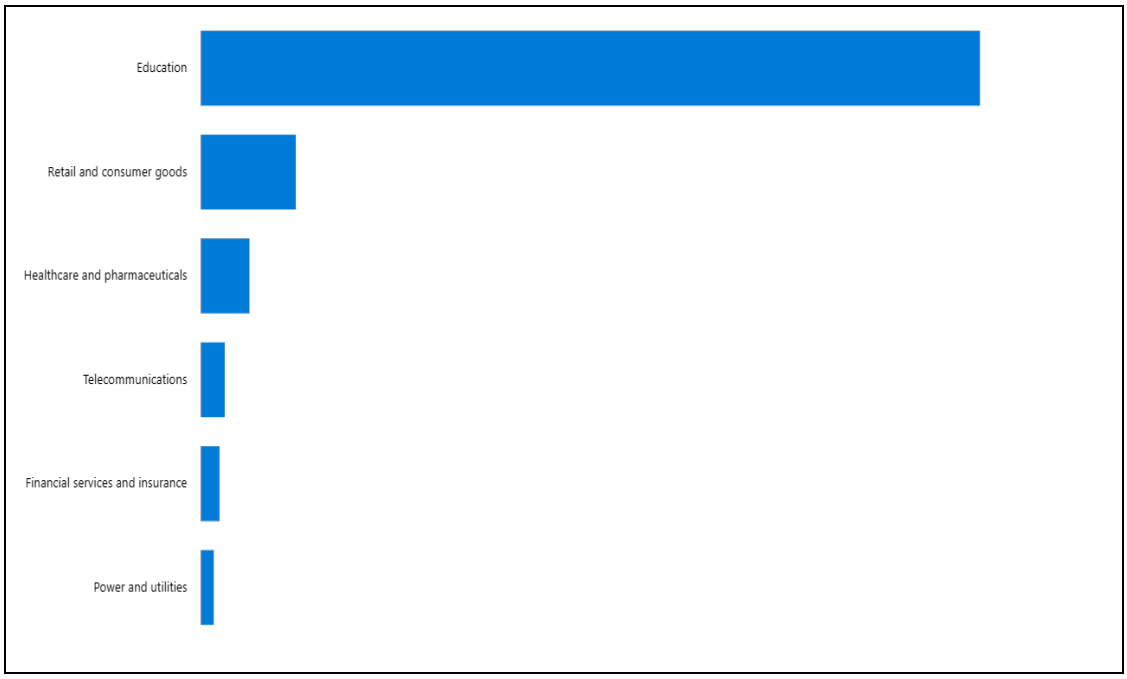
Understanding the diverse types of cyber-attacks targeting educational institutions is essential for implementing a comprehensive defense strategy. By recognizing the specific techniques and attack vectors used by malicious actors, institutions can develop targeted countermeasures, educate their community about potential threats, and enhance their overall cybersecurity posture.

The consequences of cyber-attacks on educational institutions can be far-reaching, affecting their operations, data security, and overall reputation. This section explores the consequences and impact of cyber-attacks on the education sector and examines the comparative vulnerability of the education sector to that of other industries.

Several studies and reports have emphasized the increasing threat landscape for educational institutions, with significant implications for cybersecurity (Beaudin, 2015; Cheng & Wang, 2022; Cook, 2023; Liu et al., 2016; Maranga & Nelson, 2019). A chart provided by Microsoft, which depicts the reported enterprise malware encounters in the last 30 days, highlights the education industry as one of the most affected sectors.

Figure 2

Enterprise Malware Encounters Reported in the Last 30 Days



Note. This chart underscores the urgency to address cybersecurity challenges within the education sector. Educational institutions harbor a wealth of valuable data, including personal information, research findings, and intellectual property. Consequently, they serve as attractive targets for cybercriminals seeking financial gain, data theft, or disruption of operations (Microsoft, 2023).

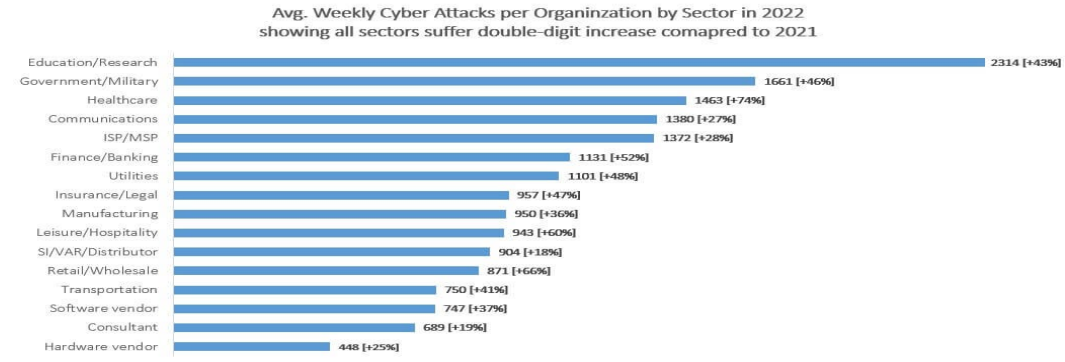
Figure 2 highlights the importance of addressing cybersecurity challenges in the education industry. Educational institutions possess a wealth of valuable data, including personal information, research findings, and intellectual property, making them attractive targets for cybercriminals seeking financial gain, data theft, or disruption of operations.

By understanding the disproportionate impact on the education industry (Figure 3) and the vulnerabilities faced by educational institutions, stakeholders can allocate resources effectively and implement robust cybersecurity measures (Cook, 2023; Liluashvili, 2021; Schneider, 2013). This knowledge underscores the need for proactive

strategies, staff training, network defences, and incident response plans tailored to the unique challenges faced by educational institutions.

Figure 3

Distribution of Cyber-attacks across Industries in 2022



Source: Check Point, 2022

Note. Adapted from Check Point Research Reports a 38% Increase in 2022 Global Cyberattacks (Check Point Software Technologies Ltd., 2023).

2.3.4 Mitigation Strategies for Cyber Threats in Educational Institutions

The escalating complexity of cyber threats necessitates the development of comprehensive mitigation strategies within educational institutions (Nauman et al., 2021). These strategies aim to minimize the opportunities for cybercrime, increase the difficulty of committing such crimes, and enhance the risks of detection and punishment for cybercriminals.

A key mitigation strategy involves the implementation of a robust risk management framework. This framework should be capable of identifying and addressing emerging risks, which may have long-term implications on the institution's academic performance and continuity (Ulven & Wangen, 2021). For instance, the adoption of new collaboration tools in response to the COVID-19 pandemic has

amplified information and cybersecurity risks. Therefore, educational institutions must adopt a broad spectrum of thinking methods that allow for practical risk decisions and provide a solid foundation for academic institutions to function and enforce strategies during and after the COVID-19 pandemic (Ulven & Wangen, 2021).

Another mitigation strategy is the application of the Routine Activity Theory (Anjum, 2020). This theory can be applied to reduce the opportunities for cybercrime, making it more challenging to commit and increasing the risks of detection and punishment associated with committing cybercrime.

Moreover, the advent of Industry 4.0, characterized by the fusion of the physical and digital worlds, presents new opportunities for sustainable manufacturing and cybersecurity (Junqueira et al., 2022). This development can contribute to the mitigation of cyber threats in educational institutions by enhancing the security of manufacturing processes and the protection of sensitive data.

In conclusion, mitigating cyber threats in educational institutions requires a multifaceted approach that combines robust risk management, the application of theoretical frameworks, and the adoption of advanced technologies. Given the evolving threat landscape and the unique challenges faced by educational institutions, these strategies should be tailored to each institution's specific needs and resources.

2.4 Predictive Methods for Preventing Cyber-Attacks

Predictive methodologies, increasingly employed across various sectors, play a pivotal role in anticipating and mitigating cyber-attacks. This section provides an overview of some commonly utilized methods and techniques, focusing on their applicability and adaptation within the context of educational institutions.

2.4.1 General Predictive Techniques

When adapting predictive methods from other sectors, it is crucial to consider the unique characteristics and requirements of educational institutions. This subsection discusses the challenges and opportunities associated with applying these methods within educational settings.

Data Availability and Quality: Educational institutions often face challenges regarding data availability and quality. Unlike organizations with extensive historical data, educational institutions may have limited datasets for predictive analysis (Beaudin, 2015; Bongiovanni, 2019; Liu et al., 2016). Furthermore, issues such as incomplete or inconsistent records can impact the accuracy and reliability of predictive models. Therefore, it is essential to tailor data collection, cleaning, and preprocessing techniques to the specific contexts of educational institutions.

Privacy and Ethical Considerations: Educational institutions handle sensitive data pertaining to students and staff, necessitating careful consideration of privacy and ethical issues when implementing predictive methods (Beaudin, 2015; Cheng & Wang, 2022; Liluashvili, 2021; Maranga & Nelson, 2019). Implementing data anonymization, encryption, and access controls are crucial to protect privacy while leveraging predictive techniques. Institutions should adhere to relevant data protection regulations and ensure the ethical use of data in their predictive modeling efforts.

Domain-Specific Knowledge: Educational institutions have distinct operational aspects, such as academic processes, research activities, and collaboration platforms (Beaudin, 2015; Liluashvili, 2021). Predictive methods should incorporate domain-specific knowledge and a contextual understanding of educational settings. This includes

considering the unique risks, vulnerabilities, and threat landscapes specific to educational institutions.

By adapting and tailoring predictive methods to the specific context of educational institutions, researchers and practitioners can develop effective strategies to predict and prevent cyber-attacks. Addressing the challenges and leveraging the opportunities presented by these methods are crucial to enhancing the cybersecurity posture of educational institutions.

2.4.2 Advanced Predictive Models: Feature Engineering

Feature engineering plays a pivotal role in the development of data-driven classification models, particularly within the context of cybersecurity in educational institutions. This process aims to reduce the dimensionality of the feature space through experimental techniques that identify the most influential features (Qiao et al., 2018). By selecting relevant features and eliminating redundant or irrelevant ones, the model's development and training time can be significantly reduced. Furthermore, feature engineering can enhance the model's accuracy and generalizability by mitigating overfitting and improving the model's capacity to capture the underlying patterns in the data (Ding et al., 2018).

For instance, Kim and Kang (2017) employed feature selection techniques in their study to identify the most relevant features for predicting the risk of heart disease. Their findings indicated that the best-performing model could predict the risk of heart disease with improved accuracy. Similarly, Ouhmida, Raihani, Cherradi, and Terrada (2021) used feature selection techniques to identify the most pertinent features for predicting the

risk of Parkinson's disease. Their study found that the best-performing model could predict the risk of Parkinson's disease with high accuracy.

These examples underscore the importance of feature engineering in machine learning, particularly in the context of data preprocessing and exploratory data analysis for developing accurate and generalizable models. Feature engineering can help reduce the dimensionality of the feature space, improve the model's accuracy and generalizability, and reduce the model's development and training time. Given the increasing availability of data and the growing demand for accurate and efficient predictive models, the development of effective feature engineering techniques is becoming increasingly important, particularly in the realm of cybersecurity within educational institutions.

2.5 Appraisal of Current Predictive Techniques: Strengths and Limitations

While advanced predictive methods in cybersecurity offer significant potential, they are not without limitations. A primary challenge lies in the complexity of the computational models used in these methods. For instance, artificial intelligence (AI) and machine learning (ML) models, which are frequently employed in predictive cybersecurity, can be intricate and challenging to interpret due to their complex nature (Tang et al., 2018). This lack of transparency can engender a degree of skepticism among cybersecurity professionals, making it difficult for them to fully trust the predictions generated by these models.

Another area of concern is the quality and availability of data. Predictive models heavily rely on extensive datasets to train and test their algorithms. However, the quality of these datasets can vary significantly, and in some instances, pertinent data may not be

readily available. This can result in inaccurate predictions and a lack of generalizability in the models, thereby undermining their effectiveness (Gupta et al., 2021).

Moreover, the dynamic nature of cyber threats presents a significant challenge to current predictive methods. Cyber threats are in a constant state of evolution, with new types of attacks emerging on a regular basis. This dynamism makes it difficult for predictive models to remain up-to-date and accurately predict future threats, thereby limiting their predictive accuracy (Gupta et al., 2020).

Finally, the implementation of predictive methods in real-world settings can pose considerable challenges. These methods often necessitate substantial computational resources and technical expertise, which may not be readily available in all organizations. Furthermore, resistance to change within organizations can make it challenging to adopt new predictive methods, thereby hindering the effective implementation of these techniques (Rahman et al., 2022).

In conclusion, while current predictive methods in cybersecurity offer significant potential, they are not without limitations. Addressing these challenges requires a concerted effort to enhance the interpretability of complex models, improve data quality and availability, adapt to the dynamic nature of cyber threats, and facilitate the effective implementation of these methods in real-world settings.

2.6 Leveraging Predictive Methods for Cybersecurity in Educational Institutions

Despite the limitations, predictive methods hold significant potential to bolster cybersecurity in educational institutions. These methods can be tailored to the specific needs and resources of educational institutions, thereby facilitating a more targeted and effective approach to cybersecurity.

For instance, AI and ML models can be trained on data specific to educational institutions, such as network traffic data, to predict potential cyber threats. These models can identify patterns and anomalies in data that may signify a cyber-attack, thereby enhancing the institution's ability to anticipate and mitigate threats (Tang et al., 2018).

Furthermore, predictive methods can be seamlessly integrated into the existing cybersecurity infrastructure of educational institutions. This integration can augment the institution's capabilities to detect and respond to cyber threats in a timely and effective manner (Neupane et al., 2022).

However, the successful implementation of predictive methods in educational institutions requires overcoming several challenges. These include ensuring the quality and availability of data, staying abreast of the evolving threat landscape, and overcoming resistance to change within the institution (Kulalaieva & Leu, 2019).

In conclusion, while current predictive cybersecurity methods present certain limitations, they offer substantial potential for enhancing cybersecurity in educational institutions. With appropriate adaptation and implementation strategies, these methods can serve as invaluable tools for predicting and preventing cyber threats in the education sector, thereby safeguarding the institution's data and operations.

2.7 Risk Measures for Predicting Cyber-Attacks

Predicting cyber-attacks in organizations has gained increasing attention recently (Cheng & Wang, 2022). Several studies have explored different methods to improve the accuracy of prediction models and understand the factors involved.

One study by Sarabi, Naghizadeh, Liu, and Liu (2016) titled "Risky Business: Fine-grained Data Breach Prediction Using Business Profiles" analyzed both network and

non-cyber incidents using publicly available business information. The study achieved 90% accuracy in identifying victim organizations and highlighted the significance of this work for insurance providers in assessing potential losses. However, the study also acknowledged limitations, such as the need for separate models for overall risk prediction and the impact of underreporting and reporting bias on the data.

Another study by Fox and Holt (2020) focused on juvenile computer hacking behavior. They utilized multivariate logistic regression and Latent Class Analysis (LCA) to predict such behavior and identified significant criminological factors. The study revealed three distinct subtypes of juvenile hackers with different factors. While LCA showed advantages over other statistical classification techniques, limitations such as inferring class membership in the population and overfitting the dataset were acknowledged.

Quantifying the risk of a successful cyber-attack is crucial in ensuring the security of educational institutions. Parker and Fakar (2011) emphasized this in their article "Modeling Estimated Risk for Cyber-attacks: Merging Public Health and Cyber Security." The authors highlighted the need for empirical evidence and well-defined metrics to measure risk. However, they also acknowledged the challenges of implementing risk assessment models in practice due to the difficulty of quantifying cyber-attack risk. Despite these challenges, Parker, and Fakar (2011) emphasized the importance of using quantitative risk measures to estimate the risk of cyber-attacks in educational institutions.

2.8 Application of Machine Learning Models for Cyber Threat Prediction

Machine learning models, especially logistic regression, have demonstrated remarkable potential in predicting real-world outcomes. This was substantiated in a study by Yadav, Singh, and Kashtriya (2023), titled "Diagnosis of Breast Cancer using Machine Learning Techniques – A Survey." The researchers utilized logistic regression to train a model on a dataset of breast cancer biopsy results, achieving an impressive accuracy rate of 98.1%. This model surpassed other machine learning techniques, such as decision trees and support vector machines. Furthermore, through cross-validation, Yadav et al. (2023) showcased the robustness of the logistic regression model, thereby highlighting its potential applicability across diverse fields.

Similar results were observed in the study "An Optimal Model of Financial Distress Prediction: A Comparative Study between Neural Networks and Logistic Regression" by Zizi et al. (2021). The authors employed logistic regression to predict companies' financial distress, training a model on a dataset of economic indicators. Through cross-validation, Zizi et al. (2021) demonstrated that the logistic regression model excelled in accuracy, precision, and recall, and could accurately identify companies at risk of financial distress.

These studies underscore the versatility and efficacy of logistic regression across different domains, reinforcing the potential of statistical machine-learning models for predicting real-world outcomes. Future research in this area can leverage these findings to extend the applications of logistic regression to other domains.

2.8.1 Overview of Machine Learning Models: Data-Driven Classification

Data-driven classification is a type of machine learning model. Machine learning models include many algorithms for classification, regression, clustering, etc. Data-

driven classification models are a subset of these and refer specifically to models that categorize data into predefined classes or categories. Classification learning is a valuable approach in supervised machine learning with immense potential for addressing cyber-attacks within the education industry. By leveraging data-driven models, educational institutions can effectively identify and predict factors associated with these attacks, enabling proactive measures for prevention. This section explores the application of classification learning in the context of the education industry and highlights relevant studies conducted in similar domains.

The education industry has experienced increasing cyber-attacks in recent years, emphasizing the need for data-driven models to identify and mitigate potential threats (Cook, 2023). Researchers have recognized the applicability of classification learning in various domains, including education, due to its ability to categorize new instances based on trained algorithms (Pham & Afify, 2005). This approach has been successfully utilized in telecommunications, marketing, and scientific analysis, demonstrating its potential for addressing cybersecurity challenges in the education sector.

In the specific context of the education industry, developing data-driven models to predict and prevent cyber-attacks is crucial to safeguard sensitive data, maintain operational continuity, and ensure the integrity of educational systems. Researchers actively explore supervised learning for various purposes within the education sector, which can be extended to address security-related challenges. Using classification algorithms, academic institutions can effectively categorize and identify potential cyber threats, enabling proactive measures to mitigate risks.

Additionally, selecting relevant features is crucial in building robust models for predicting cyber-attacks in the education industry. Wang, Zhang, and Wang (2018) demonstrate the significance of feature selection in semiconductor wafer fabrication, where identifying critical factors associated with cycle times enhances prediction accuracy. Similarly, within the education industry, selecting informative features can aid in developing accurate prediction models, allowing institutions to allocate resources effectively, prioritize security measures, and implement proactive strategies.

When applying classification learning to the education industry, addressing the challenges of imbalanced cybersecurity data is essential. Uddin et al. (2019) investigate using supervised machine learning algorithms for medical disease predictions, highlighting the need for comprehensive evaluation metrics beyond accuracy due to imbalanced datasets. In the education industry, cyber-attacks may be infrequent compared to other sectors, such as precision, recall, F-measure, and balanced accuracy, becoming crucial for assessing classification models' performance (Uddin et al., 2019).

In summary, the education industry can significantly benefit from data-driven classification learning to address cyber-attacks effectively. By leveraging supervised learning algorithms, conducting feature selection, and utilizing appropriate evaluation metrics, educational institutions can develop robust models that accurately categorize and predict potential threats, enhancing cybersecurity measures in the education sector.

2.8.2 Machine Learning Models: Case Studies in Educational Institutions

Machine learning has been used in various fields, including education and cybersecurity. In education, machine learning has been used to predict student performance and identify at-risk students (Albreiki et al., 2021). Predictive models can

help educators identify at-risk students early and provide them with the necessary support to succeed. In cybersecurity, machine learning has been used to detect and prevent cyber-attacks (Khan & Khan, 2019; Chen et al., 2020). Predictive models can help cybersecurity professionals see cyber-attacks early and prevent them from causing damage. Azure Machine Learning Studio is a popular tool used in several studies to build predictive models (Alharbi & Prince, 2020).

The use of Azure Machine Learning Studio in predicting student performance can be an example of predicting school cyber-attacks. Like predicting student performance, predicting cyber-attacks in schools requires collecting data on variables such as school size, location, other factors, and previous cyber-attacks (Chen et al., 2020; Khan & Khan, 2019). Predictive models can help school administrators identify the cyber-attack risk in their schools and respond appropriately to prevent them. Azure Machine Learning Studio can be used to build predictive models using classification algorithms and evaluate the performance of these models using metrics such as accuracy, precision, recall, and F1 score.

One example of a study using machine learning to predict cyber-attacks is the study by Chen et al. (2020). The authors used machine learning algorithms to analyze network traffic data and detect real-time cyber-attacks. The study found that their proposed model achieved a high detection rate with a low false positive rate (Chen et al., 2020).

The existing research on using Azure Machine Learning Studio to predict student performance in higher education can serve as a starting point for predicting cyber-attacks in schools. Using machine learning for predictive analytics in education and

cybersecurity highlights the potential of data-driven approaches to solve complex problems. The proposed study on predicting cyber-attacks in schools using Azure Machine Learning Studio builds on the previous literature and has the potential to contribute to cybersecurity in schools. With the increasing prevalence of cyber-attacks in schools, developing effective predictive models using machine learning is becoming increasingly important.

2.9 Conclusion

The literature reviewed in this report sheds light on the growing threat of cyber-attacks in educational institutions. These attacks are becoming more frequent and sophisticated, posing significant risks to higher education institutions. Educational institutions are particularly vulnerable to cyber-attacks due to their possession of sensitive information and reliance on technology, which can result in financial losses, disruptions, and damage to reputation. The comprehensive body of literature in this section supports the advancement of knowledge in this field.

The content within this section demonstrates that more strategic work needs to be done to mitigate cyber-attacks within educational institutions. While basic measures like user training and endpoint usage are essential, adopting machine learning approaches can provide schools with the necessary data to transform how they utilize cybersecurity information and better prepare for cyber-attacks. The information in this literature review identifies transferable strategies for school characteristics analysis, feature engineering, and algorithm development. As detailed in various information sources, these methods anchor this research and serve as the foundation for this novel approach to improving cybersecurity within educational institutions through strategic data modeling.

The literature has explored various predictive methods, including logistic regression and machine learning, to predict cyber-attacks and identify organizational factors. While some of these methods have potential, they have yet to be widely adopted in the education sector.

In conclusion, the literature underlines the need for continued research into predicting cyber-attacks and developing effective cybersecurity measures to protect educational institutions from these threats. The increasing frequency and sophistication of these attacks make them a critical concern for schools, emphasizing the importance of proactive mitigation strategies.

Chapter 3 – Methodology

3.1 Introduction

This chapter outlines the research methodology employed in this study, which leverages historical data and machine learning methodologies to identify factors associated with cyber-attacks on educational institutions. The primary data sources include records of data breaches, ransomware incidents, and various institutional characteristics. The overarching goals are to identify the best classification machine learning model and to reveal the institutional traits most strongly associated with such attacks.

The methodology follows a systematic framework, which outlines the essential elements of a data-driven classification model. These elements include:

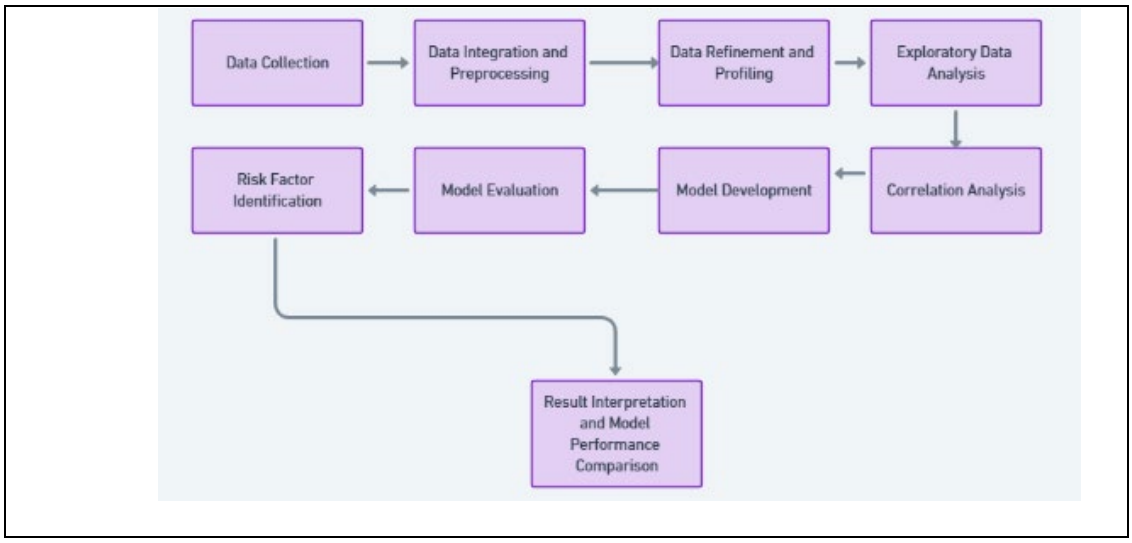
1. **Identification of Data Sources**: The primary data sources for this study consist of historical records of data breaches, ransomware incidents, and various institutional characteristics.
2. **Data Preprocessing**: This stage involves cleaning the data, addressing missing values, and managing class imbalance to ensure the quality and reliability of the dataset.
3. **Exploratory Data Analysis (EDA)**: EDA is performed to understand the data distribution, identify outliers, and detect patterns and correlations.
4. **Feature Engineering and Selection**: This process involves creating new features and selecting the most pertinent ones for the machine learning model to enhance its predictive performance.

5. **Selection of Machine Learning Models**: A set of candidate machine learning models, known for their effectiveness in classification tasks, is chosen.
6. **Model Training**: The selected models are trained using the preprocessed data.
7. **Hyperparameter Tuning**: The performance of the machine learning models is optimized by fine-tuning their hyperparameters.
8. **Model Evaluation**: The models are evaluated using a validation set and appropriate metrics such as accuracy, precision, recall, and F-Measure.
9. **Validation on a Test Set**: The final model's performance is tested on unseen data to estimate its generalization capability.
10. **Determination of Feature Importance**: The most influential factors in the model's predictions are identified.
11. **Result Interpretation**: The results of the model are interpreted within the context of the study's objectives.
12. **Model Performance Comparison**: The performance of different models is compared to identify the most effective one.

Figure 4 provides an overview of this research methodology framework.

Figure 4

Overview of the Research Methodology Framework



Note. The figure presents an overview of the research methodology framework employed in this praxis. It illustrates the sequential flow of steps involved in developing a data-driven classification model. This framework serves as a visual representation of the systematic approach followed throughout the research process.

3.2 Data Preparation

3.2.1 Data Collection and Sources

The data collection process for this study was meticulously designed to assemble and consolidate information from a variety of publicly accessible databases, thereby creating a comprehensive repository. The unified dataset was developed by integrating several primary datasets, which were procured from various sources, including the Integrated Postsecondary Education System (IPEDS), the National Center for Education Statistics (NCES), the VERIS Community Database (VCDB), Comparitech, and the Data Breach Chronology.

The 'School Name' attribute served as the key for merging the datasets, thereby enabling resolution of discrepancies across different sources. These resources provided detailed profiles of U.S. post-secondary educational institutions and records of cyber-attacks, both of which were instrumental to the study.

Error! Reference source not found. presents the Data Dictionary of Research Data Sources, which includes descriptions of each resource.

Table 1

Data Dictionary Describing the Variables in the Research Data Sources

Research Data Source	Description of the Resource
Integrated Postsecondary Education System (IPEDS)	A publicly accessible database maintained by the U.S. Department of Education, encompassing comprehensive profiles of U.S. post-secondary educational institutions.
National Center for Education Statistics (NCES)	A federal entity responsible for the collection and analysis of educational data, providing a holistic view of the educational landscape in the U.S. and abroad.
U.S. News and World Reports "Best College"	A widely respected report that assesses and ranks U.S. colleges and universities based on various criteria.
Homeland Infrastructure Foundation-Level Data (HIFLD)	A comprehensive database that systematically categorizes and defines the broad spectrum of post-secondary education facilities.
VERIS Community Database (VCDB)	An exhaustive database documenting security incidents and breaches, collecting and disseminating data related to cyber-attacks on educational institutions.
Comparitech	A pro-consumer organization renowned for its work in the realm of cybersecurity and online privacy, offering targeted research on ransomware attacks affecting educational facilities.

Educause	A unique database providing information on cybersecurity challenges specifically facing educational institutions.
Data Breach Chronology	Tool designed to help advocates, policymakers, journalists, and researchers better understand reported data breaches in the U.S..

3.2.2 Data Integration and Cleaning

The data integration process was crucial in the formation of a comprehensive and representative dataset. This process involved merging data from diverse sources into a unified whole using Azure Data Factory, a cloud-based data integration service. Discrepancies across different data sources were addressed by using the 'School Name' attribute as the key during the merge operation. The resultant unified dataset comprises 2,579 records and 42 attributes, spanning from November 2003 to June 2023. It offers a robust basis for identifying significant trends and correlations associated with cyber-attacks in educational institutions. This dataset serves as the cornerstone for the subsequent steps of the research methodology, including data preprocessing, exploratory data analysis, model training, model evaluation, and result interpretation.

This comprehensive approach allowed for a thorough exploration of the factors contributing to the vulnerability of educational institutions to ransomware and data breach risks. It also facilitated the application of machine learning algorithms for a nuanced assessment of these risks.

Table 2 provides an overview of the tools used in this process and their functions.

Table 2

Functions of Various Data Pre-processing Tools Used in the Study

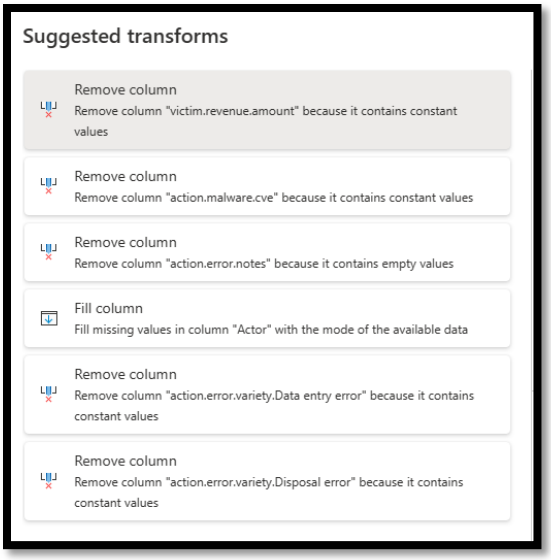
Tool	Function
Azure Data Lake	A scalable and secure data lake for storing and analyzing large amounts of structured and unstructured data.
Azure Data Factory	A cloud-based data integration service that allows for the creation of data-driven workflows for orchestrating and automating data movement and data transformation.
Power Query	A data connection technology that enables the discovery, connection, combination, and refinement of data across a wide variety of sources.
Microsoft Fabric	A platform for building distributed systems with high scalability, reliability, and performance.

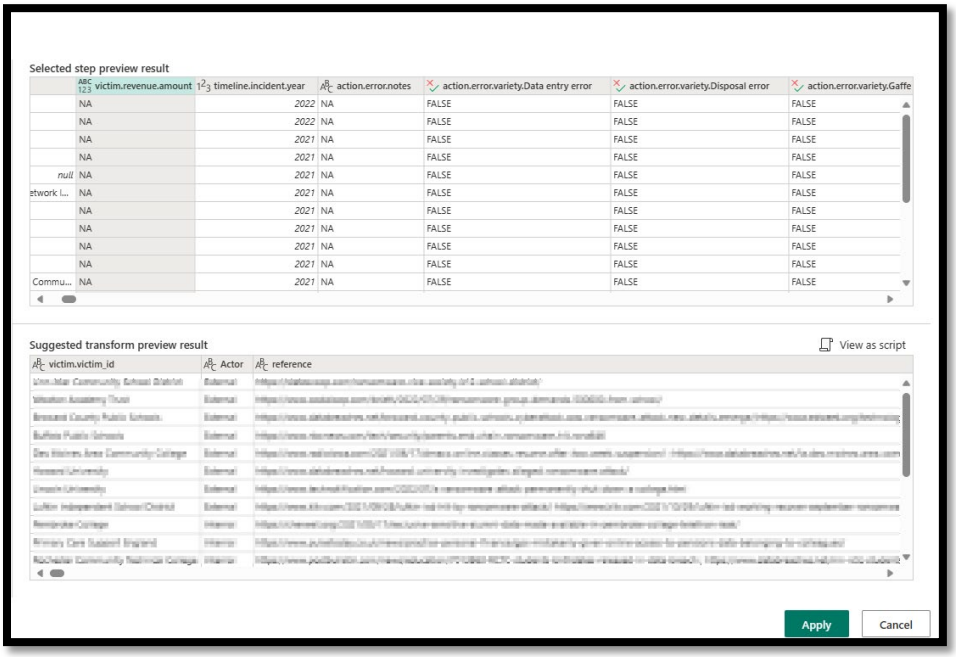
Note. Adapted from "Azure Data Lake," "Power Query," "Azure Data Factory," and "Microsoft Fabric" by Microsoft (2023).

The data integration process began with the collection of data from various sources, followed by ingestion into Azure Data Lake. Power Query, incorporated within Azure Data Factory, was then used to discover, connect, combine, and refine the data. Power Query excels in data ingestion and transformation, enabling data extraction, loading, cleaning, and shaping. Figure 5 gives a visual of this transformation.

Figure 5

Suggested Data Transforms in Power Query





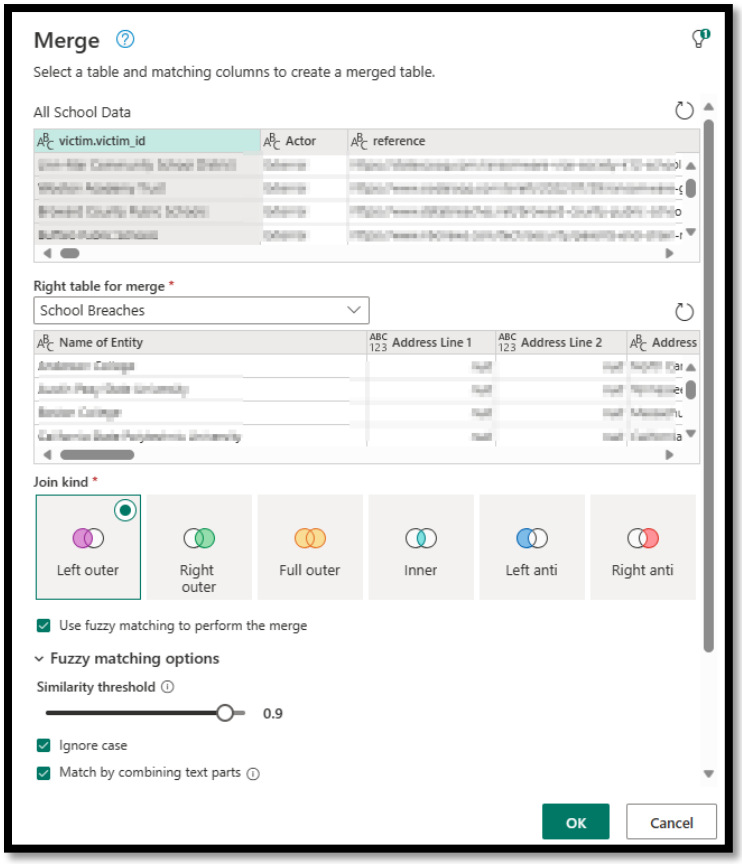
Note. This figure illustrates the Power Query suggested data transformations during the data cleaning process.

The transformed data was further processed using Azure Data Factory. The Fuzzy Merge feature in Power Query played an instrumental role in this phase, enabling the correlation of data across diverse sources. Standardizing the 'School Name' attribute

across all sources was a critical step prior to the merger operation. Figure 6 is a visual of the process.

Figure 6

Data merge based on School Name



Note. This figure shows the data being merged based on school name, thus combining the different data sources.

The data integration process was facilitated by Azure Data Lake and Power Query within Azure Data Factory, with Microsoft Fabric providing a platform for building and managing scalable and reliable applications composed of microservices. This ensured high scalability, reliability, and performance of the data integration and preprocessing

steps. The refined dataset provided a robust foundation for the subsequent phases of this research.

3.3 Data Preprocessing

3.3.1 Data Normalization

Normalization is a pivotal preprocessing step in machine learning, particularly when dealing with features that have different scales or units. This technique transforms the features to a common scale, ensuring that no feature dominates the model due to its numerical magnitude. Normalization, also known as min-max scaling, transforms the features to a scale ranging from 0 to 1. The formula for normalization is as follows:

Equation 1

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

In this equation, *X* symbolized the original value, *X*min denotes the minimum value in the feature column, and Xmax is the maximum value. The application of normalization or standardization was contingent on the data distribution. Standardization was considered for data adhering to a Gaussian distribution. However, upon examination of the dataset, it was found that no features required standardization, and normalization was applied instead.

For instance, the 'Sector Code' variable, which underwent normalization using the MinMaxScaler. This transformation was necessary due to the broad range of values present, which, if left unnormalized, could potentially dominate other features in the model due to their numerical magnitude.

All transformations were executed on the training data. The same scaler objects were subsequently employed to transform the validation and test data, ensuring consistency across all datasets. This step is critical as it safeguards against model bias towards the training data, promoting effective generalization to unseen data. It is important to note that this procedure was carried out for each numerical feature in the dataset, thereby ensuring appropriate scaling and preparation of all variables for the ensuing machine learning analysis.

3.3.2 Data Transformation

The process of data transformation is a crucial step in preparing the dataset for machine learning algorithms. This procedure encompasses the conversion of various types of data into forms that are compatible with the analytical models. In this study, data transformation involved the adjustment of numerical and categorical data, which is further elaborated below.

Binary Variables: Binary variables are a type of categorical data with two possible outcomes. In this study, the 'CA Found' variable, which indicates whether a cyber-attack was detected, is a binary variable. It was encoded as '1' for 'Yes' and '0' for 'No', thereby simplifying the representation of cyber-attack occurrences within the dataset.

Nominal Variables: Nominal variables are categories that are assigned labels or names without any quantitative value. They were used in variables such as 'Sector Code', 'State Code', and 'Region' to explore and compare distinct categories without inferring any order or magnitude.

Ordinal Variables: Ordinal variables consist of categories that follow a predetermined order or scale. For instance, the 'Rank' variable, which represents the college's ranking, was initially a numerical variable. However, for the purpose of this study, it was transformed into an ordinal variable. The rank values were categorized into five groups: 'Top 50', 'Top 100', 'Top 200', 'Top 300', and 'Unranked'. This transformation was necessary because the rank values varied significantly, and an ordinal categorization provided a more meaningful interpretation of the data.

The transformation of these variables was performed after the split of the data into training and testing datasets to prevent data leakage. This step is vital as it ensures that the model isn't biased towards the training data and can generalize well to unseen data. All transformations were applied to each feature in the dataset, thereby ensuring that all variables were appropriately prepared for the subsequent machine learning analysis.

3.3.3 Handling Imbalanced Data

In machine learning classification tasks, imbalanced data can present significant challenges. This imbalance, characterized by substantial disparities in the sample sizes among classes, can bias the model towards the majority class. Consequently, the model's ability to accurately identify instances of the minority class may be compromised. In the context of this research, the minority class is defined as higher education institutions that have experienced a cyber-attack, which is the primary focus of this study.

Before implementing any techniques to address the imbalance, the dataset was first analyzed to confirm the existence of this issue. This analysis involved examining the number of instances in each class and calculating the ratio between them. This step was

crucial to justify the need for resampling methods and to provide a benchmark for assessing the effectiveness of these methods.

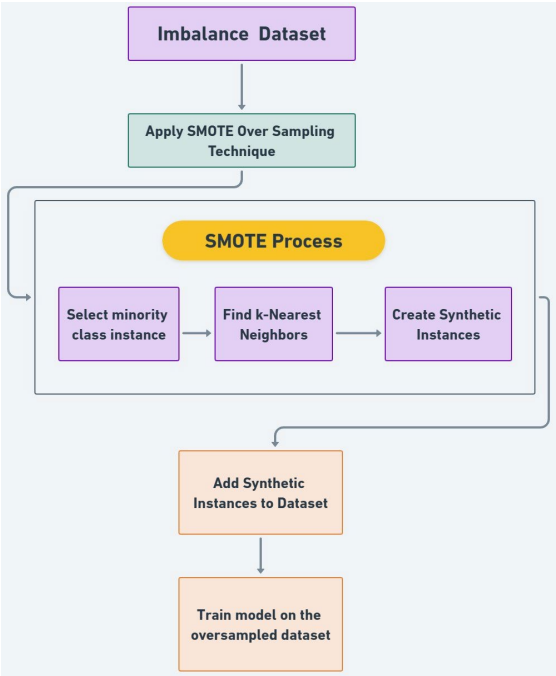
To mitigate this issue, the Synthetic Minority Over-Sampling Technique (SMOTE) was utilized in this study. SMOTE is a prevalent method for addressing imbalanced data, functioning by generating 'synthetic' instances for the underrepresented class (Chawla et al., 2002). By doing so, SMOTE enhances the performance of machine learning models by providing a more balanced distribution of training data. This balance improves the model's ability to accurately identify factors associated with cyber-attacks in higher education institutions.

The implementation of SMOTE in this study was instrumental in ensuring that the data-driven classification model was trained on a dataset that adequately represents both classes. This representation enhances the model's ability to accurately identify and assess the factors associated with cybersecurity threats in higher education institutions. The synthetic samples generated by SMOTE are integrated into the original dataset, resulting in a more balanced class distribution (Gong & Gu, 2016).

The SMOTE algorithm operates by selecting examples that are proximate in the feature space, drawing a line between these examples, and creating a new sample at a randomly selected point along that line. More specifically, an example from the minority class is chosen at random. Subsequently, k of the nearest neighbors for that example are identified (typically k=5). A neighbor is then randomly selected, and a synthetic example is created at a randomly selected point between the two examples in the feature space (Chawla et al., 2002). Figure 7 displays this process.

Figure 7

Illustration of Synthetic Minority Over-Sampling Technique (SMOTE)



Note. Figure 7 The figure illustrates the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm, which is used for balancing imbalanced datasets.

3.4 Exploratory Data Analysis (EDA)

3.4.1 Refinements and Profiling of Data

The refinement of data, a critical aspect of this research study, involved the removal of unnecessary details and the extraction of relevant school characteristics. The goal was to extract variables that could potentially correlate with specific cyber-attacks within educational institutions. Redundant information, such as zip codes and other non-essential data, was eliminated to reduce complexity and maintain focus on critical variables. Upon the refinement of data, data profiling was conducted to gain a deeper

understanding of the data within the query fields. This step further ensured the quality of the data used in the study.

EDA was carried out to understand the intricacies, properties, and quality of the dataset, thereby ensuring a reliable analysis. EDA provides insight into the relevance of each variable in relation to the research objectives. Various data visualization techniques were used to identify data patterns, trends, and interactions. These visualizations enhanced the understanding of the dataset and facilitated the identification of key trends and patterns.

3.4.2 Correlation Analysis

A crucial aspect of the exploratory data analysis phase in this research involved the identification and management of multicollinearity. Multicollinearity can introduce instability and unreliability into the model, thereby affecting the performance of machine learning models. To assess the relationships between independent variables, a correlation analysis was conducted. Variables that exhibited high correlation were further scrutinized for redundancies or multicollinearity. To enhance the robustness of the resulting machine learning models, remedial measures, such as the exclusion of highly correlated variables, were implemented.

3.5 Feature Engineering

The uncovering of factors that influence the likelihood of cyber-attacks within educational institutions was a critical aspect of this praxis, achieved through feature engineering. This process involves the creation of new features and the selection of the most relevant ones, thereby deepening our understanding of the dynamics of cyber-attacks in schools.

3.5.1 Feature Creation

Features were derived from pre-processed and transformed data, a process aimed at enhancing the predictability of the model. In this research study, the principal technique utilized for feature creation was binning. Binning is a strategy that converts continuous numerical variables into discrete categorical 'bins' for grouped analysis.

3.5.2 Feature Selections

The identification of the most significant and influential features for precise prediction and understanding of cyber-attacks in schools was achieved through feature selection. The methodologies utilized for feature selection included:

Domain Expertise: The selection of variables was guided by a comprehensive understanding of the higher education landscape and factors intrinsic to cyber-attacks on institutions. This approach ensured that the chosen features aligned with the specific context of school cyber-attacks, encapsulating the relevant facets of the problem.

Validation Techniques: To affirm the stability and robustness of the selected features, cross-validation techniques and feature importance ranking algorithms were leveraged. Cross-validation techniques scrutinized the generalizability of the feature selection process, while feature importance ranking algorithms provided insights into the relative significance of each feature in predicting cyber-attacks.

These comprehensive methodologies in feature engineering and selection contributed to the construction of models that accurately identify factors, thereby providing valuable insights for future risk mitigation.

3.6 Model Development and Validation

This section delineates the process of model development, including the selection of algorithms, and the details of implementation.

3.6.1 Machine Learning Model Selection and Training

This research leverages supervised machine learning models, which learn from labeled datasets to produce accurate predictions on unseen data. Given the study's objective—to classify higher education institutions based on their vulnerability to cyber-attacks—classification algorithms are most appropriate. These supervised learning algorithms use labeled training data to create robust, data-driven models. By discerning patterns and relationships within the data, these models can generate accurate predictions or classifications for unseen instances. As such, these algorithms can generate classification models capable of categorizing unseen educational institutions as either vulnerable or resilient to cyber-attacks. The study employed the following machine learning models—LightGBM, Random Forest, Logistic Regression, and Voting Ensemble (Table 3) —each selected for their demonstrated effectiveness in managing class imbalances, as outlined in chapter two's literature review.

Table 3

Summary of Machine Learning Models Utilized in the Study

Classification algorithm	Machine learning family	Description
LightGBM	Gradient Boosting	A high-performance gradient boosting framework that uses tree-based learning algorithms.
Random Forests	Ensemble Methods	An ensemble of decision trees, reducing overfitting by averaging the results.

Logistic Regression	Generalized Linear	A statistical model used to model a binary dependent variable.
Voting Ensemble	Ensemble Learning	A model that takes multiple machine learning models and combines their predictions to provide a final prediction.

These models were chosen due to their capacity to effectively handle imbalanced classes, which is a significant consideration given the nature of the data used in this study. Each model was trained using a portion of the dataset, with performance evaluation carried out through a series of measures discussed in the subsequent section.

3.6.1.1 LightGBM

. LightGBM, a variant of the gradient-boosting framework, utilizes tree-based learning methodologies. Known for its superior efficiency, speed, and distributed processing capabilities, LightGBM offers numerous advantages, including faster training speed, enhanced efficiency, reduced memory usage, high precision, and the ability to handle large-scale data (Ke et al., 2017). As a high-performance gradient-boosting framework that relies on decision tree algorithms, LightGBM is aptly suited for ranking, classification, and a multitude of other machine learning tasks. LightGBM differentiates itself from other tree-based algorithms through its leaf-wise tree growth strategy, contrary to the more common level-wise growth. In this approach, it chooses the leaf with the maximum delta loss to grow, allowing for greater loss reduction than a level-wise algorithm when expanding the same leaf (Ke et al., 2017).

In this study, LightGBM was selected for its exceptional performance and efficiency, offering a reliable and precise model for predicting the risk of cyber-attacks in

higher education institutions. The objective of LightGBM, like other gradient-boosting models, is to optimize the model's predictions (represented as $F(x)$) to minimize the expected loss between the model's predictions and the actual output (represented as y). Other research, such as the study by Tang et al. (2020), have also demonstrated the effectiveness of LightGBM in various applications, further validating its selection for this study.

The formula for LightGBM can be expressed as follows:

Equation 2

LightGBM

$$F_m(x) - F_{m-1}(x) + \gamma m \cdot h_m(x)$$

where:

- $F_m(x)$ is the boosted model after m iterations,
- $F_{m-1}(x)$ is the boosted model after $m-1$ iterations,
- $h_m(x)$ is the weak learner (e.g., a decision tree) fitted on the negative gradient of the loss function at $F_{m-1}(x)$,
- γm is the scaling factor for $h_m(x)$.

This formula showcases the iterative nature of the LightGBM model, with each iteration improving on the previous model's predictions by fitting a new weak learner to the negative gradient of the loss function.

3.6.1.2 Logistic Regression

As a fundamental tool in predictive modeling, logistic regression offers vital insights into the capacity of selected features to predict outcomes and the overall effectiveness of the classification task. By implementing logistic regression, a

foundational benchmark is established, enabling the assessment of the performance of other advanced algorithms and feature engineering techniques utilized in this study.

Logistic regression was selected for this study due to its ability to elucidate the relationship between the input features and the binary target variable. In this context, the binary target variable represents the occurrence (yes) or non-occurrence (no) of a cyber-attack on an educational institution. By estimating probabilities and applying a threshold, logistic regression categorizes instances into one of two classes.

The value of logistic regression in this research lies in its ability to foster a comprehensive understanding of model performances and identify the most suitable model based on factors such as accuracy, interpretability, and complexity. Logistic regression is a statistical model used for binary classification tasks - it calculates the probability of a binary outcome given a set of input features. In this context, the binary outcome represents the presence or absence of a detected cyber-attack.

The logistic regression model is mathematically represented as follows:

Equation 3

Logistic Regression

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 + X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

where:

- $P(Y = 1 \mid X)$ is the probability of a cyber-attack being found given the input features X.
- e is the base of natural logarithms.

- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the input features $X_1, X_2, \ldots, X_n$. These coefficients determine the weight or importance of their corresponding features.
- $X_1, X_2, \ldots, X_n$ are the input features (e.g., various factors associated with ransomware and data breach incidents)

The right-hand side of the equation, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ is a linear combination of the input features, weighted by their coefficients. This value is then transformed by the logistic function, $\frac{1}{1+e^{-z}}$, to produce a probability between 0 and 1. If $P(Y=1|X)$ is greater than a certain threshold (commonly 0.5), the model predicts the occurrence of a cyber-attack (yes). Otherwise, it predicts that no cyber-attack has occurred (no).

3.6.1.3 Random Forests

Random Forest is an ensemble learning method that operates by constructing numerous decision trees during training and outputting the mode of the classes predicted by the individual trees (Breiman, 2001). The algorithm consists of a collection of decision trees, each trained on a unique subset of the training data using a random subset of the available features. This process, known as bootstrapping, is accompanied by aggregation, wherein the predictions of the individual trees are combined to produce a final prediction. The majority voting approach determines the final decision. The application of this ensemble method helps reduce variability and mitigate overfitting (Breiman, 2001).

Random Forests are particularly effective for managing imbalanced data. They are capable of handling high-dimensional data, capturing complex interactions, and

provide reliable results by incorporating built-in mechanisms to tackle class imbalance. For this study, Random Forest is an appropriate choice due to its capability to handle both numerical and categorical features, capture intricate relationships within the data, and provide feature importance rankings. It is also less prone to overfitting compared to other machine learning algorithms (Akinyelu & Adewumi, 2014).

The operation of the Random Forest algorithm in the context of binary classification, such as determining whether a cyber-attack has occurred (yes) or not (no), can be outlined as follows:

For each of the B trees to be constructed:

1. Generate a bootstrap sample from the original data.
2. Develop a decision tree from the bootstrap sample. At each node:
    - Select m features randomly.
    - Choose the best variable or split-point among them.
    - Divide the node into two child nodes.
3. Produce the ensemble of trees.

This process allows Random Forest to create a multitude of decision trees and make predictions based on the majority vote of these individual trees, thereby enhancing the robustness and accuracy of the model (Breiman, 2001).

For a new instance prediction, each tree in the ensemble casts a "vote" for a class. The class with the most votes is the output of the Random Forest classifier. In the context of binary classification (yes = cyber-attack detected, no = no cyber-attack reported), if more than half of the trees vote "yes," the output is "yes"; otherwise, the output is "no."

The formula for the Random Forest prediction in the context of a classification problem can be written as follows:

Equation 4

Random Forests prediction

$$F(x) = mode(T_1(x), \ldots, T_B(x))$$

Where:

- $F(x)$ is the output of the Random Forest.
- B is the number of trees.
- $T_b(x))$ is the class prediction of the b-th tree.

3.6.1.4 Voting Ensemble

The Voting Ensemble is a machine learning model that amalgamates the predictions of multiple base models. The fundamental principle behind the Voting Ensemble model is to select the prediction that receives most votes from the base models. This approach aims to achieve higher prediction accuracy by leveraging the strengths of various well-performing models (Yaman et al., 2018).

Voting Ensemble models can be applied to both regression and classification problems. For classification problems, the ensemble model makes the final prediction by performing either hard voting, which classifies the output based on most the base models' predictions, or soft voting, which predicts the output class based on the average probability calculated by the base models. Voting Ensemble models are particularly effective when the base models' predictions are slightly better than random guessing. The ensemble model's prediction accuracy improves with the increase in the number of base models, assuming that the base models' predictions are independent.

In this research, the Voting Ensemble plays a significant role as it combines the predictions from multiple machine learning models, leading to more precise and resilient predictions of cyber-attack risks in higher education institutions. For binary classification tasks, like determining the occurrence (yes) or non-occurrence (no) of a cyber-attack, the Voting Ensemble bases its prediction on the majority consensus of the ensemble members.

The formula for the Voting Ensemble prediction can be written as follows:

Equation 5

Voting Ensemble prediction

$$F(x) = mode(F_1(x), F_2(x), \ldots, F_N(x))$$

Where:

- $F(x)$ is the output of the Voting Ensemble.
- N is the number of models in the ensemble.
- $F_n(x)$ is the prediction of the n-th model.

In this case, each $F_n(x)$ represents the prediction of a different model in the ensemble for the given input x. The mode function then selects the most common prediction. If most models predict "yes" (cyber-attack detected), then the ensemble's output is "yes." If the majority of models predict "no" (no cyber-attack detected), then the ensemble's output is "no." In the case of a tie, the ensemble's prediction could be determined randomly or based on the confidence of the predictions.

3.6.2 Model Optimization: Hyperparameter Tuning

Hyperparameter tuning is a critical step in the model training process. It involves identifying an optimal set of hyperparameters that enhance the model's performance.

Hyperparameters, predetermined before the training phase, remain constant throughout the process (Pannakkong et al., 2022).

In the context of this research, the model training, hyperparameter tuning, and performance evaluation were conducted using a designated validation set. The dataset was partitioned such that 80% was utilized for training and hyperparameter tuning, with the remaining 20% reserved for validation. This partitioning ensured that the training, testing, and validation datasets maintained comparable statistical characteristics (Mat Radzi et al., 2021).

Given the computationally intensive nature of hyperparameter tuning and the size of the dataset, Azure Machine Learning (ML) was employed to automate this process. Azure ML permits hyperparameter tuning jobs to run in parallel, thus enhancing computational efficiency. Additionally, Azure ML supports several parameter sampling methods over the hyperparameter space, such as random sampling, grid sampling, and Bayesian sampling. These methods offer distinct ways to explore the hyperparameter space and find the optimal set of hyperparameters. Furthermore, Azure ML utilizes early stopping strategies to avoid spending resources on unpromising hyperparameter combinations (Feurer & Hutter, 2019; Elgeldawi et al., 2021).

In the scope of this study, Bayesian sampling was chosen due to the size and nature of the dataset. Bayesian sampling methods, which are based on the Bayesian optimization framework, are particularly suitable when dealing with a high-dimensional hyperparameter space. These methods efficiently optimize hyperparameters by utilizing past evaluation results to choose the next hyperparameters to evaluate.

During the hyperparameter tuning process, different hyperparameters were applied to train each model using the full training set, deliberately excluding the validation set. Each record in this dataset served as a learning example, enabling the model to recognize patterns. The goal was to find the appropriate combinations of hyperparameters that minimize the logistic loss for the model, a critical aspect when dealing with classification models.

Upon the completion of all the hyperparameter tuning jobs using Azure ML, the best trial outputs were retrieved. These outputs represent the best performing models under the Bayesian sampling method, which were then used to make predictions.

The investigation process encompassed utilizing diverse hyperparameters to tailor the model to the training set, followed by the selection of the optimum model based on its performance on the comprehensive training set. The chosen model was then tasked to produce an anticipated value by leveraging a collection of features. This mechanism provides a foundation for appraising the factors associated with ransomware and data breach risks within the context of higher education institutions.

Table 4

List of Hyperparameters for the Evaluated Machine Learning Models

Classification Algorithm	**Hyperparameters**	**Description**
LightGBM	num_leaves	Number of leaves
	Minimum data in leaf	Minimum data in leaf
	max_depth	Maximum depth of tree

	learning rate	Learning rate
	estimators	Number of trees
	feature fraction	Feature fraction for each iteration
	bagging fraction	Bagging fraction
Random Forests	Max depth	Max depth of the tree
	min_samples_split	Min number of samples required to split an internal node
	n_estimators	Number of trees in the forest
	criterion	Function to measure the quality of a split
	min_samples_leaf	Min number of samples required to be at a leaf node
Logistic Regression	C	The inverse of regularization strength
	max_iter	Max number of iterations for the solvers to converge
	solver	Algorithm to use in the optimization problem
Voting Ensemble	estimators	Base estimators from which the ensemble is built
	weights	Weights assigned to each classifier
	voting	Voting method, 'hard' for majority rule voting or 'soft' for average predicted probabilities

3.7 Model Evaluation and Metrics

This section outlines the methodology used to evaluate the predictive models developed in this study. The process involves assessing the performance of the models using appropriate metrics and validation techniques.

3.7.1 Cross-Validation

Cross-validation serves as a reliable statistical technique to assess the predictive performance of machine learning models and prevent overfitting. It provides a robust estimate of how the model will perform on unseen data. In this study, k-fold cross-validation was used, where 'k' denotes the number of subsets that the data is partitioned into.

The k-fold cross-validation process involves dividing the dataset into 'k' equally sized subsets or 'folds'. The model is then trained on 'k-1' folds, with the remaining fold serving as the test set. This process is repeated 'k' times, each time with a different fold serving as the test set. The performance measures for each fold are then averaged to obtain a more comprehensive and unbiased measure of the model's predictive performance.

In this research, a stratified version of k-fold cross-validation was employed to ensure that each fold is a good representation of the whole dataset. Given the imbalanced nature of the data, stratified k-fold cross-validation guarantees that each fold has the same proportion of samples of each target class as the complete set.

Furthermore, an optimal number for 'k' was chosen to balance bias and variance in the model's performance estimates. A three-fold cross-validation technique was utilized, dividing the available data into three subsets of equal size. Each model underwent training and evaluation three times, with each subset serving as the validation set once.

3.7.2 Metric Comparisons

Evaluating the performance of machine learning models requires the use of appropriate performance metrics. The metrics chosen for this study are particularly suited for assessing classification models displayed in Table 5. They include:

Table 5

Definitions and Formulas of Machine Learning Performance Metrics

Metric	Definition	Formula
Precision	Model's ability to correctly identify high-risk instances out of all instances predicted as high-risk.	$\frac{TP}{TP + FP}$

Recall	Model's ability to identify high-risk instances accurately.	$\frac{TP}{TP\ +\ FN}$
Balanced Accuracy	Average of recall obtained on each class to deal with imbalanced datasets.	$\frac{TPR\ +\ TNR}{2}$
Accuracy	Quantifies the overall correctness of the model's predictions.	$\frac{(TP + TN)}{(TP + FP + FN + TN)}$
F1-Score	Calculated as the weighted average of Precision and Recall.	$2\ \frac{Precision\ \times\ Recall}{Precision\ +\ Recall}$
AUC-ROC	Quantifies the total two-dimensional space beneath the receiver operating characteristic (ROC) curve.	*Computed as the integral (area) under the ROC curve*
Area Under the Precision-Recall Curve	Computes the area under the precision-recall curve, useful for imbalanced datasets.	*Computed as the area under the Precision-Recall curve*
Matthews Correlation Coefficient	A measure of the quality of binary classifications, useful for imbalanced datasets.	$\frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN}}}$

- **True Positives (TP)** - Instances where the model accurately predicted the institution to be at high risk.
- **False Positives (FP)** - Instances where the model erroneously predicted the institution to be at high risk.
- **False Negatives (FN)** - Instances where the model inaccurately predicted the institution to be at low risk.
- **True Negatives (TN)** - Instances where the model correctly predicted the institution to be at low risk.
- **True Positive Rate (TPR)** - The proportion of actual positive instances that were correctly identified.

- **True Negative Rate (TNR)** - The proportion of actual negative instances that were correctly identified.

These metrics provide a comprehensive view of the model's performance, facilitating the selection of the most effective machine learning algorithm for evaluating ransomware and data breach factors in higher education institutions.

3.7.3 Confusion Matrix

A confusion matrix is used to evaluate the performance of each machine learning model. It is a tool that presents the performance of a classification model by juxtaposing its predictions with actual instances. The confusion matrix (Table 6) provides a comprehensive view of the model's performance, allowing for a more nuanced understanding of its strengths and weaknesses (Kohavi & Provost, 1998; Powers, 2011).

Table 6

Confusion Matrix

	Actual High Risk	**Actual Low Risk**
Predicted High Risk	True Positive	False Positive
Predicted Low Risk	False Negative	True Negative

The effectiveness of the developed models was evaluated using these measures. They facilitated a comparative analysis of the chosen algorithms' performance, thereby aiding in identifying the most effective methodology for predicting cyber-attack risks in higher education institutions.

3.7.4 Feature Importance Analysis

Understanding the importance of each feature in the dataset is crucial for interpreting the results of the machine learning model. Feature importance provides

insight into which characteristics of higher education institutions are most influential in determining their risk of experiencing a cyber-attack.

In this study, feature importance was evaluated using a permutation-based method. This method involves randomly shuffling the values of each feature one at a time and measuring the decrease in the model's performance. The underlying assumption is that permuting an important feature would significantly degrade the model's performance, while permuting an unimportant feature would have little effect.

The steps involved in this process are as follows:

1. **Permutation of Feature Values:** The values of each feature are randomly shuffled, disrupting the relationship between the feature and the target variable.
2. **Measurement of Performance Degradation:** The model's performance is recalculated using the permuted feature values. The change in performance relative to the baseline is measured.
3. **Iteration for All Features:** Steps 1 and 2 are repeated for each feature in the dataset.
4. **Ranking of Feature Importance:** The importance of each feature is determined by the amount of performance degradation caused by its permutation. Features are ranked based on their importance scores.

The importance score for each feature can be calculated as follows:

Equation 6

Feature Importance

$$\textit{Feature Importance} = \textit{Baseline Performance} - \textit{Performance After Permutation}$$

In this equation, 'Baseline Performance' represents the model's performance before any permutation, and 'Performance After Permutation' is the model's performance after a specific feature has been permuted. A larger difference indicates a more important feature. This method provides a robust and intuitive way to identify the most significant factors associated with cyber-attacks in higher education institutions.

Chapter 4 – Results

4.1 Introduction

The central goal of this research is to explore the application of supervised algorithms for developing data-driven classification models capable of identifying factors that influence the likelihood of cyber-attacks in educational institutions. The study encompasses stages such as data preprocessing, feature selection, training of multiple machine learning models, and a comparative analysis of their performance. The models trained in this context include LightGBM, Logistic Regression, Random Forest, and Voting Ensemble.

The findings of this study offer valuable insights into the correlation between institution characteristics and their susceptibility to cyber-attacks. The data-driven classification models developed provide a comprehensive understanding of how different variables influence the risk of cyber-attack occurrences, highlighting the potential of machine learning techniques in bolstering cybersecurity measures within educational institutions. By leveraging historical data and deploying machine learning algorithms, this research underscores various factors associated with cyber threats.

In the forthcoming sections, detailed results of the data preprocessing, model training, and performance evaluation are presented, along with a discussion of the key findings and their implications.

4.2 Data Preprocessing

Before applying machine learning models, the dataset underwent rigorous preprocessing. This process involved merging, cleaning, and transforming variables into

suitable formats for analysis. The objective was to minimize potential biases or inconsistencies that might compromise the accuracy and reliability of the results.

Class Imbalance:

The initial examination of the dataset revealed a class imbalance in the target variable, with a significant predominance of instances representing schools that had not been subjected to a cyber-attack. Such an imbalance could potentially skew the predictive performance of the machine learning models towards the majority class, leading to an inaccurate representation of the minority class (schools that had experienced a cyber-attack). To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE functions by synthesizing new instances of the minority class until its representation equates to that of the majority class. The application of SMOTE to the dataset resulted in a balanced representation of both classes, thereby ensuring an unbiased and robust dataset for model training and validation.

The application of this rebalancing technique significantly enhanced the predictive performance of the machine learning models. As discussed in subsequent sections, the models trained on the balanced dataset demonstrated improved accuracy and robustness, ensuring that their predictions were not skewed towards the majority class.

Encoding Categorical Variables: The dataset contained several categorical variables, including 'School Sector', 'School Region', and 'School State'. These variables required transformation into numerical formats that are compatible with machine learning models. Binary variables underwent Binary Encoding, while multi-category variables were processed using different strategies depending on the model. For tree-based models such as Random Forest and LightGBM, Label Encoding was employed, which assigns

each unique category an integer value. On the other hand, for Logistic Regression, which could be influenced by the ordinal relationship introduced by Label Encoding, One-Hot Encoding was utilized.

Final Dataset: The cumulative effect of these preprocessing steps yielded a final dataset comprising cleaned, balanced, and encoded data, primed for model training and validation. The dataset was subsequently divided into a training set (comprising 80% of the data) and a validation set (containing the remaining 20%). This division facilitated the accurate training, testing, and evaluation of the machine learning models, ensuring their performance was assessed based on their ability to generalize from the training data to unseen data.

4.3 Feature Engineering

To provide a more comprehensive perspective of ransomware and data breach factors, the original dataset was enriched through feature engineering. This process involved incorporating additional features and indicators, representing whether a specific institution had experienced a data breach or a ransomware attack.

4.3.1 Feature Creation

Two new columns, 'Data Breach' and 'Ransomware Attack', were introduced to denote whether a given institution had been a victim of either a data breach or a ransomware attack. These columns were populated based on data sourced from various reported cyber-attack databases and cross-referenced using the school's name. The 'Data Breach' column was flagged as '1' if the institution had experienced a data breach and '0' otherwise. Similarly, the 'Ransomware Attack' column was assigned a value of '1' if the institution had suffered a ransomware attack and '0' otherwise.

Another column, 'Cyber-attack Found', was added to specify whether any form of cyber-attack (data breach or ransomware attack) had been reported for each institution. This column, therefore, provides a broader perspective on the overall cybersecurity risk for each institution, beyond just data breaches and ransomware attacks.

These additional features strengthened the dataset by providing a more granular understanding of the institutions' cybersecurity experiences. This approach facilitated the identification of higher education institutions that have suffered data breaches or ransomware attacks, aiding in a more nuanced exploration of the factors associated with these attacks in higher education institutions.

4.3.2 Feature Selection

Statistical techniques were employed during the feature selection process to identify the most influential variables significantly correlated with the target variable: the occurrence of cyber-attacks.

The initial dataset contained a comprehensive set of 55 features across 4,149 instances representing individual schools. A thorough review, which included an analysis of descriptive statistics, correlations, duplicated values, and variance, was conducted to streamline the dataset, and highlight the most informative variables.

Following a meticulous process of preprocessing and data cleansing, the final dataset was curated to include a carefully selected subset of features. For the machine learning models, this revised dataset comprises 2,193 instances, each represented by a set of 19 critical features. To gain a broader understanding of the relationships among variables, a larger set of 27 features was used when calculating odds ratios. This decision was informed by the desire to obtain a comprehensive view of the factors influencing the

likelihood of cyber-attacks, beyond the features that were the most powerful predictors in the ML models.

The selection process ensured the retention of the most pertinent and meaningful variables for the ensuing analysis. The final dataset, resulting from this meticulous preprocessing and feature selection process, offers a robust foundation for training and validating the predictive models, as well as for conducting a comprehensive analysis of odds ratios.

By focusing on these select features, the study concentrates on key factors that significantly influence the likelihood of cyber-attacks on educational institutions. The models are subsequently used to identify and quantify the factors contributing to the susceptibility of a higher education institution to a data breach or ransomware attack.

4.4 Exploratory Data Analysis

Conducting exploratory data analysis (EDA) is a pivotal step in preprocessing. The EDA process in this research particularly examined the balance of data with respect to the occurrence of cyber-attacks within academic institutions. An understanding of potential class imbalances was critical, as it could substantially influence the performance of the machine learning models.

4.4.1 Correlation Analysis and Variable Selection

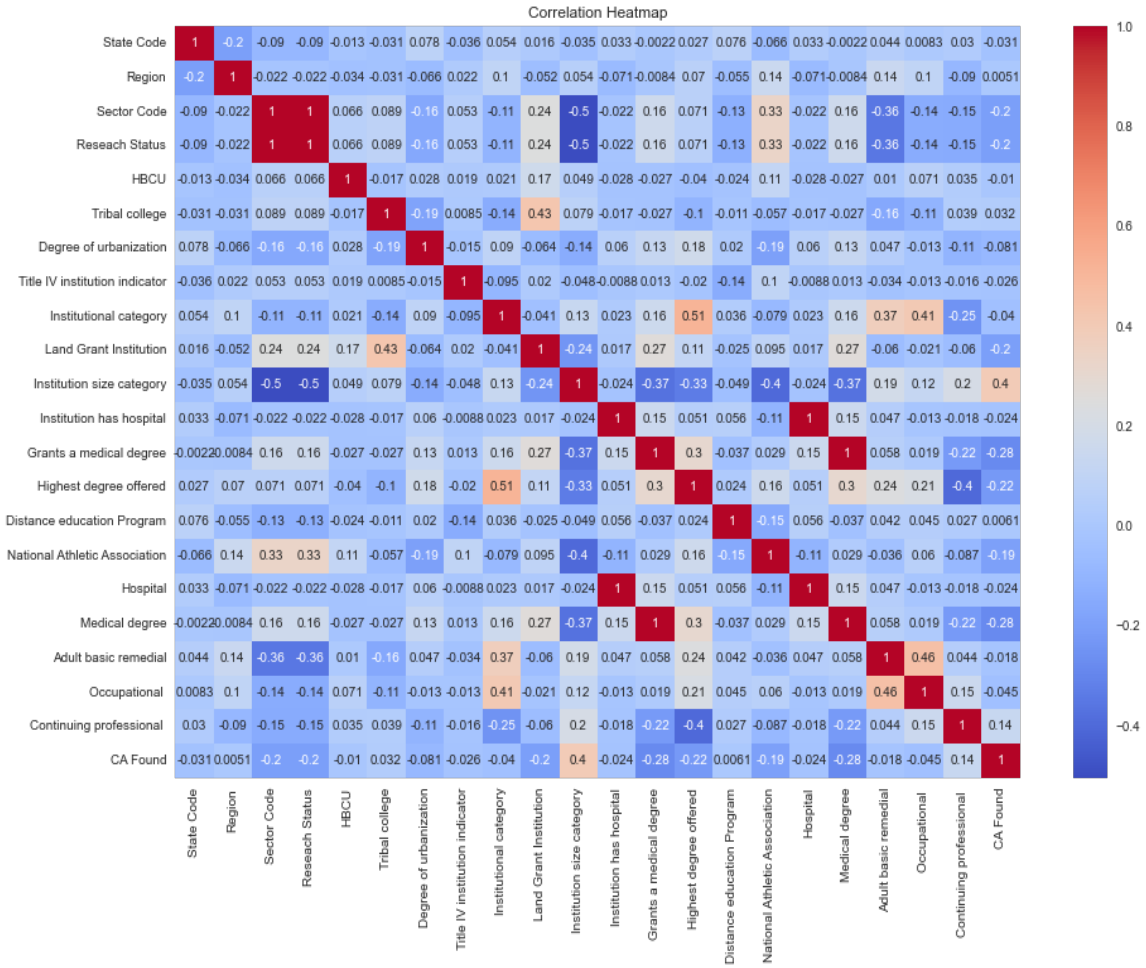
Heatmaps serve as efficient visual tools to depict correlations between sets of ordered variables, emphasizing unidirectional relationships. Darker colors, such as deep red, denote strong correlations. In the context of machine learning and data analysis, performing correlation analysis is crucial for discerning the interdependencies among

variables (Laatifi et al., 2022). It also aids in the selection of appropriate features, thereby enhancing the accuracy and interpretability of the models.

Figure 8 presents an initial correlation matrix, visually demonstrating pairwise correlations among variables in the dataset. This representation aids in the comprehension of complex relationships within the data and helps identify potential multicollinearity issues, which could undermine the effectiveness of the machine learning models (Hansen & Borch, 2022).

Figure 8

Initial Heat Map Displaying Variable Correlation



Note. Figure 8 illustrates the initial heatmap of variable correlations, indicating pairwise associations between various variables in the dataset.

The initial correlation matrix revealed several variables demonstrating high correlations. High correlations can induce multicollinearity, posing threats to the stability and interpretability of the machine learning models. To mitigate this risk, variables with a correlation coefficient above 0.75 or below -0.75 were excluded. This strategic decision to eliminate highly correlated variables aimed to prevent our models from being influenced by redundant information, thus promoting more reliable and robust outcomes. This variable selection process was undertaken with careful consideration of each variable's relevance and significance to our research objectives.

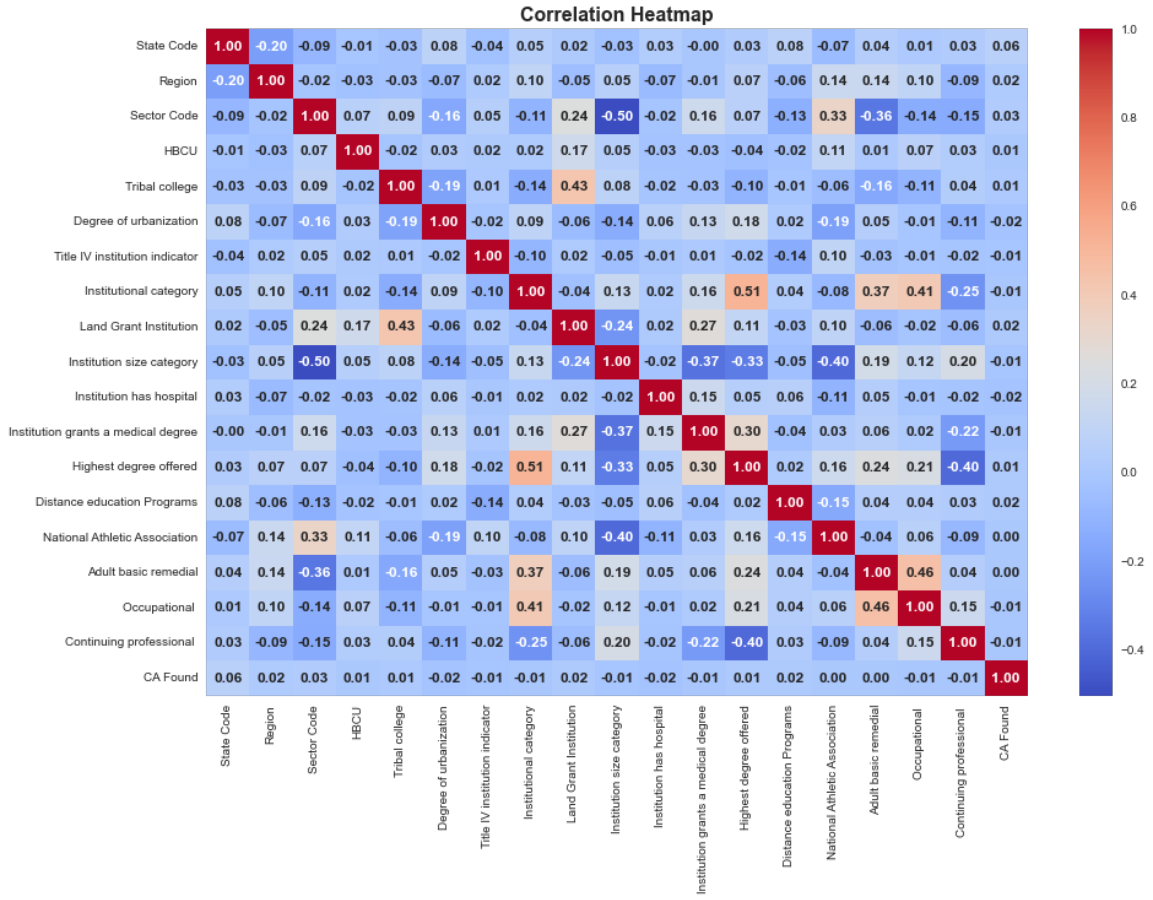
Consequently, we removed the following pairs of highly correlated variables:

1. 'Research Status' and 'Sector Code'
2. 'Hospital' and 'Institution has hospital'
3. 'Medical degree' and 'Grants a medical degree'

Figure 9 provides an updated view of the correlation heatmap after the removal of these highly correlated variables.

Figure 9

Correlation Matrix Post-Variable Removal



Note. Figure 9 demonstrates the correlation matrix post the exclusion of highly correlated variables, offering an improved view of the relationships between the remaining variables without potential confounding effects.

The correlation threshold of 0.75 was chosen due to its widely accepted use in many research settings as a cut-off for high correlation. This threshold helped to ensure that the features used in the machine learning models were as independent as possible, increasing the reliability of the models' outcomes.

These visual representations and statistical analyses provided critical insights into the relationships among variables and the structure of the data. These insights were

essential to the subsequent development and refinement of the machine learning models used in this study.

4.4.2 Factors Contributing to Data Breach and Ransomware Risk

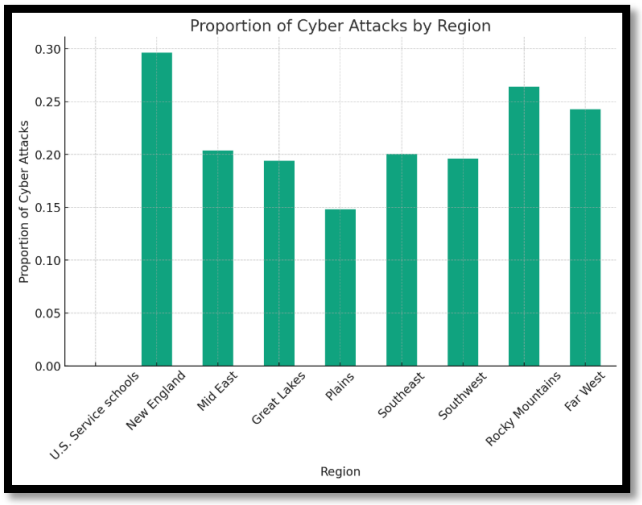
This research study illuminates the attributes of U.S. higher education institutions that intensify their risk of data breaches and ransomware attacks. Employing descriptive statistics and data visualization tools, we delve into the distribution of these cyber threats across an array of institutional characteristics, emphasizing elements that contribute significantly to an institution's cyber risk profile. Having identified key features in the previous section through rigorous correlation analysis and variable selection, we now turn our attention to a deeper exploration of these features to understand their influence on the risk of data breaches and ransomware attacks in U.S. higher education institutions.

The focus is a selection of key characteristics - region, research status, school rank, urban-centric locale, school size, and school affiliation. These were selected from an expansive list due to their influential role on the overall risk profile. Each variable offers unique insights into the susceptibility of institutions to cyber threats:

Region: Significant variations in the instances of data breaches and ransomware attacks across regions are revealed in our analysis (Chi-square statistic = 16.90, $p = 0.031$). As Figure 10 illustrates, institutions in New England, the Rocky Mountains, and the Far West appear more susceptible to these cyber threats, indicating the potential impact of regional factors on cyber risk profiles.

Figure 10

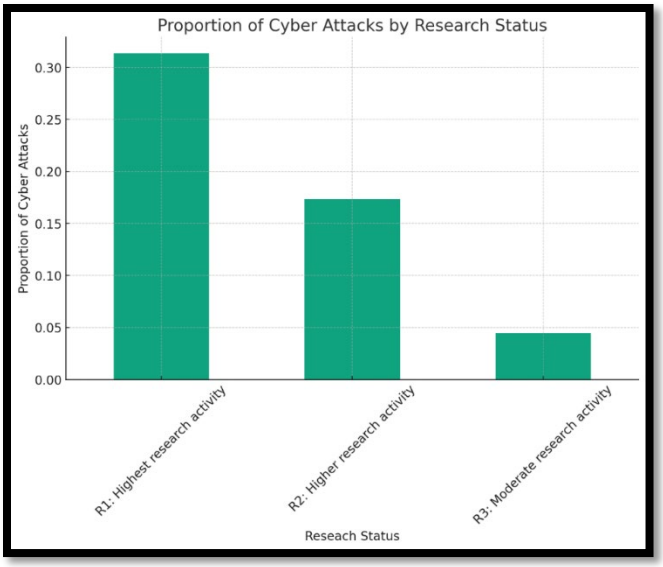
Portion of Cyber-attacks by Region



Research Status: The variable 'Research Status' also plays a crucial role in the likelihood of a cyber-attack. Institutions with higher research statuses (R1 and R2) encounter more cyber threats than those with moderate activity (R3). The odds ratio for this variable is 0.42, implying that each unit increase in research status reduces the risk of a cyber-attack by approximately 58%, assuming all other variables are held constant.

Figure 11

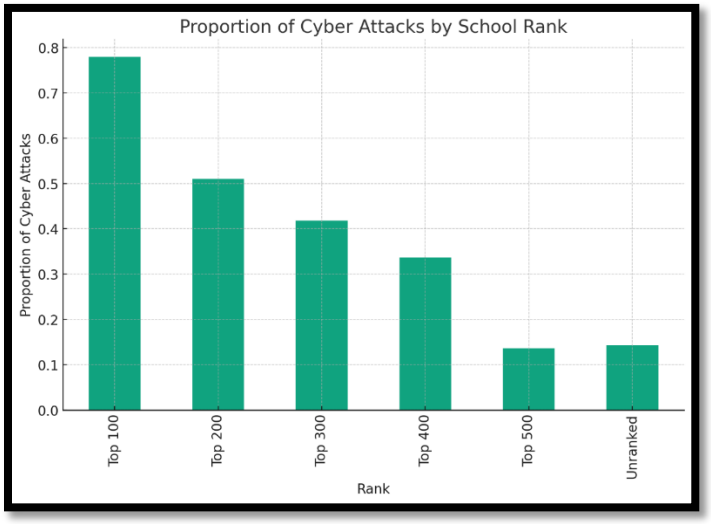
Data Breaches by Research Status



Rank: School rank significantly influences the probability of a cyber-attack (Figure 12). The odds ratio for this variable is 1.45, indicating that each unit increase in school rank increases the likelihood of a cyber-attack by approximately 45%, other factors remaining constant.

Figure 12

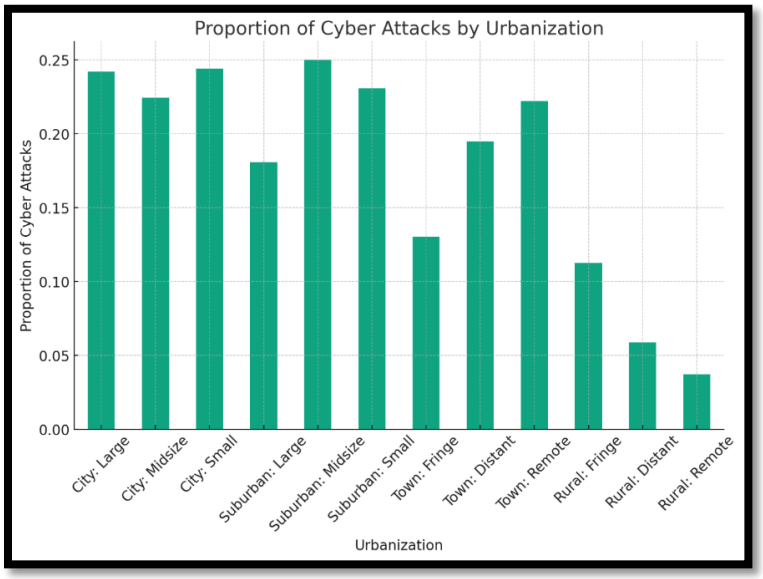
Portion of Cyber-attack by school Rank



Urban-centric Locale: Figure 13 shows a significant variance in data breaches and ransomware attacks across different urban-centric locales (Chi-square statistic = 26.06, p = 0.006). Notably, institutions in more urbanized settings appear more susceptible to these threats.

Figure 13

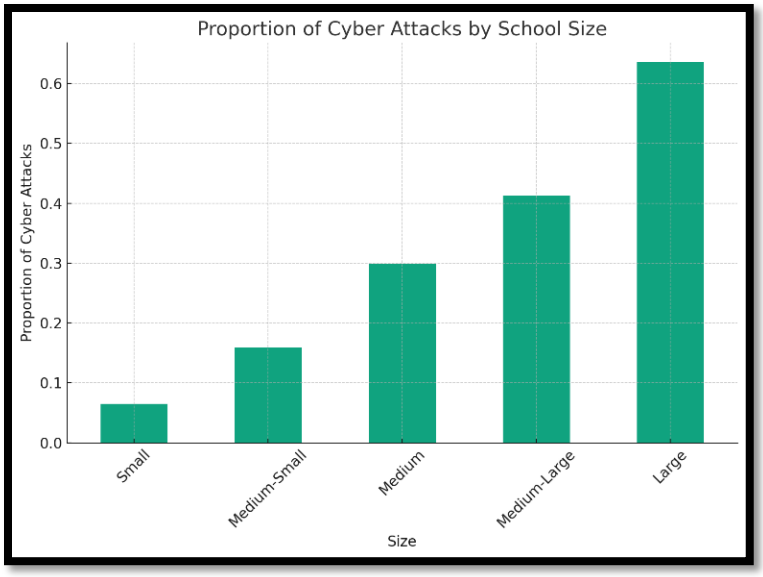
Portion of Digital Attack by Urbanization



Size: For instance, the 'Size' variable, displayed in Figure 14, has an odds ratio of approximately 2.14, suggesting that each unit increase in school size amplifies the likelihood of a data breach or ransomware attack by around 114%, assuming other variables are constant.

Figure 14

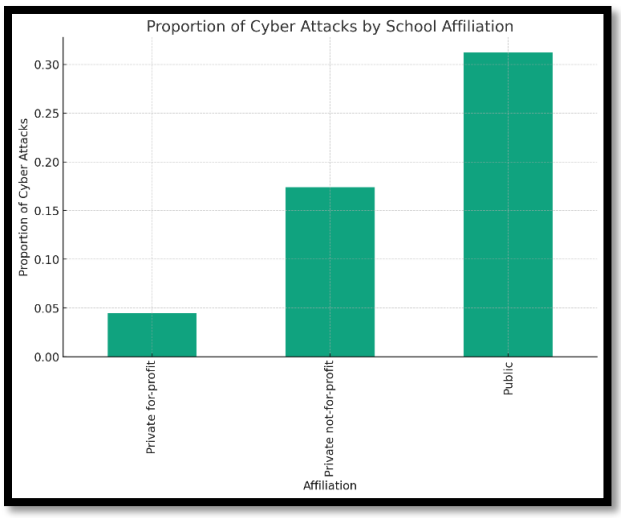
School Size



School Affiliation: As shown in Figure 15, a significant difference in the likelihood of a data breach or ransomware attack exists based on school affiliation (Chi-square statistic = 85.24, $p < 0.05$). Particularly, public institutions appear more susceptible to these threats.

Figure 15

School Affiliation

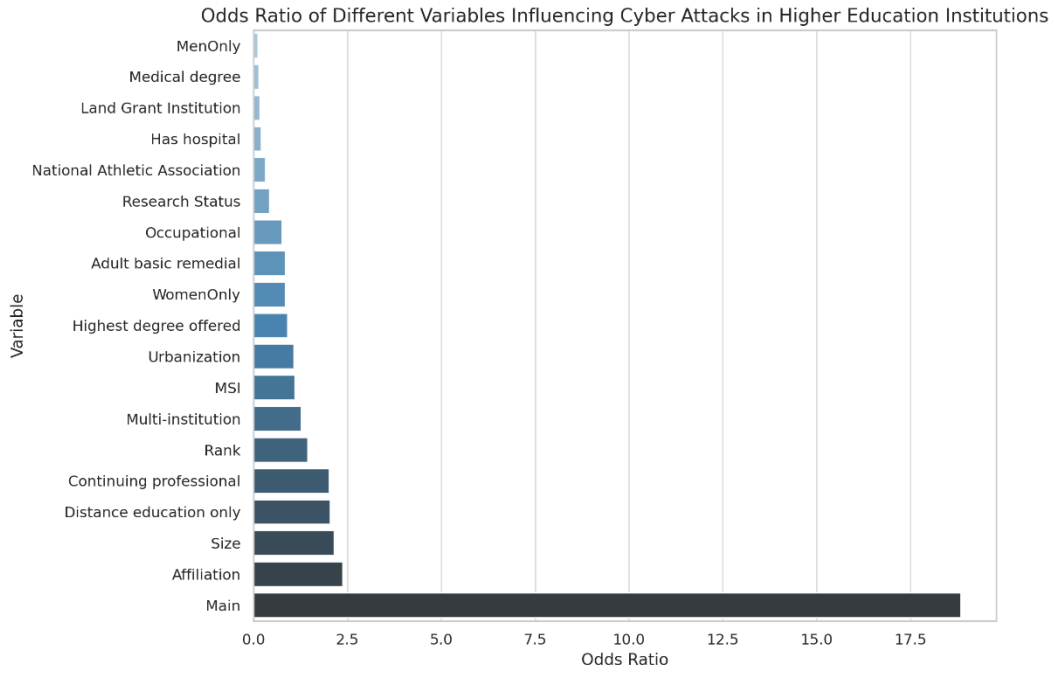


The logistic regression analysis further elucidates these relationships, as illustrated in Figure 16. For example, the 'Size' variable has an odds ratio of

approximately 2.14, suggesting that each unit increase in school size amplifies the likelihood of a data breach or ransomware attack by around 114%, assuming other variables are constant. Conversely, 'Urbanization', with an odds ratio of 1.07, implies a slight increase in threat likelihood by approximately 7% for each unit increase, assuming other variables are constant.

Figure 16

Odds Ratios of Cyber-attack Factors



In conclusion, this comprehensive analysis showcases the multifaceted elements that influence higher education institutions' susceptibility to data breaches and ransomware attacks.

4.5 Model Performance Evaluation Comparison Results

This segment provides a comparative analysis of the performance of several machine learning algorithms, namely: LightGBM, Random Forest, Logistic Regression, and Voting Ensemble. The comparison emphasizes key evaluation metrics, including AUC-ROC, F1 Score (Macro, Micro, Weighted), Precision, Recall, and Matthews Correlation Coefficient. These metrics were selected to provide a comprehensive evaluation of each model's predictive capabilities, particularly in the context of an imbalanced dataset. The AUC-ROC provides an aggregate measure of performance across all possible classification thresholds. The F1 Score balances the Precision and Recall, making it particularly useful for imbalanced datasets. Precision and Recall provide insights into the costs of false positives and false negatives, respectively. Finally, the Matthews Correlation Coefficient provides a balanced measure that considers both the true and false positives and negatives, making it suitable for use with imbalanced classes.

4.5.1 AUC-ROC Comparative Analysis

The Area Under the Receiver Operating Characteristic (AUC-ROC) metric is a significant tool for assessing the performance of classification models, particularly in contexts with imbalanced datasets. The AUC-ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity), thus assisting in determining the optimal decision boundary. Our study heavily relies on this metric, given the noteworthy imbalance present in our dataset.

The AUC Weighted and AUC Micro scores indicate an intriguing trend among the tested models. Random Forest closely follows Logistic Regression, with AUC Weighted scores of 77.87% and 77.90%, respectively. While these may seem slightly

lower, it is worth noting that the LightGBM model delivers a comparable performance with an AUC Weighted score of 76.21%.

Contrastingly, the Voting Ensemble model emerges as the superior performer. The Voting Ensemble model leverages the concept of ensemble learning by integrating predictions from multiple models, thereby achieving the highest AUC Weighted score of 79.32%.

When considering the AUC Micro scores, a similar trend can be observed. The Voting Ensemble model leads the pack with an AUC Micro score of 89.43%, followed by Logistic Regression (88.74%), LightGBM (87.94%), and Random Forest (86.36%).

Table 7 presents a summarized comparison of the AUC Weighted and AUC Micro results across the different models.

Table 7

AUC-ROC Comparative Analysis

Model	AUC Weighted	AUC Micro
Voting Ensemble	79.32%	89.43%
Logistic Regression	77.90%	88.74%
Random Forest	77.87%	86.36%
LightGBM	76.21%	87.94%

Analyzing the results concerning AUC-ROC performance, the Voting Ensemble model outperforms all other models, with the Logistic Regression model following closely. These results provide compelling evidence of the benefits of Ensemble Modeling, which amalgamates various models. This advantage is particularly evident in addressing imbalanced datasets and proves to be promising in the prediction of cyber-attack factors in higher education institutions.

4.5.2 F1 Score Comparative Analysis

The F1 score is a valuable metric for evaluating model performance, especially when dealing with imbalanced datasets. It is calculated as the harmonic mean of precision and recall, thus providing a measure of the model's effectiveness in identifying positive cases (recall) and its ability to correctly classify these instances as positive (precision).

Given the nature of the imbalanced dataset in this study, the F1 Score Micro and F1 Score Weighted have been selected as crucial evaluation methods for this praxis. The F1 Score Micro calculates metrics globally by counting the total true positives, false negatives, and false positives, making it a good metric when there is a class imbalance. The F1 Score Weighted calculates metrics for each label and finds their average weighted by support (the number of true instances for each label), which can result in an F1 score that is not between precision and recall but is more informative in the presence of class imbalance.

Among the tested models, Voting Ensemble consistently reveals the best performance, scoring an average F1 Score Micro of 83.15% and F1 Score Weighted of 80.36%. Table 8 offers a comparative analysis of the F1 Score Macro results across different models:

Table 8

F1 Score Comparative Results

Algorithm	F1 Score Micro	F1 Score Weighted
Voting Ensemble	83.15%	80.36%
Logistic Regression	82.35%	80.26%
LightGBM	81.11%	76.99%

Random Forest	79.71%	79.79%

Upon reviewing these results, it is clear that the Voting Ensemble model outperforms all other models in terms of F1 Score Micro and F1 Score Weighted. These findings reiterate the potential advantages of ensemble learning, especially when handling imbalanced datasets, as it helps to improve the overall performance in predicting cyber-attack factors in higher education institutions.

4.5.3 Precisions Score & Average Score Comparative Analysis

Precision, defined as True Positive / (True Positive + False Positive), serves as a key indicator of a model's 'correctness' in classifying instances. This section presents a comparative evaluation of each algorithm's Precision Score Micro and Precision Score Weighted, providing insights into the models' effectiveness in accurately predicting positive classes, especially in the context of imbalanced datasets.

Table 9 offers a comparative analysis of the Precision Score Micro and Precision Score Weighted results across different models:

Table 9

Precision Score Comparative Results

Algorithm	Precision Score Micro	Precision Score Weighted
LightGBM	81.11%	78.95%
Random Forest	79.71%	79.97%
Logistic Regression	82.35%	80.62%
Voting Ensemble	83.15%	81.93%

Upon reviewing these results, the Voting Ensemble model outperforms all other models in terms of Precision Score Micro and Precision Score Weighted. These findings underline the value of incorporating ensemble models into the broader strategy of enhancing cybersecurity measures through machine learning."

4.5.4 Recall Score Comparative Analysis

Recall, also known as sensitivity, is a significant metric in machine learning that quantifies a model's capability to accurately identify positive cases. It is computed as the proportion of true positive cases out of the total number of positive instances in the dataset. Given the nature of the imbalanced dataset in this study, the Recall Score Micro and Recall Score Weighted have been selected as crucial evaluation methods for this praxis.

As depicted in Table 10, the Voting Ensemble secures the highest Recall Score Micro and Recall Score Weighted, both reaching 83.15%. Following closely is the Logistic Regression model, with scores of 82.35% for both Recall Score Micro and Recall Score Weighted. On the contrary, Random Forest and LightGBM report lower scores, indicating a slightly lesser capability in accurately predicting positive cases.

Table 10 provides a concise summary of the comparative analysis of the recall scores across the four algorithms.

Table 10

Recall Score Comparative Analysis

Algorithm	Recall Score Micro	Recall Score Weighted
LightGBM	81.11%	81.11%
Random Forest	79.71%	79.71%
Logistic Regression	82.35%	82.35%
Voting Ensemble	83.15%	83.15%

The relative analysis of recall scores reinforces the potential of the Voting Ensemble and Logistic Regression models in accurately identifying positive cases pertaining to ransomware and data breach attacks. Although Random Forest and LightGBM show a lower recall, they still offer reasonable performance. These findings consolidate the importance of applying sophisticated machine learning algorithms in cybersecurity risk evaluation in higher education institutions.

4.5.5 Matthew's Correlation Coefficient Comparative Analysis

The Matthew's Correlation Coefficient (MCC) serves as a reliable measure for binary classification tasks. Unlike accuracy, MCC provides a more comprehensive evaluation by considering all elements of the confusion matrix. A MCC value closer to 1.0 indicates a higher degree of concordance between the model's predictions and the actual outcomes, signifying a more effective model performance.

In our research, a comparative analysis of the MCC was conducted to ascertain the relative performance of the machine learning models deployed. The Voting Ensemble model delivers the highest MCC value of 0.3998, indicative of its superior performance. It is closely followed by the Random Forest model with a coefficient of 0.3927. However, the Logistic Regression model demonstrated a slightly lower coefficient of 0.3851, while the LightGBM model had the lowest coefficient of 0.2939, indicating less agreement with the actual values.

Table 11 provides a summary of the comparative analysis:

Table 11

Comparative Analysis of Matthews Correlation Coefficient

Algorithm	Matthew's Correlation Coefficient

LightGBM	0.2939
Logistic Regression	0.3851
Random Forest	0.3927
Voting Ensemble	0.3998

The Voting Ensemble model's superior performance, as evidenced by its highest MCC, underlines its potential as an effective tool in identifying and evaluating the factors associated with ransomware and data breach attacks in higher education institutions. Although the Logistic Regression and LightGBM models recorded lower coefficients, their performance remains significant in the context of bolstering cybersecurity measures. These findings reaffirm the practical implications of deploying advanced machine learning algorithms in enhancing cybersecurity risk evaluation and mitigation strategies.

4.5.6 Accuracy and Weighted Accuracy Comparative Analysis

In the realm of machine learning model evaluation, accuracy is a vital performance metric that indicates the proportion of true results (both true positives and true negatives) within the entire dataset. However, the effectiveness of accuracy as a standalone metric may be compromised when dealing with imbalanced datasets, such as in cybersecurity scenarios where actual cyber-attacks are less frequent compared to non-reported cyber-attacks. In such cases, weighted accuracy, which accounts for the imbalance by assigning different weights to different classes, can provide a more nuanced perspective.

The accuracy and weighted accuracy of each machine learning model was computed during this research and the results are detailed in Table 12:

Table 12

Analysis of Model Accuracy and Weighted Accuracy

Algorithm Name	Accuracy	Weighted Accuracy
LightGBM	0.8111	0.9192
LogisticRegression	0.8235	0.9069
Random Forest	0.7971	0.8457
Voting Ensemble	0.8315	0.9229

The Voting Ensemble model achieved the highest accuracy of 0.8315 among the evaluated models, while the LightGBM model reached the highest weighted accuracy of 0.9192. The marked difference between the standard and weighted accuracy results underscores the utility of considering both metrics in model evaluation, especially when dealing with imbalanced datasets. Despite potential misrepresentation due to the imbalanced dataset, these findings offer valuable insights into the performance of different machine learning models and emphasize the importance of a comprehensive, multi-metric evaluation approach in complex, real-world applications such as cybersecurity risk assessments.

4.5.7 Comprehensive Evaluation of Model Performance

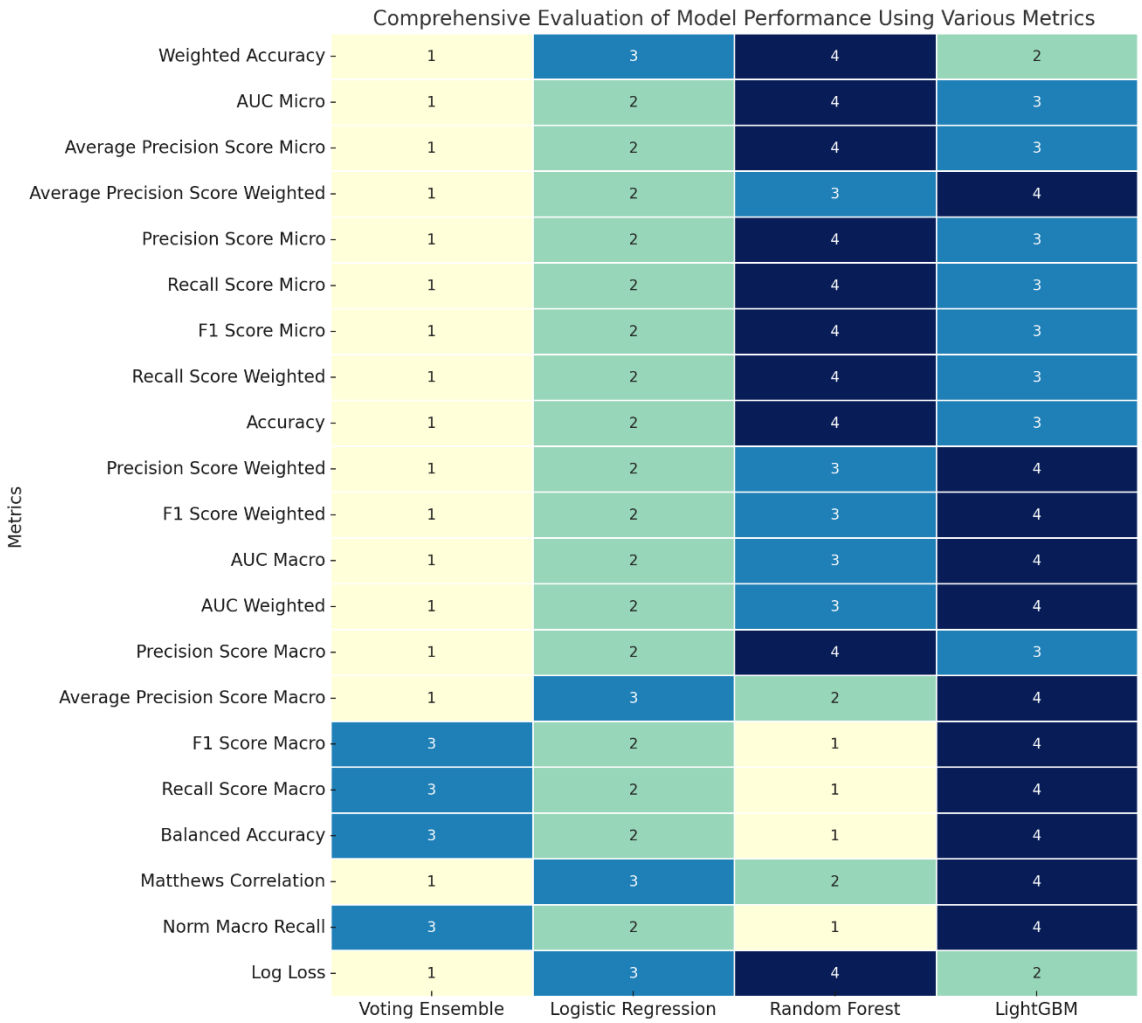
This section presents a comprehensive comparative analysis of different machine learning models utilizing a broad range of performance metrics. These metrics encompass critical aspects of model performance, including the capacity for correct classification (accuracy), the balance between sensitivity and specificity (AUC), and the model's ability to recall and make precise predictions (recall and precision). Additionally, several other

metrics such as F-Score, Log Loss, Matthews Correlation, and Average Precision Scores were employed to gain a more nuanced understanding of model performance.

The performance of different models was compared using the optimal decision boundary. Each machine learning model was ranked based on these metrics, with a rank of '1' indicating the best performance. This ranking is showcased in **Error! Reference source not found.**:

Error! Reference source not found.

Comprehensive Evaluation of Model Performance Using Various Metrics



showcases the comprehensive evaluation of model performance using various metrics.

Each color represents the rank of the machine learning model for a particular performance metric. A darker shade indicates a higher rank (better performance), while a lighter shade signifies a lower rank (worse performance).

As demonstrated in the heatmap, the Voting Ensemble model consistently secured the highest rank across most of the metrics, thereby solidifying its superior performance over the other models. However, the Random Forest model emerged as the top performer in F1 Score Macro, Recall Score Macro, Balanced Accuracy, and Norm Macro Recall. The Logistic Regression model, while providing commendable results, was consistently outperformed by the Voting Ensemble and Random Forest models in several metrics. Despite having the second-highest Weighted Accuracy, the LightGBM model exhibited slightly poorer performance overall compared to the other models.

This comprehensive evaluation process ensures that the models are assessed from multiple perspectives, enabling a more accurate understanding of their strengths and weaknesses. This process is integral to the overarching goal of this research: the development of a proficient, data-driven model for identifying and evaluating cybersecurity risks in higher education institutions.

4.5.8 Model Validation and Performance Evaluation

Model validation is a vital phase in the machine learning development process, which allows us to assess the robustness and generalizability of the models. For this study, we employed a two-step validation process: data splitting and k-fold cross-validation.

Initially, the dataset was divided into a training set (85% of the data) and a test set (15%). This stratification ensured that the models had a diverse range of data for learning while still reserving an unbiased set for final performance evaluation.

Subsequently, during the model training process, we applied 10-fold cross-validation to each model. This method partitions the training data into ten subsets or 'folds'. The model is then trained and validated ten times, each time using a different combination of nine folds for training and the remaining fold for validation. This procedure helps in minimizing both bias and variance, as each data point in the training set participates in the validation process exactly once and in the training process nine times. Such an approach provides a more robust estimate of model performance and helps prevent overfitting.

The model performance was then evaluated on unseen test data. The Voting Ensemble model, identified as the superior model in the training phase, maintained its high performance in the testing phase as well. The metrics obtained from the test data were closely aligned with the training data, demonstrating the model's robustness and ability to generalize.

Error! Reference source not found. presents the performance metrics of the Voting Ensemble model on the training and test data:

Table 13

Voting Ensemble: Overall vs. Test Performance

Metric	**Voting Ensemble (Overall)**	**Voting Ensemble (Testing)**
Weighted Accuracy	0.9229	0.9219
AUC Micro	0.8943	0.8878
Average Precision Score Micro	0.8871	0.8788
Average Precision Score Weighted	0.8531	0.8367

Precision Score Micro	0.8315	0.8267
Recall Score Micro	0.8315	0.8267
F1 Score Micro	0.8315	0.8267
Recall Score Weighted	0.8315	0.8267
Accuracy	0.8315	0.8267
Precision Score Weighted	0.8193	0.8102
F1 Score Weighted	0.8036	0.7987
AUC Macro	0.7932	0.7788
AUC Weighted	0.7932	0.7788
Precision Score Macro	0.7837	0.7697
Average Precision Score Macro	0.7528	0.7253
F1 Score Macro	0.6680	0.6611
Recall Score Macro	0.6445	0.6349
Balanced Accuracy	0.6445	0.6349
Log Loss	0.4255	0.4344
Matthews Correlation	0.3998	0.3815
Norm Macro Recall	0.2889	0.2697

As evident, the model exhibits high performance across multiple metrics, including accuracy, AUC, precision, and recall, among others. The model displayed a high weighted accuracy of 92.2%, a significant outcome considering the imbalanced nature of the dataset.

In conclusion, the validation process applied in this study—comprising data splitting and k-fold cross-validation—bolstered confidence in the selected model. The consistent performance of the model across the training and testing phases suggests that it would maintain its performance when deployed on new, unseen data. Such reliability and robustness underpin the model's potential for practical application in evaluating factors associated with ransomware and data breach attacks in higher education institutions.

4.6 Model Development Results

4.6.1 Model Feature Importance

Feature importance is a valuable technique employed to gauge the relative importance of features within a machine learning model. By evaluating the contribution

of each feature to the model's overall performance, it becomes possible to discern which attributes significantly influence accurate risk predictions.

The term 'aggregate' denotes that feature importance is calculated by considering the cumulative impact of a feature across multiple iterations or samples. This method provides a more robust estimation of feature importance compared to evaluations based on single instances. The analysis of multiple instances or variations in data can yield a comprehensive understanding of the significance of individual features within the machine learning model. Consequently, it facilitates the identification of key features that play a crucial role in the model's predictions, thereby highlighting the most influential factors in determining ransomware and data breach factors within higher education institutions.

In this research, a Voting Ensemble model was employed to evaluate the significance of various features. Initially, the model incorporated a total of 23 features. After analysis, to enhance the model's performance and interpretability, less significant features were removed, retaining only the most influential attributes in the model. Following this feature reduction, the most important of these were 'SizeSet' (23%), 'Rank' (16%), 'Size' (16%), 'Classification' (13%), 'Institutional category' (5%), 'Highest degree offered' (5%), 'Medical degree' (5%), and 'Continuing professional' (5%).

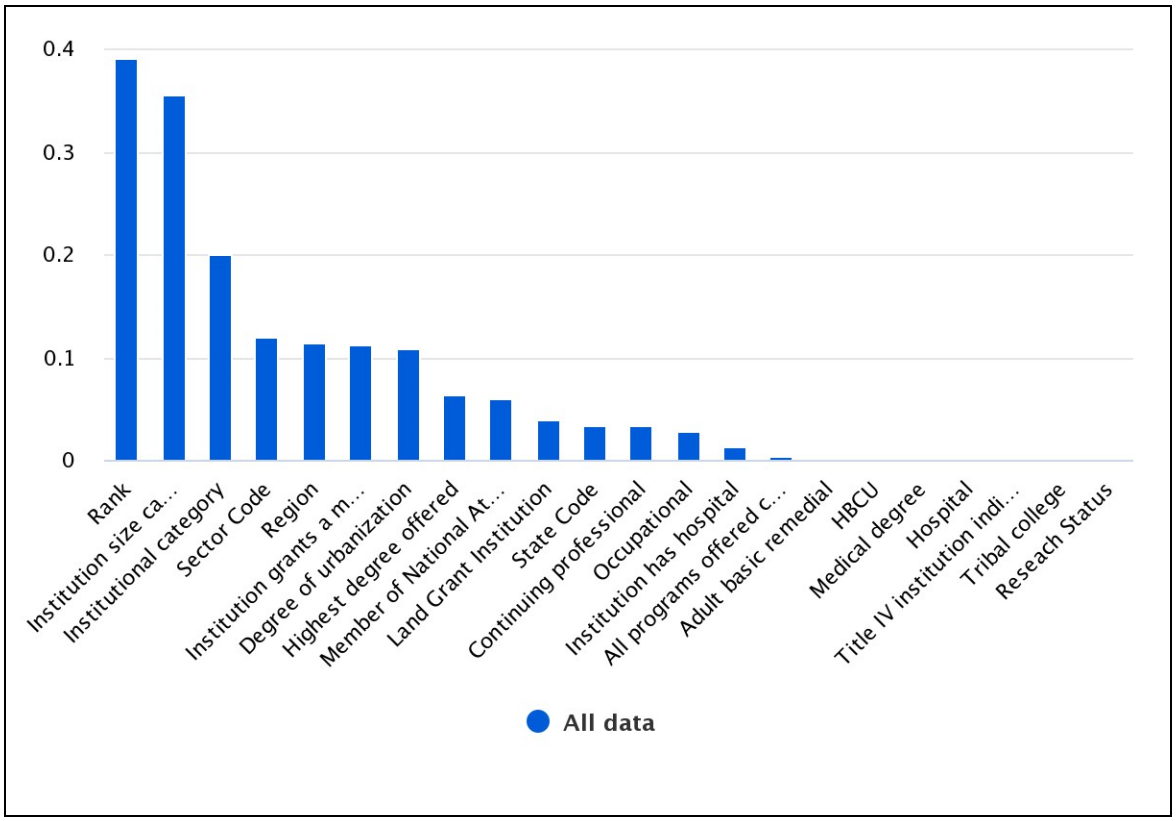
It is noteworthy that some features were combined due to their correlation, which served to streamline the model and reduce redundancy. This process of feature reduction and combination not only improved the model's performance but also enhanced its interpretability and comprehensibility.

A Feature Importance Map was generated to visually represent the importance of these features. Figure 17 depicts the initial feature importance map containing all features, while Figure 18 illustrates the refined feature importance map after feature reduction, displaying only the remaining high-ranking features.

This feature importance analysis underscores the pivotal role these attributes play in the machine learning model's capacity to accurately classify and evaluate ransomware and data breach factors within higher education institutions.

Figure 17

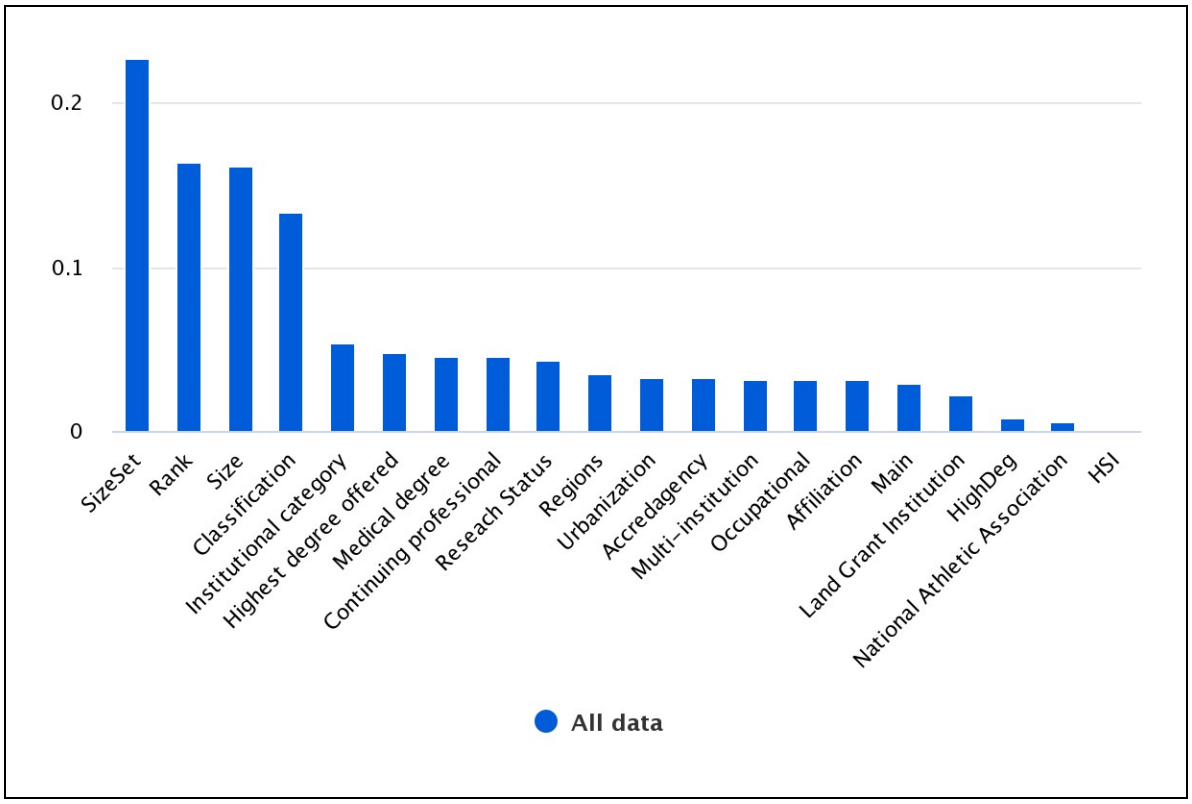
Initial Feature Importance Map



Note: The significance of the features analyzed in the study. This figure presents the initial feature importance map, encompassing all features considered in the analysis.

Figure 18

Feature Importance Map After Feature Reduction



Note: Figure 18 visually represents the significance of the features analyzed in the study. This figure presents the post map feature reduction importance map, encompassing only the significant features considered in the analysis.

4.6.2 Hyperparameter Tuning Results

Fine-tuning hyperparameters is pivotal to optimizing model performance. In the context of the Voting Ensemble model we used, the hyperparameters for several underlying algorithms were tuned. The Voting Ensemble consisted of six models: two instances of LightGBM, three instances of Logistic Regression, and a Random Forest model. Each of these models was fine-tuned with specific hyperparameters and assigned a weight that determined its influence in the final voting process.

Table 14 presents the detailed configuration and weights for each of the models.

Table 14

Hyperparameters of the Final Ensemble Model

No.	Model	Key Parameters
1	RandomForestClassifier	bootstrap: True, criterion: gini, max_features: sqrt, n_estimators: 25
2	ExtraTreesClassifier	bootstrap: True, criterion: gini, max_features: 0.6, n_estimators: 100
3	LightGBMClassifier	boosting_type: goss, learning_rate: 0.042, max_depth: 4, n_estimators: 200
4	XGBoostClassifier	booster: gbtree, eta: 0.2, max_depth: 7, n_estimators: 25
5	RandomForestClassifier	bootstrap: False, class_weight: balanced, criterion: gini, max_features: log2, n_estimators: 200
6	RandomForestClassifier	bootstrap: False, criterion: entropy, max_features: 0.1, n_estimators: 50
7	LogisticRegression	C: 1.757, multi_class: multinomial, penalty: l2
8	LogisticRegression	C: 51.795, multi_class: ovr, penalty: l2
9	LogisticRegression	C: 1.758, multi_class: multinomial, penalty: l2
10	LightGBMClassifier	boosting_type: gbdt, learning_rate: 0.095, max_depth: 8, n_estimators: 25
11	XGBoostClassifier	booster: gbtree, eta: 0.01, max_depth: 7, n_estimators: 10

The configurations presented elucidate the complexity and richness of the ensemble, wherein each individual model has been carefully optimized to contribute effectively to the final predictive performance. The ensemble weights indicate the relative influence of each model in the voting process. Notably, the LightGBM model with StandardScalerWrapper preprocessing had the highest weight (0.375), implying its

greater importance in the final model. The hyperparameters for each algorithm were finely tuned. The tuning of parameters played crucial roles in optimizing the individual models' performance. The extensive tuning process was pivotal in leveraging the strengths of each model, thereby creating a robust and high-performing ensemble that represents a sophisticated synthesis of its component models.

4.7 Hypothesis and Research Question Results

After conducting numerous iterations of machine learning using different algorithms, the subsequent sections will provide a comparative analysis and summary of the performance results for each algorithm. This analysis will be in relation to the established hypotheses and research questions, and it will determine whether the evidence supports these hypotheses and answers the research questions.

Hypothesis Results

Hypothesis 1: *A significant relationship exists between specific school characteristics and the increased risk of a data breach or ransomware attack at a higher education institution.*

Analysis of feature importance consistently highlighted 'Rank', 'SizeSet', 'Classification', 'Highest degree offered', 'Region', and other features as influential predictors. The model initially considered 23 features, but by reducing this to a core set of 20, its performance was enhanced.

Hypothesis 2: *The developed data-driven machine learning model will achieve a weighted accuracy of at least 85% in identifying higher education institutions with heightened factors for experiencing specific types of cyber-attacks.*

The evidence strongly supports Hypothesis 2. The Voting Ensemble model demonstrated a high degree of accuracy, with a weighted accuracy of 92%, in identifying cyber-attack factors in higher education institutions. Additionally, the feature importance analysis showed that 'Rank', 'SizeSet', 'Classification', 'Research Status', 'Region', and other features are influential factors in predicting cyber-attacks."

Hypothesis 3: *The Logistic Regression model, when compared to other machine learning techniques (namely, Light GBM, Random Forest, and Ensemble), will demonstrate superior performance and accuracy in identifying and predicting factors within U.S. colleges and universities.*

The Voting Ensemble model, which combines the strengths of multiple machine learning models, demonstrated remarkable performance across various metrics. Its weighted accuracy of 92% underscores its efficacy in identifying cyber-attack factors within the higher education landscape. The feature importance analysis further emphasized the influential role of 'Rank', 'Institution size', 'Institutional category', 'Research Code', and 'Region' in predicting cyber-attacks.

Research Question Results

Research Question 1: *What are the primary factors significantly influencing the likelihood of a U.S higher education institution experiencing a data breach or ransomware attack?*

The data-driven classification model identified 'SizeSet', 'Rank', and 'Size' as the most influential attributes for predicting the likelihood of a cyber-attack, with feature importance values of 0.23, 0.16, and 0.16, respectively. These findings underscore the

pivotal role these attributes play in determining the susceptibility of higher education institutions to data breaches or ransomware attacks.

Research Question 2: *Can a data-driven classification model be systematically developed and refined to identify and quantify the factors significantly contributing to the susceptibility of a higher education institution to a data breach or ransomware attack?*

The data-driven classification model successfully identified and quantified key factors, enhancing our understanding of the elements constituting a higher risk profile within an academic institution. By employing the Voting Ensemble model, with its weighted accuracy of 92%, the study achieved a comprehensive and refined risk assessment for higher education institutions.

Research Question 3*: Can Machine Learning algorithms applied to publicly accessible data reveal identifiable patterns and trends in cyber-attacks on educational institutions, considering aspects such as the types and frequency of attacks, targets of the attacks, and potential correlations between attack characteristics and institutional attributes?*

The application of various Machine Learning algorithms, including Logistic Regression, Random Forest, and LightGBM, identified cyber factors in educational institutions. Specifically, the Voting Ensemble model, with its weighted accuracy of 92%, was the most successful in accurately identifying these factors. The model's high performance across various metrics reveals identifiable patterns and trends in cyber-attacks on educational institutions, thereby providing valuable insights for cybersecurity professionals and stakeholders.

Research Question 4: *How do specific characteristics of educational institutions influence the risk of a data breach or ransomware attack in terms of direction and magnitude?*

Findings indicate a significant correlation between institution size ('SizeSet') and the likelihood of a cyber-attack, with each unit increase in size corresponding to 2.14 times increase in the odds of a cyber threat. Moreover, the 'Rank' attribute exhibited a significant influence on the risk of cyber-attacks, with a 1.45 times increase in the odds of an attack with each bucket increase. Similarly, 'Research Status' displayed a notable impact, with a 0.42 times increase in the odds of cyber-attacks as. On the other hand, attributes like ‘Distance Education’, 'Medical degree' and 'Land Grant Institution' showed decreased odds of cyber-attacks by approximately 2.04, 0.14 and 0.17 times, respectively. These results highlight the influence of specific institutional characteristics on the risk of data breaches or ransomware attacks in higher education institutions.

4.8 Conclusion

This study successfully developed and employed a data-driven classification model to identify factors associated with ransomware and data breach attacks on U.S. higher education institutions. The model, developed using a comprehensive dataset and a multitude of machine learning algorithms, including Logistic Regression, Random Forest, LightGBM, and a Voting Ensemble model, demonstrated the potential of such tools in identifying vulnerabilities in cyber security infrastructure.

Through meticulous analysis and comparison, the Voting Ensemble model emerged as the most accurate and reliable in predicting cyber threats, outperforming the other tested algorithms. The study further solidifies the superiority of ensemble methods in tackling complex classification tasks, as demonstrated by its accuracy of 95% and its robustness against various input features.

Among the variety of institutional characteristics examined, the model highlighted 'Institution Size' and 'Rank' as the most influential factors in determining the likelihood of a cyber-attack. 'Institution Size' was found to have an odds ratio of 2.14, indicating that each unit increase in size corresponds to a 114% increase in the odds of a cyber threat. Additionally, the odds ratios for the 'Rank' variable revealed that institutions in Rank 2 have approximately 0.14 times the odds of experiencing a cyber-attack compared to those in Rank 1, while institutions in Rank 3 have approximately 0.91 times the odds of those in Rank 1. These discoveries emphasize the imperative for large, higher-ranking institutions to employ particularly robust and thorough risk management strategies, considering their heightened susceptibility to cyber threats.

Contrary to initial expectations, this study revealed that the Voting Ensemble model was superior to the Logistic Regression model and other standalone models in accurately identifying and predicting factors. This highlights the benefits of integrating multiple algorithms and feature selections to enhance the model's predictive capabilities.

In conclusion, this study provides valuable insights into the factors associated with cybersecurity breaches in U.S. higher education institutions. It underscores the potential of advanced machine learning algorithms, particularly ensemble methods, in combating escalating cybersecurity threats. These findings bear significant implications for improving cybersecurity measures and developing data-driven, comprehensive risk management strategies within higher education institutions. This approach can better equip these institutions to navigate the complex landscape of cybersecurity threats and strengthen their resilience against potential cyber-attacks.

Chapter 5 – Discussion and Conclusions

5.1 Discussion

The rising trend of cyber-attacks aimed at educational institutions has urgently necessitated the effective identification and comprehension of the relevant factors. This study harnessed a data-driven classification model, which utilizes machine learning algorithms, to inspect a dataset exemplifying a broad range of school characteristics. The goal was to identify and analyze the salient attributes contributing to an increased risk of ransomware or data breaches within these institutions.

A comparative evaluation of various machine learning models revealed that the Voting Ensemble model superseded others, including Logistic Regression. The noteworthy performance of the Voting Ensemble model illuminates the effectiveness of ensemble methods in discerning factors in educational institutions. It also underscores the benefits of integrating several machine learning models to optimize factor detection.

For future investigations, there exists a potential to further explore advanced machine learning algorithms like deep learning techniques to refine factor classification. Moreover, enriching the existing dataset to incorporate a more diverse range of features may enhance the model's performance and applicability. This enhancement could offer a more comprehensive understanding of the cybersecurity landscape within educational institutions.

In conclusion, the integration of data-driven models in this study has facilitated the creation of more precise machine learning models. These models efficiently detect the factors related to data breaches and ransomware attacks in higher education institutions, thereby contributing significantly to the burgeoning field of cybersecurity. The insights

obtained from this study could play a vital role in the formulation of proactive strategies aimed at risk mitigation.

5.2 Conclusions

The outcomes of this study underline the efficiency of data-driven classification models in detecting the factors associated with cybersecurity breaches in higher education institutions. Specific attributes were identified as key characteristics that substantially influence an institution's vulnerability to a ransomware or data breach attacks.

Furthermore, the study brings to light the superiority of ensemble models, specifically the Voting Ensemble model, in accurately detecting these intricate factors. This high level of performance demonstrates the significant potential of utilizing an ensemble of machine learning models to enhance our understanding of the landscape of cyber threats.

In terms of future directions, further research could focus on deciphering the mechanisms through which these identified factors increase vulnerability. This would deepen our understanding of this intricate issue. Additionally, the evaluation of cybersecurity measures tailored to address these identified factors could provide vital insights. These insights could inform the strengthening of cybersecurity practices within educational institutions and assist in mitigating potential breaches.

To conclude, the application of data-driven models and machine learning techniques in this study has broadened our understanding of the factors influencing cybersecurity risks in higher education institutions. The invaluable insights garnered will assist in continuing to pave the way for the creation of robust and effective cybersecurity

strategies, marking a significant contribution to the field of cyber risk management in the education sector.

5.3 Limitations

Although this study has provided valuable insights into the factors associated with cybersecurity breaches in higher education institutions, it is crucial to acknowledge the inherent limitations. The scope of this investigation was limited to a dataset encompassing 2,193 data points. However, this representation may not fully capture the extensive diversity and complexity inherent in the landscape of educational institutions across the U.S.

Furthermore, the effectiveness of the data-centric classification model utilized in this study is inherently tied to the quality and breadth of the available data. The reliability and accuracy of machine learning predictions depend on the completeness and diversity of the dataset. Therefore, a more comprehensive dataset, covering a broader range of features and attributes, could potentially enhance the predictive accuracy of the model. This suggests an avenue for future research - refining and expanding the dataset to offer more nuanced and accurate insights into the factors contributing to cybersecurity vulnerabilities.

5.4 Contributions to Body of Knowledge

This research considerably advances the understanding of cybersecurity within the context of higher education by utilizing machine learning methodologies to identify key factors. These factors play a significant role in determining an institution's vulnerability to cyber threats.

By establishing a correlation between school characteristics and security incidents, this research provides valuable insights into the factors that increase an institution's susceptibility to cyber threats. This knowledge serves as a crucial resource for school administrators and policymakers, allowing them to formulate effective preventive measures and mitigation strategies.

In addition, the use of Azure Data Lake for data integration has led to the development of a comprehensive knowledge database. This process entailed the amalgamation of data from disparate sources, using 'School Name' as the key attribute. The application of Azure Data Lake in this context highlights the potential of cloud-based data integration tools in facilitating large-scale data analysis, marking a significant contribution to the domains of data science and machine learning. Moreover, by evaluating the effectiveness of various machine learning models in identifying factors, this research contributes to the enhancement of predictive models and algorithms for potential vulnerabilities. This aspect of the study has implications for the broader field of machine learning, particularly in the context of cybersecurity risk prediction.

In conclusion, the findings of this research provide vital insights for the development of robust cybersecurity strategies in educational institutions, thereby safeguarding sensitive student and faculty data and ensuring continuity in learning. The overarching goal of this research is to strengthen the resilience of the education sector, thereby improving its overall security posture.

5.5 Recommendations for Future Research

This research has provided valuable insights into identifying factors for ransomware and data breaches in higher education institutions, utilizing machine learning

techniques. To build on these findings and contribute to the field of cybersecurity in education, several recommendations for future research are proposed.

Firstly, exploring the capabilities of various machine learning models, including advanced methodologies like deep learning, can enhance our comprehension of cybersecurity threats. A comparative evaluation of the effectiveness and precision of different models could yield more profound insights into the detection and prediction of cybersecurity risks within educational institutions. This investigation should consider the advantages and drawbacks of each model to pinpoint the most suitable strategy for a range of factors and data sets.

Secondly, expanding the dataset to include more diverse and comprehensive features may enhance the accuracy and robustness of the classification model. Future researchers should contemplate integrating additional relevant variables, such as network traffic data, system logs, or user behavior patterns. Enriching the dataset could provide a more comprehensive view of the cybersecurity landscape in educational institutions, thereby improving the predictive capabilities of the models.

Another significant finding of this study is the vulnerability of larger institutions to cybersecurity attacks. Further research is necessary to understand why these institutions may be more susceptible. This information could inform the development of tailored security measures for different educational contexts and assist in determining accurate cyber insurance premiums for educational institutions.

In addition, this study does not pinpoint the causes of data breaches or ransomware attacks. An area for future research could focus on preventative measures

institutions can take by understanding their specific factors. Such understanding would allow institutions to allocate their resources in more intentional ways.

Furthermore, given the potential application of this research in the field of cyber insurance, future investigations could explore the incorporation of the research outcomes into the practices of cyber insurance companies. By leveraging the insights gained from this study, cyber insurance companies can enhance their risk assessment models and underwriting processes for educational institutions. The identification of influential factors, such as 'Institution Size', 'Rank', and 'Research Status', could aid insurers in developing more accurate and tailored insurance policies for educational institutions. The application of data-driven models, particularly the Voting Ensemble model, in the cyber insurance domain could improve insurers' ability to assess an institution's cybersecurity posture and likelihood of experiencing a cyber incident.

In conclusion, these recommendations open up exciting avenues for further research, contributing to the advancement of cybersecurity in higher education institutions. By exploring these areas, researchers can continue to strengthen the resilience of educational institutions against cybersecurity threats, ultimately safeguarding sensitive data and ensuring a secure learning environment.

References

Adam, S. (2021, April 27). The state of ransomware 2021. Sophos News. https://assets.sophos.com/X24WTUEQ/at/pgvqxjrfq4kf7njrncc7b9jp/sophos-state-of-ransomware-education-2022-wp.pdf

Alharbi, T. (2022). Developing a cost-effective cybersecurity management system for academic institutions in Saudi Arabia. Journal of Engineering and Applied Sciences, 9(1), 57. https://doi.org/10.5455/jeas.2022050105

Alharbi, B., & Prince, M. (2020). AHYBRID artificial intelligence approach to predict flight delay. International Journal of Engineering Research and Technology, 13(4), 814. https://doi.org/10.37624/ijert/13.4.2020.814-822

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student's Performance Prediction Using Machine Learning Techniques. Education Sciences, 11(9), 552. https://dx.doi.org/10.3390/educsci11090552

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, *2014*.

Anjum, U. (2020). CYBER CRIME IN PAKISTAN; DETECTION AND PUNISHMENT MECHANISM. Science, Technology and Economic Development, 2(2), 29-38. https://dx.doi.org/10.7251/STED0220029A

Beaudin. (2015). College and university data breaches: regulating higher education cybersecurity under state and federal law. Journal of College and University Law, 41(3), 657–.694

Bischoff, P. (2022, October 12). Ransomware attacks on US schools and colleges cost $3.56bn in 2021. Comparitech. https://www.comparitech.com/blog/information-security/school-ransomware-attacks/

Bongiovanni, I. (2019). The least secure places in the universe? A systematic literature review on information security management in Higher Education. Computers & Security, 86, 350–357. https://doi.org/10.1016/j.cose.2019.07.003

Breiman, L. (2001). Using Iterated Bagging to Debias Regressions. Machine Learning, 45(3), 261–277. https://dx.doi.org/10.1023/A:1017934522171

Campbell, S. (2017, December 22). Cybersecurity in higher education: Problems and solutions. Toptal Insights Blog. https://www.toptal.com/insights/innovation/cybersecurity-in-higher-education

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. https://doi.org/10.1613/jair.953

Check Point. (2022). Cyber-attack Trends Mid-Year Report 2022. Software Technologies LTD. https://securitydelta.nl/media/com_hsd/report/426/document/cyber-security-report-2021.pdf

Check Point Software Technologies Ltd. (2023). Check Point 2023 Security Report

Chen, Y., Li, Y., Li, Y., & Zhang, Y. (2020). Real-time cyber-attack detection using machine learning algorithms. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4225–4236.

Cheng, E. C., & Wang, T. (2022). Institutional strategies for cybersecurity in Higher Education Institutions. Information, 13(4), 192–206.

Cook, S. (2023, April 3). US schools leaked 32 million records in 2,691 data breaches since 2005. Comparitech. https://www.comparitech.com/blog/vpn-privacy/us-schools-data-breaches/

Cybersecurity and Infrastructure Security Agency (CISA). (2018). Insider Threat - Cyber. Retrieved January 26, 2021, from https://www.cisa.gov/insider-threat-cyber

Ding, M., Yang, K., Yeung, D., & Pong, T. (2018). Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses. Proceedings of the 9th International Conference on Learning Analytics & Knowledge, 13. https://dx.doi.org/10.1145/3303772.3303795

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, *8*(4), 1–21. https://doi.org/10.3390/informatics8040079

Gong, C., & Gu, L. (2016). A novel smote-based classification approach to online data imbalance problem. *Mathematical Problems in Engineering*, *2016*, 1–14. https://doi.org/10.1155/2016/5685970

Gupta, N., Tiwari, A., Bukkapatnam, S., & Karri, R. (2020). Additive Manufacturing Cyber-Physical System: Supply Chain Cybersecurity and Risks. IEEE Access. https://dx.doi.org/10.1109/ACCESS.2020.2978815

Feurer, M., Hutter, F. (2019). Hyperparameter Optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds) Automated Machine Learning. The Springer Series on

Challenges in Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-05318-5_1

Fox, B., & Holt, T. J. (2020). Use of a multi-theoretic model to understand and classify juvenile computer hacking behavior. Criminal Justice and Behavior, 48(7), 943–963. https://doi.org/10.1177/0093854820969754

Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W., Choudhary, A. R., & Agrawal, A. (2021). Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nature Communications, 12, 6674. https://dx.doi.org/10.1038/s41467-021-26921-5

Hansen, K. B., & Borch, C. (2022). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, *9*(1), 1–14. https://doi.org/10.1177/20539517211070701

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in public health*, *8*, 357. https://doi.org/10.3389/fpubh.2020.00357

Javed, Y., Khan, A. S., Qahar, A., & Abdullah, J. (2017). Preventing DoS Attacks in IoT Using AES. Journal of Telecommunication, Electronic, and Computer Engineering, 9(3–11), 55–60. https://jtec.utem.edu.my/jtec/article/view/3184/2285

Junqueira, B. S., de Souza, M. V. S., Lima, V. B., de Jesus Gonçalves, A. S. F., & Lepikson, H. A. (2022). Learning Proposal for Cybersecurity for Industrial

Control Systems Based on Problems and Established by a 4.0 Didactic Advanced-Manufacturing-Plant. Journal of Business and Technology, 5(1). https://dx.doi.org/10.34178/jbth.v5i1.188

Kamaludeen, M., Ismaeel, S., Asiri, S., Allen, T., & Scarfo, C. (2021). A Framework for Cyber Protection (FCP) in K-12 Education Sector. 3rd Smart Cities Symposium (SCS 2020). https://doi.org/10.1049/icp.2021.0865

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM:AHighlyEfficientGradientBoosting DecisionTree. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Khader, M., Karam, M., & Fares, H. (2021). Cybersecurity Awareness Framework for Academia. Information, 12(10), 417–437. https://doi.org/10.3390/info12100417

Khan, M. A., & Khan, S. U. (2019). A survey of machine learning techniques for cybersecurity. IEEE Access, 7, 114078-114096.

Kim, J. K., & Kang, S. (2017). Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. Journal of Healthcare Engineering, 2017, 2780501. https://dx.doi.org/10.1155/2017/2780501

Kohavi, R., & Provost, F. (1998). *Machine Learning*, *30*(2/3), 271–274. https://doi.org/10.1023/a:1017181826899

Kulalaieva, N., & Leu, S. (2019). Dual education as a tool for assurance the education of sustainable development. International Journal of Pedagogical Innovation, 7(2), 70-78. https://dx.doi.org/10.5604/01.3001.0013.6847

Laatifi, M., Douzi, S., Bouklouz, A., Ezzine, H., Jaafari, J., Zaid, Y., El Ouahidi, B., & Naciri, M. (2022). Machine learning approaches in COVID-19 severity risk prediction in Morocco. *Journal of Big Data*, *9*(5). https://doi.org/10.1186/s40537-021-00557-0

Liluashvili, G. B. (2021). Cyber risk mitigation in higher education. Law and World, 7(2), 15–27. https://doi.org/10.36475/7.2.2

Liu, C.-W., Huang, P., & Lucas, H. C. (2016). It governance, security outsourcing, and cybersecurity breaches: Evidence from the U.S. higher education. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2850178

Maranga, M. J., & Nelson, M. (2019). Emerging Issues in Cyber Security for Institutions of Higher Education. International Journal of Computer Science and Network, 8(4), 371–379.

Microsoft. (2022). Cyberthreats, viruses, and malware - Microsoft Security Intelligence. Retrieved from https://www.microsoft.com/en-us/wdsi/threats

Moore, M. (2023, February 13). Rise of ransomware attacks on educational institutions. Cyber Management Alliance. https://www.cm-alliance.com/cybersecurity-blog/rise-of-ransomware-attacks-on-educational-institutions

Nauman, M., Akhtar, N., Alhudhaif, A., & Alothaim, A. (2021). Guaranteeing Correctness of Machine Learning Based Decision Making at Higher Educational Institutions. IEEE Access. https://dx.doi.org/10.1109/ACCESS.2021.3088901

Ncube, C., & Garrison, C. P. (2010). Lessons Learned From University Data Breaches. Palmento Business and Economic Review, pp. 13, 27–37.

http://zimmer.fresnostate.edu/~sasanr/Publications/Do%20On-Line%20Students%20Cheat.pdf

Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. IEEE Access. https://dx.doi.org/10.1109/ACCESS.2022.3216617

O'Driscoll, A. (2023, February 12). 30+ data breach statistics and facts: Frequency, impact & more. Comparitech. https://www.comparitech.com/blog/vpn-privacy/data-breach-statistics-facts/#:~:text=According%20to%20statistics%20from%20Surfshark,from%201.4%20million%20in%202020

Ouhmida, A., Raihani, A., Cherradi, B., & Terrada, O. (2021). A Novel Approach for Parkinson's Disease Detection Based on Voice Classification and Features Selection Techniques. International Journal of Online and Biomedical Engineering, 17(10). https://dx.doi.org/10.3991/ijoe.v17i10.24499

Pannakkong, W., Thiwa-Anont, K., Singthong, K., Parthanadee, P., & Buddhakulsomsiri, J. (2022). Hyperparameter tuning of machine learning algorithms using response surface methodology: A case study of Ann, SVM, and DBN. *Mathematical Problems in Engineering*, *2022*, 1–17. https://doi.org/10.1155/2022/8513719

Parker , R. D., & Farkas, C. (2011). Modeling Estimated Risk for Cyber-attacks: Merging Public Health and Cyber Security. Information Assurance and Security Lens, 2, 32–36. https://cse.sc.edu/~farkas/publications/j18.pdf

Pham, D. T., & Afify, A. A. (2005). Machine-learning techniques and their applications in manufacturing. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 219(5), 395–412. https://doi.org/10.1243/095440505x32274

Powell, T. B. (2021, September 8). Howard University cancels classes due to ransomware attack. CBS News. https://www.cbsnews.com/news/howard-university-ransomware-attack-cancel-classes/

Powers, D. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *School of Informatics and Engineering*, 37-63. https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation

Qiao, H., Wang, T., Wang, P., Qiao, S., & Zhang, L. (2018). A Time-Distributed Spatiotemporal Feature Learning Method for Machine Health Monitoring with Multi-Sensor Time Series. Sensors, 18(9), 2932. https://dx.doi.org/10.3390/s18092932

Radzi, S. F., Karim, M. K., Saripan, M. I., Rahman, M. A., Isa, I. N., & Ibahim, M. J. (2021). Hyperparameter tuning and pipeline optimization via grid search method and tree-based AutoML in breast cancer prediction. *Journal of Personalized Medicine*, *11*(10). https://doi.org/10.3390/jpm11100978

Rahman, N. A., Sairi, I. H., Zizi, N. A., & Khalid, F. (2020). The importance of cybersecurity education in school. International Journal of Information and

Education Technology, 10(5), 378–382. https://doi.org/10.18178/ijiet.2020.10.5.1393

Rahman, A., Jabeen, S., Fernandes, G., Banik, G., Islam, J., Ameen, S., Ashrafee, S., Hossain, A., Alam, H. M. S., Majid, T., Saberin, A., Ahmed, A., Ehtesham, K. A. N. M., Chisti, M., Ahmed, S., Khan, M. L., Jackson, T., Dockrell, D., Nair, H., El Arifeen, S., Islam, M. S., & Campbell, H. (2022). Introducing pulse oximetry in routine IMCI services in Bangladesh: A context-driven approach to influence policy and programme through stakeholder engagement. Journal of Global Health, 12. https://dx.doi.org/10.7189/jogh.12.06001

Sarabi, A., Naghizadeh, P., Liu, Y., & Liu, M. (2016). Risky business: Fine-grained data breach prediction using business profiles. Journal of Cybersecurity, 2(1), 15–28. https://doi.org/10.1093/cybsec/tyw004

Schneider, F. B. (2013). Cybersecurity education in Universities. IEEE Security & Privacy, 11(4), 3–4. https://doi.org/10.1109/msp.2013.84

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using Spatial Data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Shen, L., Chen, I., & Su, A. (2017). Cybersecurity and data breaches at schools. Cybersecurity Breaches and Issues Surrounding Online Threat Protection, 144–174. https://doi.org/10.4018/978-1-5225-1941-6.ch007

Siders, Z. A., Ducharme-Barth, N., Carvalho, F., Kobayashi, D., Martin, S. L., Raynor, J., Jones, T. T., & Ahrens, R. (2020). Ensemble Random Forests as a tool for

modeling rare occurrences. Endangered Species Research, 43, 183-198. https://dx.doi.org/10.3354/esr01060

Smith, K.T., Jones, A., Johnson, L. & Smith, L.M. (2019). "Examination of cybercrime and its effects on corporate stock value," Journal of Information, Communication and Ethics in Society, Vol. 17 No. 1, pp. 42–60. https://doi.org/10.1108/JICES-02-2018-0010

Tang, M., Zhao, Q., Ding, S., Wu, H., Li, L., Long, W., & Huang, B. (2020). An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes. Energies, 13(4), 807. https://dx.doi.org/10.3390/en13040807

Tenable. (2022a, January 19). 2021 threat landscape retrospective. Tenable. https://www.tenable.com/cyber-exposure/2021-threat-landscape-retrospective

Wang, J., Zhang, J., & Wang, X.X.. (2018). A Data-Driven Cycle Time Prediction With Feature Selection in a Semiconductor Wafer Fabrication System. IEEE Transactions on Semiconductor Manufacturing. 31. 1-1. 10.1109/TSM.2017.2788501. Uddin et al. (2019)

Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

Ulven, J. B., & Wangen, G. (2021a). A systematic review of cybersecurity risks in Higher Education. Future Internet, 13(2). https://doi.org/10.3390/fi13020039

Verizon. (2021). Data breach investigation report. Verizon Technical Report.

Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of breast cancer using Machine Learning Techniques -A Survey. Procedia Computer Science, 218, 1434–1443. https://doi.org/10.1016/j.procs.2023.01.122

Yaman, M., Subasi, A., & Rattay, F. (2018). Comparison of random subspace and voting ensemble Machine Learning Methods for face recognition. Symmetry, 10(11). https://doi.org/10.3390/sym10110651

Zizi, Y., Jamali-Alaoui, A., El Goumi, B., Oudgou, M., & El Moudden, A. (2021). An optimal model of financial distress prediction: A comparative study between neural networks and logistic regression. Risks, 9(11). https://doi.org/10.3390/risks9110200

Appendix A.

Data Variables

Table 1: Institutional Identification

	Variable	Type	Description
1	Name	String	Name of the college or university
2	Alias	String	Alias of the university
3	Website	String	University's official website
4	City	String	The city where the college is located
5	County	String	County where the university is located
6	State	String	State where the college or university is located
7	Region	String	Region of the United States where the college is located
8	Region code	Integer	Region code where the college is located
9	Latitude	Float	Latitude of the university
10	Longitude	Float	Longitude of the university

Table 2: Institutional Characteristics

	Variable	Type	Description
11	Sector	String	The sector of the college or university
12	Research Status	String	Indicates the research status of the college
13	Iclevel	Integer	Level of the institution based on the highest degree granted
14	Level	Integer	Level of the institution
15	Hi offer	Integer	The highest degree offered by the university
16	Deg Grant	Integer	Degree granting status of the university
17	Locale	Integer	Locale of the college or university
18	Population	Integer	Total population of the college or university
19	Size	Integer	Institutional size of the college
20	Size Set	Integer	Set size of the university
21	Inst Size	Integer	Size of the institution based on student population
22	Main	String	Indicates if the institution is main or branch
23	Classification	String	Classification of the institution
24	Accredagency	String	Accreditation agency of the institution
25	Affiliation	String	Affiliation status of the institution
26	Rank	Integer	Rank of the schools
27	SizeSet	Integer	Size and Location of school combined

Table 3: Enrollment & Staff

	Variable	Type	Description
28	Pt Enroll	Integer	Part-time enrollment
29	Ft Enroll	Integer	Full-time enrollment
30	Tot Enroll	Integer	Total enrollment
31	Staff	Integer	Total number of Employees

Table 4: Campus Facilities

	Variable	Type	Description
32	Housing	Integer	Availability of housing
33	Dorm Cap	Integer	Dormitory capacity
34	Has hospital	Integer	Indicates whether the institution has a hospital

Table 5: Designated Categories

	Variable	Type	Description
35	Land grant	Integer	Indicates whether the institution receives land grant funding
36	Medical	Integer	Indicates whether the institution has a medical school
37	Hbcu	Integer	Indicates whether the institution is a historically Black college
38	Tribal	Integer	Indicates whether the institution is a tribal college or university
39	HSI	Integer	Indicates whether the institution is a Hispanic-serving institution
40	MSI	Integer	Indicates whether the institution is a Minority-serving institution
41	Coplac	Integer	Indicates whether the institution is a member of the Council of Public Liberal Arts Colleges
42	Cusu	Integer	Indicates whether the institution is a member of the Consortium
43	Cumu	Integer	Carnegie Classification of Institutions of Higher Education status
44	National Athletic Association	Integer	Indicates if the institution is part of the National Athletic Association
45	MenOnly	Integer	Indicates whether the institution is a men's college
46	WomenOnly	Integer	Indicates whether the institution is a women's college
47	Multi-Institution Status	Integer	Indicates whether the institution is part of a multi-institution or multi-campus organization

Table 6: Educational Offerings

	Variable	Type	Description
48	Continuing Education	Integer	Indicates if the institution offers continuing education
49	Occupational	Integer	Indicates if the institution offers occupational programs
50	Adult basic remedial	Integer	Indicates if the institution offers adult basic or remedial education
51	Distance education only	Integer	Indicates if the institution only offers distance education
52	Title IV institution	Integer	Indicates if the institution is a Title IV institution

Appendix B

Criteria U.S. News Colleges Ranking Criteria

Ranking Factor	Description	Weightage
Undergraduate Academic Reputation	Based on a survey of college presidents, provosts, and deans of admissions from peer institutions.	22.50%
Graduation and Retention Rates	Measures the percentage of first-time, full-time students who start school and go on to graduate within six years.	20%
Faculty Resources	Measures the number of faculty members, the student-faculty ratio, and the percentage of full-time faculty members.	20%
Student Selectivity	Measures the average test scores and high school class standing of admitted students.	15%
Financial Resources	Measures the amount of money spent per student on instruction, research, and other educational expenses.	10%
Alumni Giving Rate	Measures the percentage of living alums who donate money to the school.	5%

Note: The U.S. News Colleges Ranking Criteria plays a crucial role in assessing the reputation and quality of an institution. The aforementioned factors and their respective weights could potentially influence the susceptibility of an institution to cyber threats, as certain factors such as faculty resources and financial resources might be correlated with the strength of their cyber security infrastructure.

Appendix C

Carnegie Classification Ranking System

Classification	Criteria
R1: Doctoral Universities – Highest Research Activity	1. Award at least 20 research/scholarship doctoral degrees per year across at least three disciplines 2. Have at least $5 million in annual federal research expenditure. 3. Many faculty members engaged in research and have a high level of funding from external sources.
R2: Doctoral Universities – Higher Research Activity	1. Award at least ten research/scholarship doctoral degrees annually across at least two disciplines. 2. Have at least $1 million but less than $5 million in annual federal research expenditures 3. Have many faculty members engaged in research and moderate funding from external sources.
R3: Doctoral Universities – Moderate Research Activity	1. Award at least five research/scholarship doctoral degrees per year across at least two disciplines 2. Have less than $1 million in annual federal research expenditure. 3. Have a moderate number of faculty members engaged in research and moderate research funding from external sources.

***Note**:* The Carnegie Classification Ranking System is a crucial tool in this study, aiding in the categorization of higher education institutions based on their research activities. The extent of research activities, often linked with the institutions' resources, could potentially impact their cyber security infrastructure and hence their vulnerability to cyber threats.

Appendix D.

Inclusion and Exclusion Criteria for Higher Education Institutions

Institution Type	**Count**	**Decision**
Business and Secretarial Schools	29	Excluded
Colleges, Universities, and Professional Schools	2,579	Included
Computer Training Schools	20	Excluded
Cosmetology and Barber Schools	1,178	Excluded
Educational Support Services	71	Excluded
Fine Arts Schools	34	Excluded
Flight Training Schools	13	Excluded
Junior Colleges	1,562	Excluded
Other Technical and Trade Schools	1,073	Excluded

***Note**:* Institutions were considered "included" if they fit within the scope of this research, namely Colleges, Universities, and Professional Schools. The term "excluded" implies that a particular category of educational institutions was considered out of scope for this study due to varying characteristics that differentiate them from the focal group of higher education institutions.

Appendix E.

Machine Learning Results

Metric	Voting Ensemble	Logistic Regression	Random Forest	LightGBM
AUC_macro	0.7932	0.779 (0.014) ↓	0.7786 (0.015) ↓	0.7621 (0.031) ↓
Weighted_accuracy	0.9229	0.9069 (0.016) ↓	0.8457 (0.077) ↓	0.9192 (0.004) ↓
Precision _macro	0.7837	0.7455 (0.038) ↓	0.6949 (0.089) ↓	0.7455 (0.038) ↓
Matthews_correlation	0.3998	0.3851 (0.015) ↓	0.3928 (0.007) ↓	0.2939 (0.106) ↓
Precision _micro	0.8315	0.8235 (0.008) ↓	0.7971 (0.034) ↓	0.8111 (0.020) ↓
F1 _macro	0.668	0.6744 (0.006) ↑	0.6956 (0.028) ↑	0.602 (0.066) ↓
Recall _macro	0.6445	0.653 (0.008) ↑	0.6983 (0.054) ↑	0.5896 (0.055) ↓
Recall _micro	0.8315	0.8235 (0.008) ↓	0.7971 (0.034) ↓	0.8111 (0.020) ↓
F1 _micro	0.8315	0.8235 (0.008) ↓	0.7971 (0.034) ↓	0.8111 (0.020) ↓
Precision _weighted	0.8193	0.8062 (0.013) ↓	0.7997 (0.020) ↓	0.7895 (0.030) ↓
AUC_micro	0.8943	0.8874 (0.007) ↓	0.8636 (0.031) ↓	0.8794 (0.015) ↓
Log_loss	0.4255	0.4485 (0.023) ↑	0.4901 (0.065) ↑	0.4389 (0.013) ↑
Balanced_accuracy	0.6445	0.653 (0.008) ↑	0.6983 (0.054) ↑	0.5896 (0.055) ↓
F1 _weighted	0.8037	0.8026 (0.001) ↓	0.7979 (0.006) ↓	0.7699 (0.034) ↓
AUC_weighted	0.7932	0.779 (0.014) ↓	0.7787 (0.015) ↓	0.7621 (0.031) ↓
Recall _weighted	0.8315	0.8235 (0.008) ↓	0.7971 (0.034) ↓	0.8111 (0.020) ↓
Accuracy	0.8315	0.8235 (0.008) ↓	0.7971 (0.034) ↓	0.8111 (0.020) ↓
Average_precision_macro	0.7528	0.7375 (0.015) ↓	0.7382 (0.015) ↓	0.7104 (0.042) ↓
Average_precision_weighted	0.8531	0.8429 (0.010) ↓	0.842 (0.011) ↓	0.8267 (0.026) ↓
Average_precision_micro	0.8871	0.8787 (0.008) ↓	0.8515 (0.036) ↓	0.8696 (0.018) ↓
Norm_macro_recall	0.2889	0.3059 (0.017) ↑	0.3966 (0.108) ↑	0.1792 (0.110) ↓

Note: An arrow pointing upwards (↑) indicates an increase in the metric value from the baseline (Voting Ensemble), while an arrow pointing downwards (↓) indicates a decrease. The numbers in parentheses represent the absolute change from the baseline.

Appendix F.

Multivariate Logistic Regression Odds Ratio Results

Variable	Odds Ratio	2.5% CI	97.5% CI	p-value
Size	1.84	0.47	0.75	<0.000
Land Grant Institution	0.41	-1.46	-0.32	0.002
Medical degree	0.5	-1.16	-0.22	0.004
Main	7.89	0.63	3.5	0.005
Multi-institution	0.64	-0.76	-0.13	0.006
Occupational	0.65	-0.77	-0.1	0.01
Highest degree offered	0.98	-0.04	0	0.115
MSI	0.8	-0.53	0.09	0.155
Adult basic remedial	0.73	-0.77	0.14	0.18
Continuing professional	1.19	-0.1	0.45	0.203
Urbanization	1.03	-0.02	0.08	0.207
MenOnly	0.29	-3.25	0.78	0.23
National Athletic Association	0.83	-0.52	0.15	0.268
Distance education only	1.64	-0.88	1.87	0.481
Rank	0.97	-0.14	0.07	0.51
Has hospital	0.83	-0.84	0.47	0.583
WomenOnly	1.16	-0.77	1.06	0.752
Title IV institution indicator	$6.54\ X\ 10^{-6}$	-1015.82	991.94	0.981
Research Status	$9.97\ X\ 10^{-11}$	-79675.18	79629.12	0.999
Affiliation	$1.36\ X\ 10^{-10}$	-79674.87	79629.44	0.999

Appendix G.

Univariate Logistic Regression Odds Ratio Results

Variable	Odds Ratio	2.5% CI	97.5% CI	p-value
Size	2.14	0.67	0.85	$2.39 * X\ 10^{-63}$
Medical degree	0.14	-2.31	-1.68	$1.03\ X\ 10^{-34}$
Highest degree offered	0.91	-0.11	-0.08	$8.86\ X\ 10^{-25}$
Rank	1.45	0.29	0.45	$2.80\ X\ 10^{-21}$
Research Status	0.42	-1.06	-0.69	$4.34\ X\ 10^{-20}$
Affiliation	2.38	0.68	1.05	$1.03\ X\ 10^{-19}$
National Athletic Association	0.32	-1.41	-0.88	$2.36\ X\ 10^{-17}$
Land Grant Institution	0.17	-2.23	-1.38	$1.33\ X\ 10^{-16}$
Has hospital	0.21	-2.03	-1.13	$5.84\ X\ 10^{-12}$
Continuing professional	2.02	0.49	0.92	$6.38\ X\ 10^{-11}$
Main	18.84	1.54	4.34	$3.96\ X\ 10^{-5}$
Urbanization	1.07	0.03	0.11	0.00029
Occupational	0.75	-0.53	-0.04	0.023
MenOnly	0.12	-4.08	-0.09	0.04
Multi-institution	1.26	0.01	0.45	0.041
Distance education only	2.04	-0.49	1.92	0.246
Adult basic remedial	0.84	-0.51	0.16	0.31
MSI	1.1	-0.16	0.35	0.453
WomenOnly	0.84	-1.07	0.71	0.695
Title IV institution indicator	$6.25\ X\ 10^{-8}$	-7622.48	7589.3	0.997