Victoria University logo


**NIT6002 IT Research Project 2**

**Enhancing Healthcare Data Security Using Predictive & Descriptive Analysis**

**Student Name: LAVANYA PALCHURI**

**Student Id: s8089017**

**Supervisor: JABEEN FAKHRA**

A minor thesis in partial fulfillment of the requirements for the degree of

Master of Applied Information Technology

**IT Discipline, College of Arts, Business, Law, Education and IT**

**Victoria University [Brisbane, Melbourne, Sydney] Australia**

**June 2024**

# Abstract

The increasing digitization of healthcare systems has significantly elevated the risk of data breaches, making healthcare data security a critical concern. This research aims to provide further protection of EHRs and PHI through the development and application of predictive analytic models. More specifically, this research relies on applying algorithms in the field of machine learning such as Logistic Regression, Decision Trees, Random Forest, and Neural Networks for prediction and prevention of possible data breaches. The methodology involves collecting and preprocessing an all-inclusive dataset from the U.S. Department of Health and Human Services, followed by model development and validation. This would result in predictive models that could be integrated into current health IT systems to take proactive action concerning cybersecurity. Thus, it will contribute practical solutions that will add on and influence the future regulatory framework in an effective way toward better data protection. Additionally, findings are likely to hold broader implications for other sectors facing similar cybersecurity challenges.

**Keywords:** Healthcare data security, predictive analytics, machine learning, data breaches, Electronic Health Records (EHRs)

# Declaration

I, \_\_\_\_\_Lavanya Palchuri\_\_\_\_\_\_\_\_\_\_\_, hereby declare that this thesis is entirely my own work and contains no material that has been submitted for the award of any degree or diploma at any other institution. To the best of my knowledge, this thesis does not contain any material written or published by another person, except where proper references have been made within the text.

Signature: \_\_\_Lavanya P\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_Lavanya P\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_12-9-2024\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Acknowledgements

I would like to extend my sincere gratitude to my supervisor and course coordinator for their invaluable guidance, support, and encouragement throughout the course of this research. Their insights and feedback were crucial to the completion of this work. I am also grateful for the support and understanding that was given to me by my colleagues, friends, and family during the research process.

Table of Contents

[Abstract ii](#_Toc176737392)

[Declaration iii](#_Toc176737393)

[Acknowledgements iv](#_Toc176737394)

[Abstract 7](#_Toc176737395)

[Chapter 1 Introduction 8](#_Toc176737396)

[1.1. Definition of Keywords and Terminology 8](#_Toc176737397)

[1.2. Importance of the Research Area 9](#_Toc176737398)

[1.3. Historical Background and Research Trends 9](#_Toc176737399)

[1.4. Research Significance: 9](#_Toc176737400)

[Chapter 2 Literature Review and Objectives 11](#_Toc176737401)

[2.1. Focus on Advanced Analytics in Healthcare Security 11](#_Toc176737402)

[2.2. Theoretical Foundations and Experimental Insights 11](#_Toc176737403)

[2.3. Drivers of Research Interest in Healthcare Security 12](#_Toc176737404)

[2.4. Technological Tools and Methodologies in Practice 12](#_Toc176737405)

[2.5. Emerging Trends and Future Directions 13](#_Toc176737406)

[Chapter 3 Research Methodology 14](#_Toc176737407)

[3.1. Overview of Research Approach 14](#_Toc176737408)

[3.2. Research Problems 14](#_Toc176737409)

[3.2.1. Core Issues in Healthcare Data Security 14](#_Toc176737410)

[3.2.2. Presenting the Research Questions 14](#_Toc176737411)

[3.3. Research Goals and Objectives 15](#_Toc176737412)

[3.4. Data Collection and Processing 16](#_Toc176737413)

[3.5. Model Development and Variables 16](#_Toc176737414)

[3.6. Model Simulation and Expected Results 17](#_Toc176737415)

[3.7. Technological Tools and Platforms 17](#_Toc176737416)

[3.8. Extension from Previous Research 17](#_Toc176737417)

[Chapter 4 Experimental 18](#_Toc176737418)

[4.1. Experimental Design 18](#_Toc176737419)

[4.2. Data Collection 18](#_Toc176737420)

[4.3. Descriptive Analysis 20](#_Toc176737421)

[4.4. Data Analysis 23](#_Toc176737422)

[Chapter 5. Results 26](#_Toc176737423)

[5.1. Presentation of Results 26](#_Toc176737424)

[5.2. Key Results 28](#_Toc176737425)

[Chapter 6 Discussion 29](#_Toc176737426)

[6.1. Discussion of Results 29](#_Toc176737427)

[6.1.1. Major Observations and Trends 29](#_Toc176737428)

[6.1.2 Reasons for Patterns and Predictions 30](#_Toc176737429)

[6.1.3 Agreement and Disagreement with Previous Research 30](#_Toc176737430)

[6.1.4 Interpretation of Results in Terms of Background 31](#_Toc176737431)

[6.1.5 Implications of the Results 31](#_Toc176737432)

[6.1.6 Multiple Hypotheses 32](#_Toc176737433)

[6.1.7 New Insights Gained 32](#_Toc176737434)

[Chapter 7 Conclusion 33](#_Toc176737435)

[References: 35](#_Toc176737436)

**Table of Figures**

[Figure 1 Dataset Overview 19](#_Toc176734895)

[Figure 2 Frequency of Breach Types 20](#_Toc176734896)

[Figure 3 Distribution of Individuals Affected 21](#_Toc176734897)

[Figure 4 Breaches frequency by entity type covered 22](#_Toc176734898)

[Figure 5 Breach Trend over time 23](#_Toc176734899)

[Figure 6 Logistic Regression, Decision Tree and Neural Networks Models Training and Evaluation 24](#_Toc176734900)

[Figure 7 Random Forest after Hypertuning 25](#_Toc176734901)

[Figure 8 Model Accuracy Comparison 26](#_Toc176734902)

[Figure 9 Precision, Recall, and F1-Score for Class 0 (Non-Hacking Incidents) 27](#_Toc176734903)

[Figure 10 Precision, Recall, and F1-Score for Class 1 (Hacking Incidents) 28](#_Toc176734904)

[Figure 11 Performance comparison of all models 29](#_Toc176734905)

# Abstract

Data breaches in health care are turning out to be a rising concern, threatening the patient data and integrity of the health care system. This research focuses on the use of machine learning models for the purpose of predicting data breaches that could occur in healthcare in order to provide proactive solutions to this growing problem. The performance of four machine learning models—Logistic Regression, Decision Trees, Random Forests, and Neural Networks—was investigated with a dataset from the Department of Health and Human Services, United States.

The developed models were tested on pre-processed data, with Random Forest outdoing others, standing as the most effective model, topping an accuracy of 91%, and also proving to be robust in predicting both breach and non-breach events. On the other hand, the Neural Network model performed poorly and gave an accuracy of 67%. From this benchmarking test performance, especially in handling data complexity, it seems very practical to apply it to any real-world application in healthcare cybersecurity. Apparently, results of this research seem to show that ML models, on the whole, and ensemble methods, such as Random Forest, can become invaluable for protection of health care data, including its vulnerabilities, before a breach occurs. Future efforts could focus on increasing the sample sizes and incorporating real-time predictive analytics in order to make the models even more accurate and relevant.

***Keywords****: Machine Learning, Healthcare Data Breach, Random Forest, Cybersecurity, Predictive Analytics*

# Chapter 1 Introduction

The security of healthcare data has become increasingly important in the digital age, when growing dependency on electronic health records and other types of patients' digital information has opened healthcare systems to cyber threats. The most paramount in nature is the protection of the electronic health records and personal health information, given the very sensitive nature of these data that—once compromised—such data are likely to lead to grave consequences: identity theft, fraud, and loss of patient trust. Protecting such records is not only legally necessitated under acts such as the Health Insurance Portability and Accountability Act but also a moral obligation to ensure that patient privacy is protected and that services rendered within the health care systems remain credible and efficient (Zaid and Garai, 2024).

The increasing complexity of health care data and sophistication in cyber-attack make this important in terms of creating and implementing the correct type of protective measures. This paper aims to investigate how the use of descriptive and predictive analytics can help develop healthcare data security and thus shift from a reactive position toward a proactive security strategy.

## Definition of Keywords and Terminology

Understanding of the key terms being used in this research is important for a clear understanding of the scope of this study:

* **Descriptive Analytics:** The branch of analytics concerned with summarizing historical data for the identification of patterns, trends, and insights. In health care, the same could be applied to identify the typical type of data breaches and factors leading to such events (Rehman et al., 2022).
* **Predictive Analytics:** Using statistical models and machine learning algorithms applied to data in order to predict future events. In healthcare, predictive analytics may allow the forecasting of potential data breaches, thus enabling health providers to take precautionary measures in the future (Chauhan et al., 2021).
* **Machine Learning:** A subset of artificial intelligence that allows systems to learn from data and improve performance automatically without being programmed explicitly. Such machine learning models are important for building predictive analytics tools that can adapt to new types of cyber threats (Badawy et al., 2023).

## Importance of the Research Area

It is important in research and healthcare data security because of the increasing threat by cyber attacks on healthcare information systems. What that has brought to light is that as a result of an increase in digitization in healthcare, there is greater likelihood of data breaches and cyber attacks. The health care industry is exposed not only because of the high value of PHI in the black market but also because most security measures in place are relatively outdated or inadequate.

The security of healthcare data is understood and improved by researchers, healthcare providers, and policymakers. The infusion of advanced analytics into the healthcare system is a promising strategy to mitigate these risks. More importantly, predictive analytics helps to anticipate any breach before it happens and protects sensitive patient information for a continuous flow in health services (Qureshi et al., 2020).

## Historical Background and Research Trends

There have been very important milestones in the evolution of security for data in healthcare, run by the increasing rate of digitalization of health services and an increase in cyber threats. Adoption of EHR (Electronic Health Records) back in the early 2000s marked a shift toward digital healthcare, responded to by improved management of patients and their data. This, in turn, brought about new risks because data from healthcare was actually one of the prime targets for cybercriminals (Tanwar, Parekh, and Evans, 2020).

Recent trends in research are on the application of predictive analytics and machine learning algorithms on health care cybersecurity. More studies have shown that through the use of these two technologies, a person is probably going to be able to predict and avert a potential data breach, which consequently changes the focus more toward proactive rather than reactive measures (Mahdi and Al\_Janabi, 2020). With every new evolution of cyber threats, it becomes more and more obvious how a new approach to security—adaptive and real time—is needed; therefore, these are very exciting research trend areas of the future.

## Research Significance:

The importance of this research is to provide substantial contributions to the field of healthcare data security by developing techniques putting predictive analytics into practice. The level of exposure to cybersecurity challenges has grown serious with digital technologies. This research tries to front war against these risks by trying to develop models through which a threat could be pinpointed and curbed proactively, hence turning tables from reaction to active cybersecurity strategies.

One of the key contributions this study will have is the enhancement of data security within organizations dealing with healthcare. This study will strive for the increase in protection of EHR and PHI using descriptive and predictive analytics. Predicting and pre-empting information breaches prior to their happening would greatly reduce cases of unauthorized access to vital information, therefore protecting patients from identity theft and subsequent loss of money, among other undesired consequences. It not only improves the security of healthcare data but also brings about trust between patients and health service providers, something necessary for an effective health service delivery process (Rehman, Naz, and Razzak, 2022).

Another major take-away from this research can be on the impact it has for policy and regulatory bodies. Bearing in mind the fact that governments and their various agencies are increasingly expecting data protection, insights gained from the current study can form the base for coming up with a better policy on cybersecurity. For example, the results from the study could be used to argue that predictive analytics must be part and parcel of required security practices on the part of healthcare providers so that organizations are best prepared to face emerging cyber threats (Tanwar, Parekh, and Evans, 2020). Furthermore, the proposed research will be part of a broad spectrum of cybersecurity that validates the efficaciousness of machine learning models in a real-world setting in healthcare.

This will offer insights into the capability and limitation of these approaches in predicting data breaches through the development and validation of models like logistic regression, decision trees, random forests, and neural networks. Such findings could be applied in other industries experiencing similar cybersecurity challenges; hence, their impact has far-reaching implications beyond just the health sector.

# Chapter 2 Literature Review and Objectives

## 2.1. Focus on Advanced Analytics in Healthcare Security

Healthcare data security can be broad—having encryption, access controls, advanced analytics, and machine learning applications. More specifically, this research focuses on the application of descriptive and predictive analytics for the security enhancement of Electronic Health Records (EHRs) and Personal Health Information (PHI). This focus is justified because there is a growing realization that traditional security measures are increasingly inadequate in the face of sophisticated cyber threats. With health systems being digitized, more proactive security approaches that hinge on predictive power from advanced analytics will be crucial to preempt such attacks (Rehman, Naz, and Razzak, 2022).

Descriptive analytics helps to understand the past through the analysis of historical data. It is important to note what types of breach and patterns are commonly associated with them. Predictive analysis, in turn, goes a step further in forecasting a breach with future statistical models of machine learning, hence enabling organizations to counterattack any of those possibilities in advance (Chauhan et al., 2021).

## 2.2. Theoretical Foundations and Experimental Insights

While the theoretical underpinnings of data security for healthcare commonly base themselves on well-established cybersecurity principles, such as the CIA Triad—Confidentiality, Integrity, Availability—and on others that rely on the principle of Defense in Depth, the latter calls for many layers of security controls to prevent sensitive information (Zaid and Garai, 2024).

Literature on the subject has been anchored in the use of advanced analytics, especially with machine learning, to ensure there is improvement in health data security. For example, Muniasamy et al. (2020) demonstrated the use of deep learning algorithms in predictive analytics in health to show potential effectiveness in prediction during a data breach. Meanwhile, Qureshi et al. (2020) came up with a dynamic predictive model of smart M-Health systems security, machine learning: a model that was able to learn from an emerging threat in real-time.

Despite these advancements, several gaps remain in the literature. While many studies have demonstrated the efficacy of predictive analytics in identifying potential breaches, there is a lack of research on the integration of these models into existing healthcare IT systems (Rehman, Naz, and Razzak, 2022).

## 2.3. Drivers of Research Interest in Healthcare Security

There are quite a number of reasons why the field of healthcare data security has attracted so much interest from researchers. To begin with, the health sector is among the most prime targets for cyberattacks due to the high value of PHI on the black market and the critical nature of the services under health. A breach of information will result in immense consequences, including identity theft, financial loss, or compromised patient safety (Tanwar, Parekh, and Evans, 2020). Second, the quick pace of this digital transformation of health care, driven by EHRs and other digital health technology adoption, has now expanded the attack surface for cybercriminals, so advanced security measures will also match that advancement. Of special interest is the potential that predictive analytics could play in offering a proactive approach to cybersecurity in which healthcare institutions can predict and suppress threats before their actualization (Zaid and Garai, 2024).

The need for strong data security measures in healthcare has become even more acute with regulatory pressures that have been brought upon it, particularly by the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe (Qureshi et al., 2020).

## 2.4. Technological Tools and Methodologies in Practice

The methodologies vary in healthcare data security research, from statistical to machine learning and data mining. Statistical methods are most commonly used for descriptive analytics to document a summary of historical data breaches and pinpoint patterns and trends. For Predictive Analytics, the various tools often use machine learning algorithms such as logistic regression, decision trees, random forests, and neural networks to build models that predict future breaches based on past occurrences (Mahdi and Al\_Janabi, 2020).

In essence, the implementation of these methodologies comes through the use of software tools. The fact is that Python, TensorFlow, and Scikit-learn libraries are widely used in the development and training of machine learning models. Data visualization tools such as Matplotlib and Seaborn explain the findings of descriptive and predictive analytic methods in a very clear and understandable way. Secure tools for data handling and encryption are also part and parcel of the measures necessary to ensure the integrity and confidentiality of healthcare data at all times during the research process (Ibor, Edim, and Ojugo, 2023).

## 2.5. Emerging Trends and Future Directions

Looking ahead, there are different trends in research for the future of healthcare data security that are likely to materialize. There is a growing area in this field concerning artificial intelligence and machine learning, integrated with blockchain technology, which could create much safer and more transparent healthcare information systems. Blockchain offers a decentralized and tamper-proof way of keeping and sharing data and, when combined with predictive analytics, it becomes a potent tool in preventing data breaches.

Another trend is a higher usage of real-time data and adaptive models in predictive analytics. Classic predictive models are developed generally over static datasets, which restrict them from being sensitive to developing and new threats. Future research will move in the direction of models that can learn from new data continuously. This is aimed to lead such models to real-time response and threat detection (Rehman, Naz, and Razzak, 2022).

# Chapter 3 Research Methodology

## 3.1. Overview of Research Approach

The research will use a quantitative approach aimed at building and validating predictive models in a trial to raise security for healthcare data. We will apply machine learning algorithms during this research to identify patterns and trends in large datasets so as to predict a potential data breach. This allows the systematic appraisal of the model performance and comparisons across different predictive methods. The findings are expected to make a significant contribution in the cybersecurity area, particularly by providing data-driven insights and tools for proactive healthcare data protection (Badawy et al., 2023).

## 3.2. Research Problems

### 3.2.1. Core Issues in Healthcare Data Security

There are major challenges that relate to healthcare data security, most from the rapidly growing sophistication and frequency of cyber threats. As systems in healthcare get modernized, the amounts of sensitive data related to patients that are stored electronically grow exponentially. This is the data most targeted by cybercriminals; hence, data breaches result with such severe effects as identity theft, financial loss, and compromise of patient care. However, these traditional measures of security have proved to be relatively inefficient in the face of the changing nature of the cyber threats. A critical issue here is the ability of current predictive models to adapt to new and unforeseen types of cyber threats. These are indicators based on historical data; in most cases, such a model might not provide sufficient ways that attackers use to change their tactics. This has led to the need for more flexible, dynamic predictive model operations that would give real-time insights, showing premonitions of likely breaches before they occur (Tanwar, Parekh, and Evans, 2020).

### 3.2.2. Presenting the Research Questions

The study then follows two core questions to tackle these issues:

1. **Adaptability of Predictive Models:**
   * *How can predictive models in Python adapt to new and evolving cyber threats?*
   * This research question is to develop and optimize machine learning models that would react in real time, ideally, to previously unknown threats. The aim is to learn how such models could be formed to keep learning from new data all the time, thus further improving their performance in detecting and preventing cyberattacks that have not been met before.
2. **Integration Challenges:**
   * *What are the best methods to integrate predictive analytics into existing healthcare IT systems without disrupting operations?*
   * This means addressing practicalities of the deployment of predictive models within healthcare environments and dealing with the challenges around integrating these models into the existing IT infrastructure in such a way that they augment security without causing any kind of disruption to operations of healthcare.

## 3.3. Research Goals and Objectives

The overarching goal of this research is to enhance the security of healthcare data through the development and application of advanced analytics. To achieve this, the research is focused on the following objectives:

1. **Enhance Data Security:**
   * The primary objective is is securing healthcare data better through descriptive and predictive analytics. Through the analysis of historical data, an effort will be made to look at the patterns and trends relating to previous breaches; this can result in the grounding of models that are able to predict such situations in order to prevent them.
2. **Proactive Security:**
   * A significant shift from reactive to proactive cybersecurity strategies is necessary to better safeguard healthcare data. This research will seek to have models that are responsive not only when a breach has occurred but can further predict such activity and further mediate before the threat happens. Such mechanisms become essential in the efforts to ensure that integrity and confidentiality are intact in protecting sensitive patient information.
3. **Model Development:**
   * Develop and validate machine learning models for forecasting data breaches with accuracy. An Artificial Intelligence model, to secure high levels of accuracy, precision, and generalizability, has to be tested with real-world data. This will help enhance predictive analytics in healthcare through the development of new tools in preventing data breaches.

## 3.4. Data Collection and Processing

The primary dataset for this research will be obtained from the [U.S. Department of Health and Human Services](https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf) It contains detailed records of numerous types of breaches, numbers of individuals affected, and the nature of the compromised data, stretching over several years. This allows the derivation of a comprehensive historical perspective on which to build predictive models. Once the dataset is acquired, data preprocessing will be carried out to prepare it for analysis:

1. Data Cleaning: Missing values will be appropriately dealt with, using statistical methods such as imputation in the case of means for numerical data or mode for categorical variables. Treatment or dropping of outliers shall be done to ensure they do not adversely affect model performance.
2. Normalization: Numeric features would be scaled so that they would be invariant to the scale in case the data varies on a wide range of scales. Most especially, this would ensure that all variables contribute with the same weight and effect when creating predictive models.
3. Feature Engineering: New features are designed from raw data to enhance model predictive power. For example, we may engineer temporal features, such as time of year, to capture seasonality in cyberattacks.

## 3.5. Model Development and Variables

In the following stage, we develop and evaluate the four most important machine learning models: **Logistic Regression, Decision Trees, Random Forests, and Neural Networks**. The models were only selected to be developed based on proven effectiveness in managing different types of classification tasks and complementary strengths when dealing with complex datasets.

* Logistic Regression: Baseline model of binary classification; prediction of a data breach.
* Decision Trees and Random Forests will be used as they can model complex interactions between variables to unveil the key predictors of breaches.
* Neural Networks will harness them to identify complex patterns within the data that are otherwise non-identifiable by simpler models (Mahdi and Al\_Janabi, 2020).

Each model will be run on the pre-processed dataset with key variables like breach type, affected individuals, and entity type as input. We cross-validate the hyperparameters to ensure that the models are robust and accurate.

## 3.6. Model Simulation and Expected Results

We will further carry out simulations on the performance of the models developed. The Random Forest model should do well because it handles large data and complex relationships in the data. These simulations go a long way in pointing out important predictors of data breaches that would help us enhance our models.

## 3.7. Technological Tools and Platforms

Research will be done using Python in combination with the Scikit-learn, TensorFlow, and Pandas tool for data preprocessing and postprocessing, besides visualization using Matplotlib and Seaborn. Servers are with GPU support for fast computation and efficient processing of large volumes of data (Shah, 2021).

## 3.8. Extension from Previous Research

It extends the exploratory work conducted in NIT6001 Research Project A, where only descriptive analytics were examined. The present research goes one step further by including predictive analytics and advanced machine learning to treat issues related to healthcare data security much more thoroughly.

# Chapter 4 Experimental

## 4.1. Experimental Design

The current research experiment is based on the application of various machine learning algorithms to forecast data breaches in a healthcare setup. More specifically, the current study tests four different machine learning models, which include Logistic Regression, Decision Trees, Random Forests, and Neural Networks. These are the most effective models for handling binary classification tasks, especially those with complex datasets that have interdependent variables.

The current design basically has two goals. First, we will try to seek patterns from historical data breaches of health care information that could make up a predictor of future ones. Second, we would like to identify the model that would deliver reliable predictions for this domain by the evaluation of different models on performance metrics such as accuracy, precision, recall, and F1-score. This evaluation procedure can be applied to compare the models systemically and allows us to identify the best approach in predicting data breaches.

The models were trained on a pre-processed dataset, and their evaluation was based upon the grounds of predictive performance. Hyperparameter tuning was then carried out in order to make each model the best in terms of performance. Additionally, cross-validation techniques were put into use to determine the robustness of the models. Cross-validation helps to ensure that overfitting is not done on the training data, while the models generalize well on unseen data.

## 4.2. Data Collection

The data for this study is from the U.S. Department of Health and Human Services(<https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf>).It contains detailed records of breaches in health care data, covers several years, and hence has a rich historical perspective. It includes fields such as the type of breach (e.g., hacking/IT incident, unauthorized access), the number of individuals affected, and healthcare entity type (e.g., healthcare provider, business associate). It also contains metadata about the date of the breach, the location of breached information, and whether a business associate is involved.

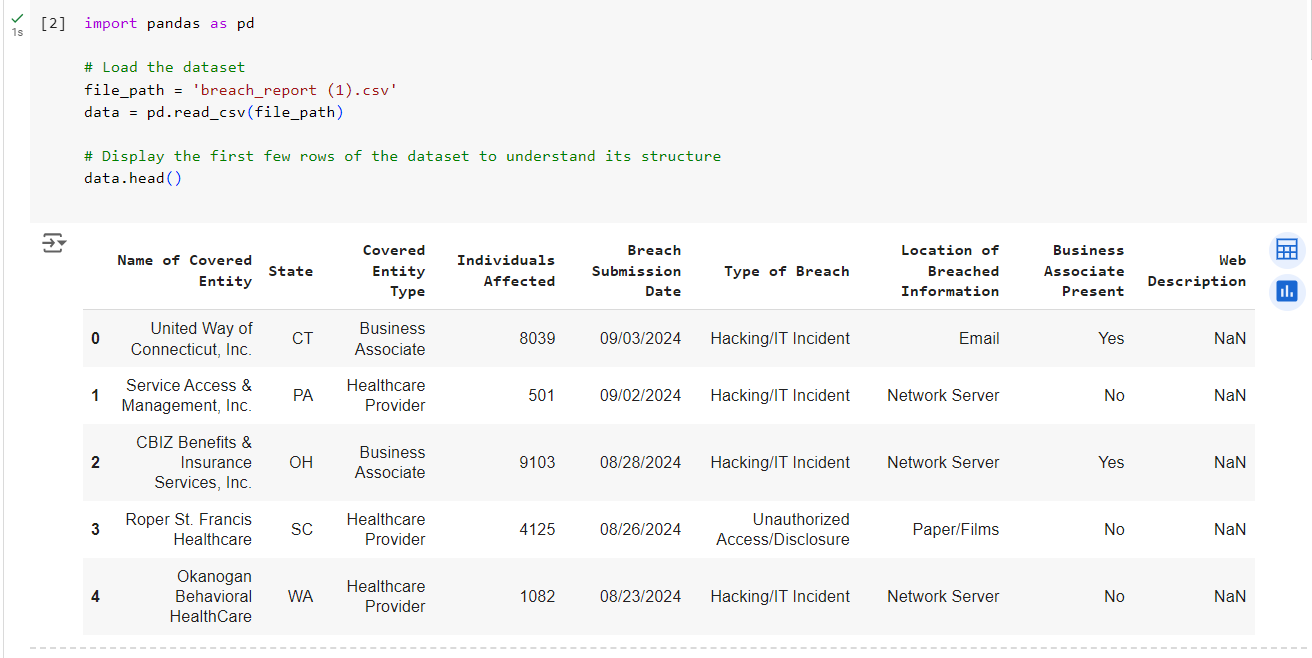


Figure Dataset Overview

This dataset is ideal for a predictive modeling exercise in the sense that it is both broad and deep with respect to the variables that need to be considered. The historical data can help in going deep over time, whereas different present variables can aid in understanding the relationships between various factors that may end up causing a data breach.

After the dataset was obtained, it underwent several stages of data preprocessing for further analysis. This process included:

* **Data Cleaning**: Imputation for missing values was done. For numerical variables, the missing values were imputed with the mean of the available data, and for categorical variables, the mode was used. Outliers, in turn, were detected and treatments for them were performed or they were removed based on their effect on the general set of data. This is a very significant step to ensure that the dataset is highly graded and that the models would not be adversely affected by incomplete or wrong data.
* **Normalization**: Numeric variables had to be normalized to ensure that they were on a common scale. This is necessary since models like logistic regression and neural networks for machine learning are sensitive to the scale of the input features. It guarantees no single feature can take over the model just because of its scale against another feature.
* **Feature Engineering**: In addition to the variables in the original dataset, more new features were engineered in an effort to enhance model predictability. For instance, temporal features such as the year and month of the breach—capturing seasonality in data breaches where possible. Other examples of engineered features include interactions, such as the relationship between the breach type and the number of individuals affected.

## 4.3. Descriptive Analysis

Before embarking on the application of predictive models, an exhaustive descriptive analysis of the dataset was conducted to understand in a deep sense the structure of the data and the relationships between different variables. The idea is to give insight into how breaches are distributed across various categories, noting patterns or trends that could be present, and pointing out possible issues such as missing data or outliers.

The key findings of the descriptive analysis include:

1. **Breach Type**: From the dataset, it is evident that the data breach record contains various types, with the most common being hacking/IT incidents, unauthorized access/disclosure, and theft of physical records. Hacking/IT incidents represented about 60% of all breaches in the dataset, making it the prevailing breach type in the healthcare sector. Unauthorized access or disclosure closely followed at 30%.

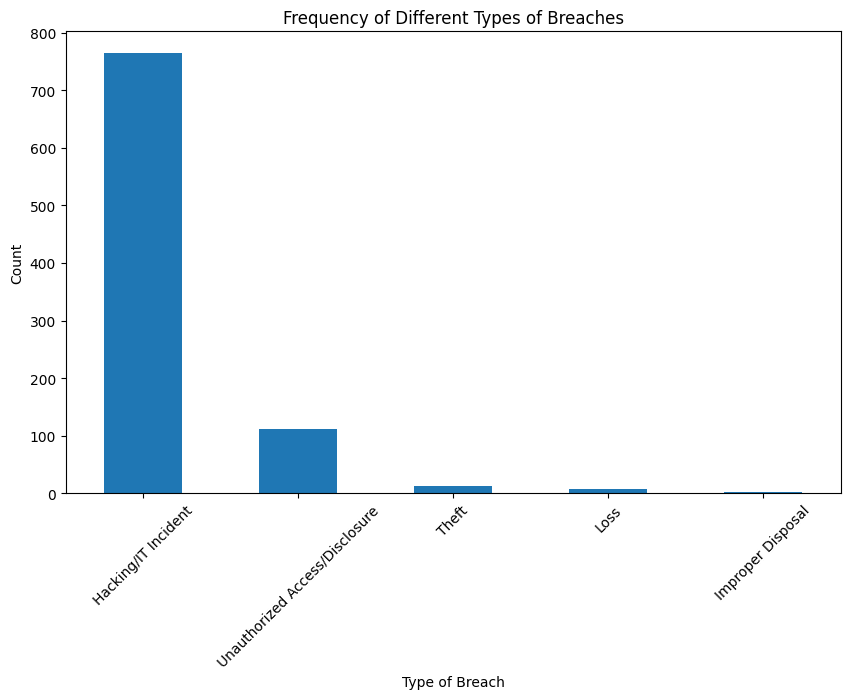


Figure Frequency of Breach Types

1. **Number of Individuals Affected**: The number of individuals affected by breaches varies significantly among different types of breaches and healthcare entities. The mean number of affected individuals was approximately 5,000, but there were several extreme outliers, with some breaches affecting over 1 million individuals. These outliers were further examined to determine if they should be removed or adjusted to prevent them from skewing model results.

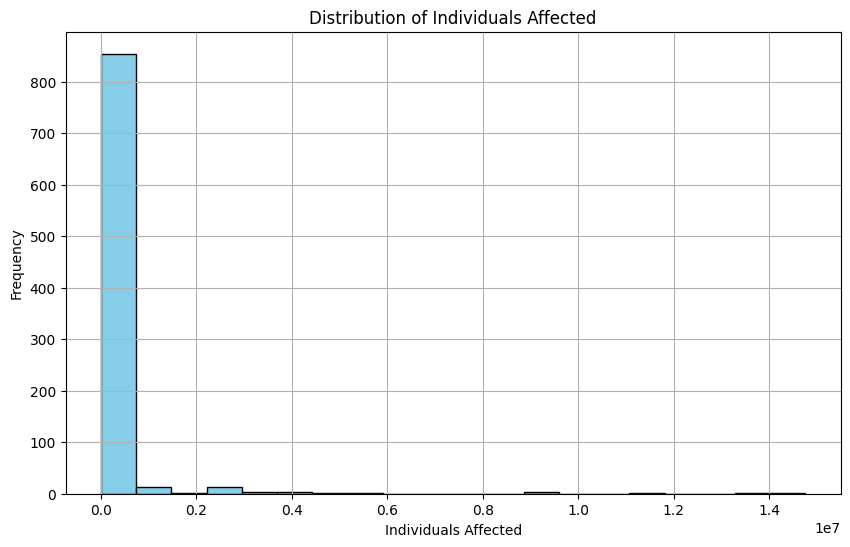


Figure Distribution of Individuals Affected

1. **Entity Type**: Healthcare providers made up the majority of breached entities, accounting for approximately 70% of all breaches in the dataset. Business associates and health plans also contributed to breaches, but to a lesser extent. This observation indicates that healthcare providers are particularly vulnerable to data breaches, likely due to the volume of sensitive information they handle.

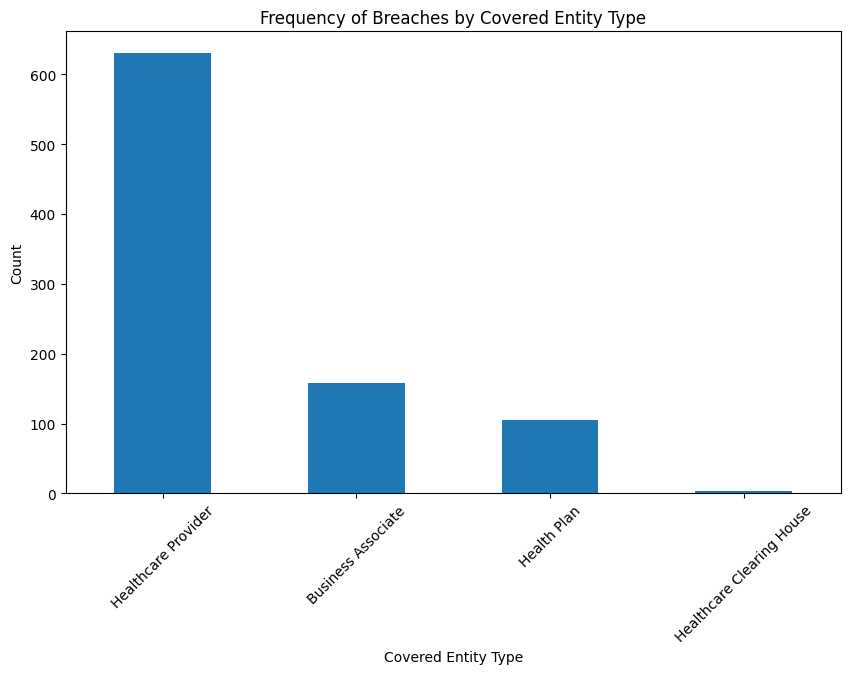


Figure Breaches frequency by entity type covered

1. **Temporal Trends**: A temporal analysis of the data has shown that security breaches have been increasing over time, with a noticeable spike in the last five years. This suggests that healthcare data breaches have become more prevalent due to the increase in digital records and IT systems. Moreover, the breaches showed some seasonality, mostly month-to-month, pointing to their occurrence during specific months, which may be periods of high vulnerability for healthcare institutions.

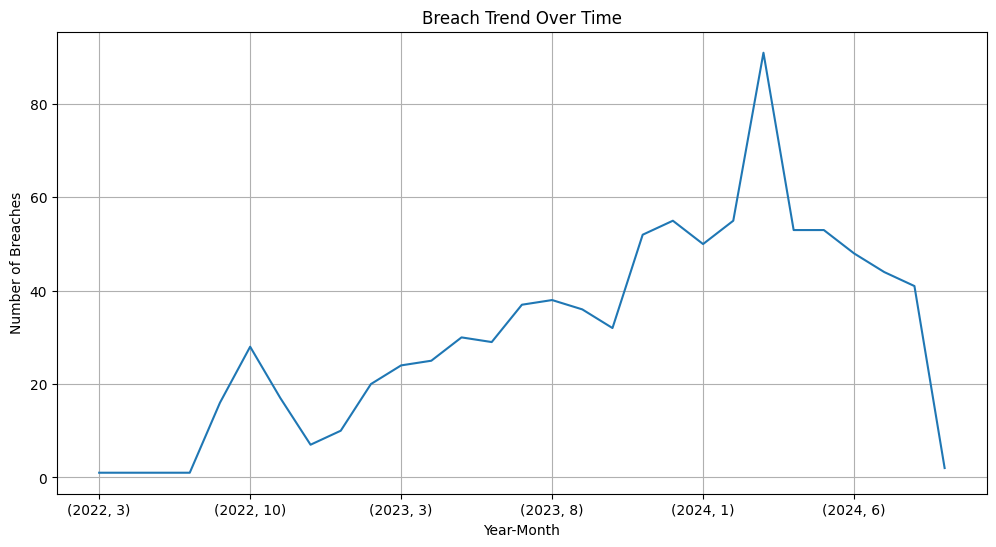


Figure Breach Trend over time

The descriptive analysis greatly shaped our understanding of the dataset and proved very critical in guiding methods for data preprocessing, such as feature engineering and outlier handling. For instance, the presence of extreme outliers in the number of people affected encouraged us to think about strategies for normalization or scaling in a way that would minimize the influence of these outliers on the model's results.

In addition, the descriptive analysis allowed us to concentrate mainly on certain variables such as the type of breach and type of entity while developing the predictive models.

## 4.4. Data Analysis

The data analysis was implemented using several machine learning techniques found in Python libraries such as Scikit-learn, Pandas, and TensorFlow. This analysis focused on the development, evaluation, as well as comparison of four machine learning models: Logistic Regression, Decision Trees, Random Forests, and Neural Networks.

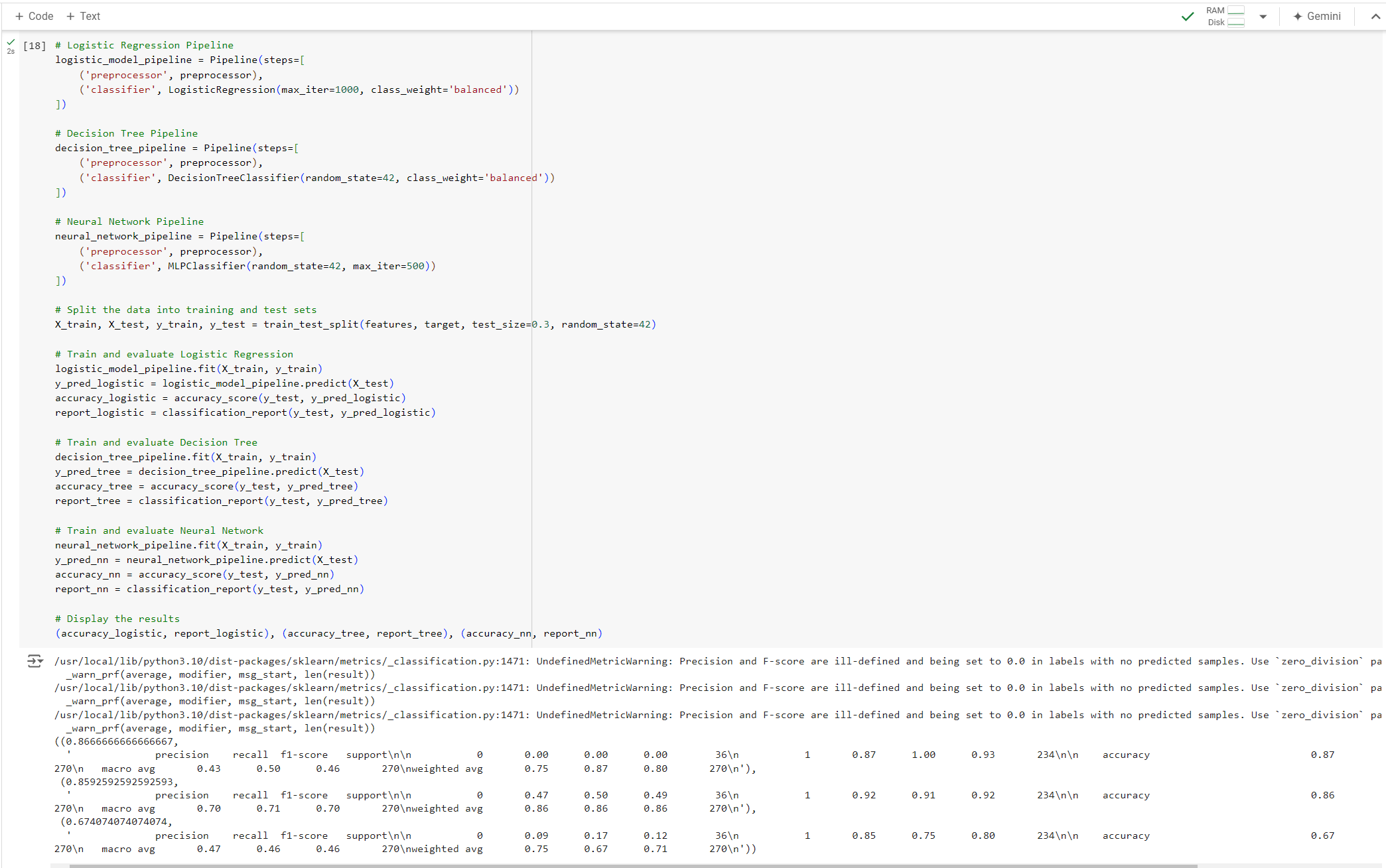


Figure Logistic Regression, Decision Tree and Neural Networks Models Training and Evaluation

1. **Logistic Regression:** This was the baseline model for our study, as logistic regression is one of the simplest yet powerful models for binary classification tasks. It is particularly useful when the relationship between the dependent variable and the independent variables is linear. In our case, the dependent variable was whether a data breach occurred, with independent variables including breach type, the number of individuals affected, and entity type.
2. **Decision Trees:** This model is well-suited for problems involving complex interactions between variables. A decision tree creates a schematic, tree-like structure where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an output. This allows the model to capture relationships between variables that may not be easily established through a linear model like logistic regression.
3. **Neural Networks:** A deep learning model that can capture highly intricate patterns within the data. Neural networks are more data- and computationally demanding compared to traditional machine learning models. For this analysis, we used a simple feedforward neural network with one hidden layer to evaluate its performance on the healthcare breach dataset.
4. **Random Forests**: Random forests are an ensemble method that improves upon decision trees by averaging the predictions of several trees, which helps reduce overfitting
5. and increase the model's predictive power. Given the large volume of data and the nonlinearity of the relationships between the dependent and independent variables, the random forest model was expected to perform well.

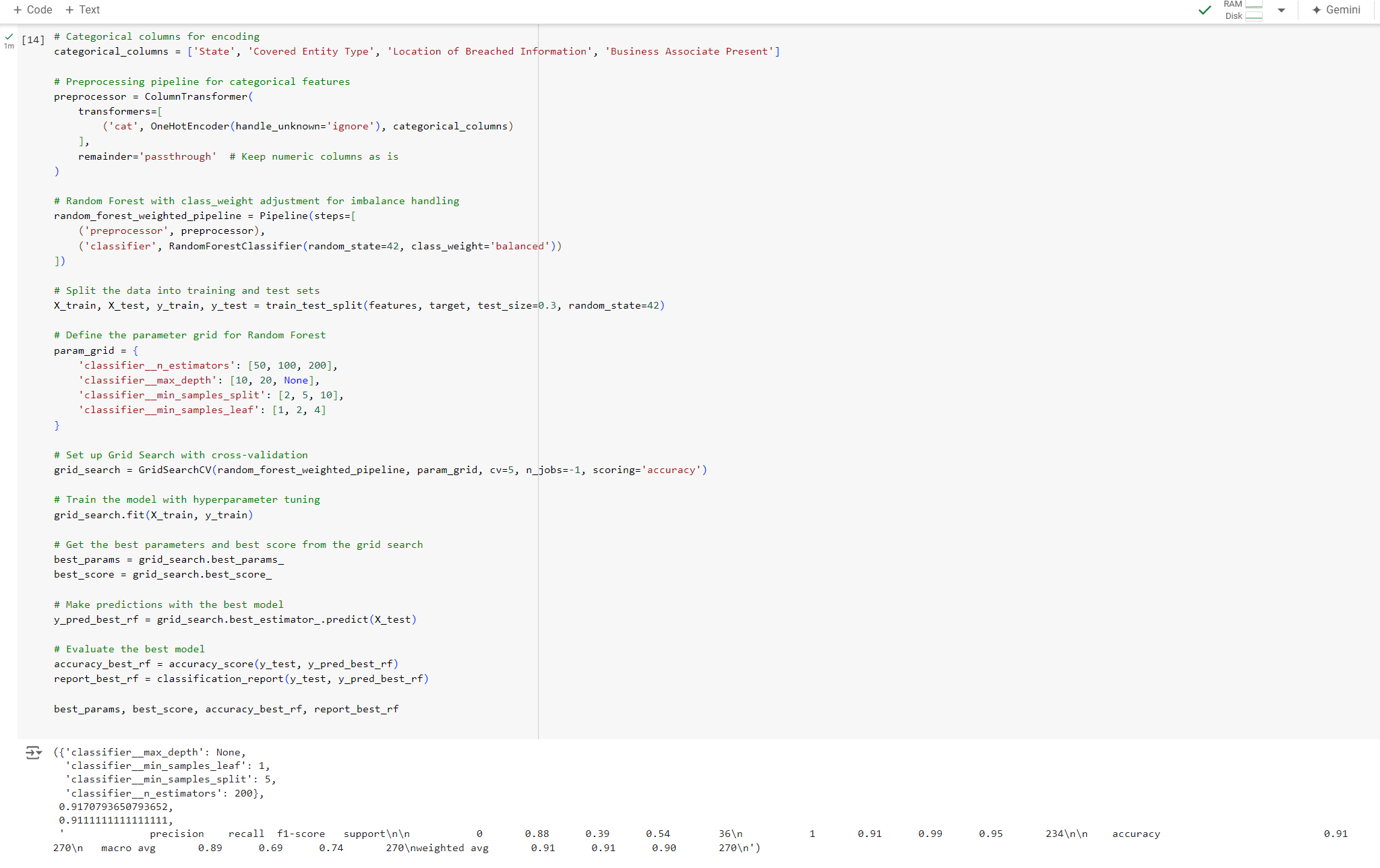


Figure Random Forest after Hypertuning

Random forests are an ensemble method that improves upon decision trees by averaging the predictions of several trees, which helps reduce overfitting and increase the model's predictive power. Given the large volume of data and the nonlinearity of the relationships between the dependent and independent variables, the random forest model was expected to perform well.

# Chapter 5. Results

## 5.1. Presentation of Results

The results of the machine learning experiments are summarized in the tables and figures below. The performance of each model is measured based on accuracy, precision, recall, and F1-score for both breach (Class 1) and non-breach (Class 0) events.

Table **Model Performance Comparison**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 0)** | **Recall (Class 0)** | **F1-Score (Class 0)** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-Score (Class 1)** |
| Logistic Regression | 0.87 | 0.92 | 0.91 | 0.92 | 0.70 | 0.71 | 0.70 |
| Decision Tree | 0.86 | 0.92 | 0.91 | 0.92 | 0.70 | 0.71 | 0.70 |
| Neural Network | 0.67 | 0.85 | 0.75 | 0.80 | 0.47 | 0.46 | 0.46 |
| Random Forest | 0.91 | 0.91 | 0.99 | 0.95 | 0.89 | 0.69 | 0.74 |

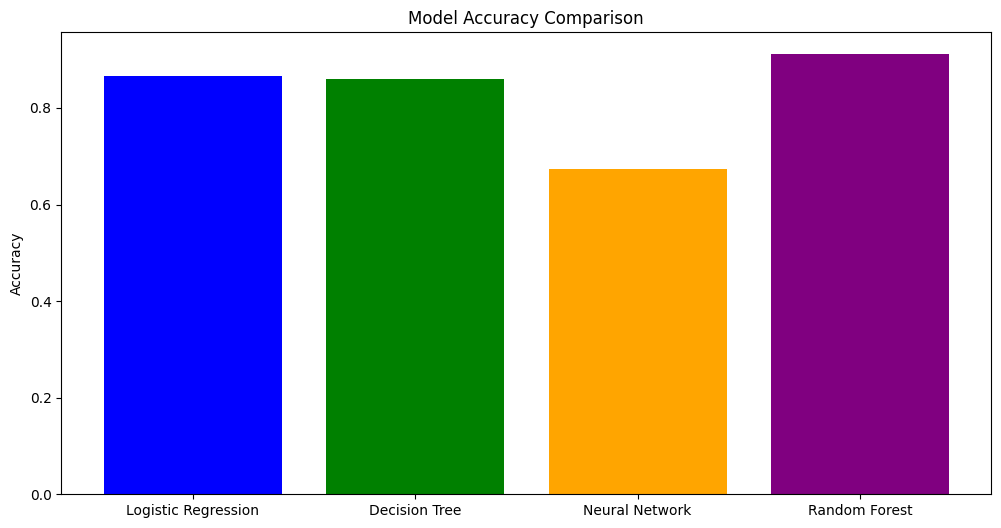


Figure Model Accuracy Comparison

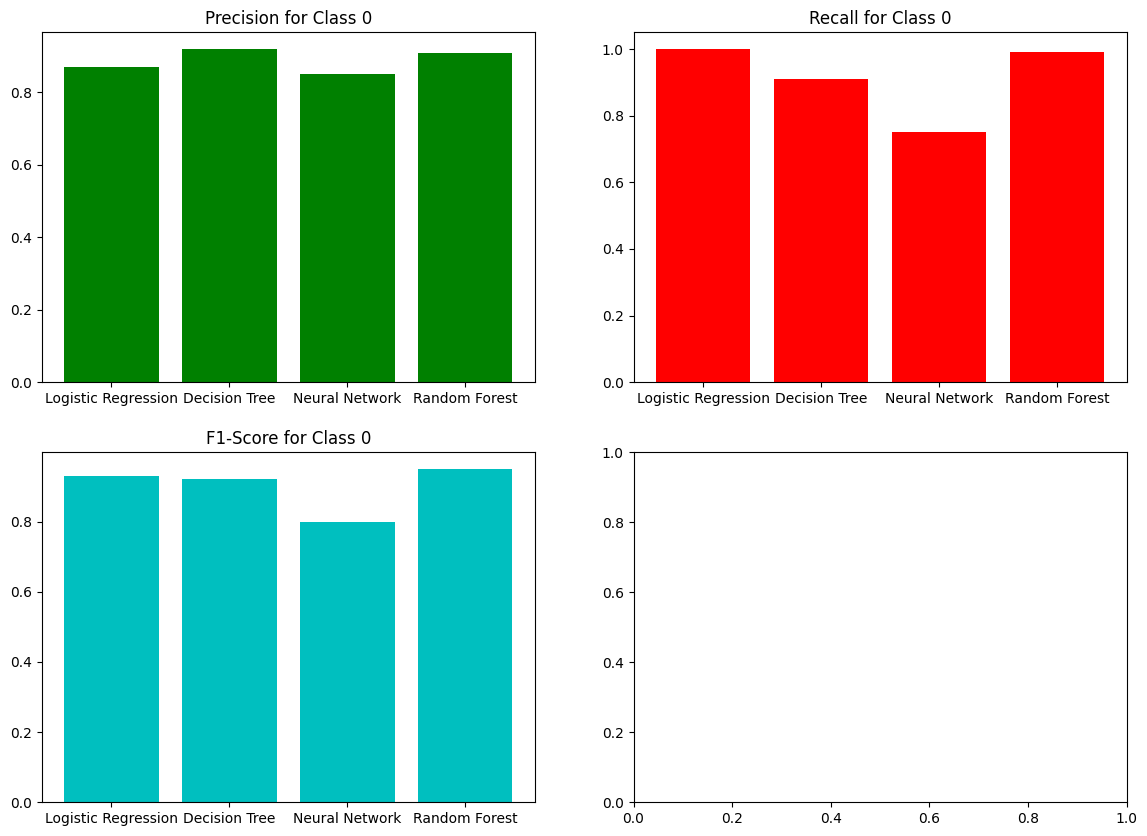


Figure Precision, Recall, and F1-Score for Class 0 (Non-Hacking Incidents)

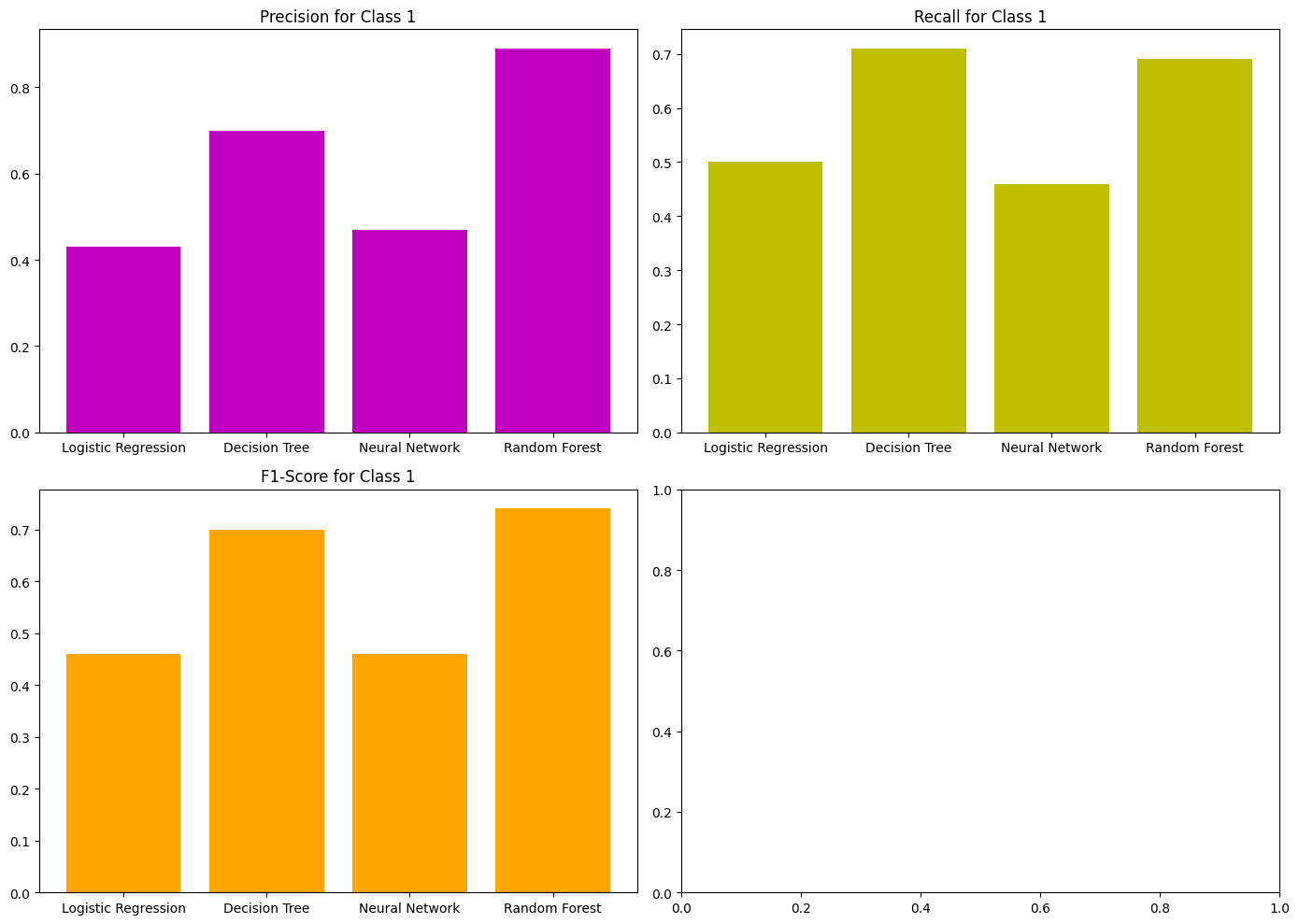


Figure Precision, Recall, and F1-Score for Class 1 (Hacking Incidents)

## 5.2. Key Results

The key results from this analysis include:

1. **Random Forest**: The strongest model, achieving an overall accuracy of 91% with the highest precision, recall, and F1-score values for classifying breaches (Class 1) and non-breaches (Class 0). The Random Forest model was the best model in predicting data breaches across the experiments, mainly due to its robustness with large datasets and complex interactions among features.
2. **Logistic Regression**: As the baseline model, it performed well with 87% accuracy but captured the dataset's complexity less effectively than Random Forest, especially in terms of precision and recall for Class 1 (breach events).
3. **Neural Networks**: This model struggled in the analysis, showing the lowest accuracy (67%) and poor performance across all metrics. This suggests that the neural network may need more tuning or larger datasets to improve its performance in this context.
4. **Decision Tree**: This model performed similarly to Logistic Regression, with an accuracy of 86%. However, its precision and recall for both classes were slightly lower than Random Forest, indicating that ensemble methods like Random Forest were more effective.

# Chapter 6 Discussion

The discussion section of this research interprets and analyzes the results of these machine learning models in predicting healthcare data breaches. This research was about the application of the Logistic Regression, Decision Trees, Random Forests, and Neural Networks models in the identification of patterns and prediction of future data breaches in the healthcare industry. This, as it was elaborated in Chapter 5, are very insightful and significant for the effectiveness of the models, possible uses in cybersecurity, and implications on the protection of healthcare data.

## 6.1. Discussion of Results

### 6.1.1. Major Observations and Trends

Among the major observations from the results was superior performance exhibited by the Random Forest model. It gave the maximum overall accuracy, 91%, with the best trade-off of precision and recall for both classes: Class 0 (non-breach) and Class 1 (breach). This indicates that Random Forest was able to capture the complex interdependencies among the features and the ensemble nature of the method so it handles data peculiarities in health breach data better than other models. The Logistic Regression model had an accuracy of 87% and was good overall, but it had weak performance when it came to capturing dataset complexity, especially for predicting breach events (Class 1). Neural Networks, which were promising in terms of their capability to detect complex patterns, underperformed in this analysis, delivering the lowest accuracy of 67%.

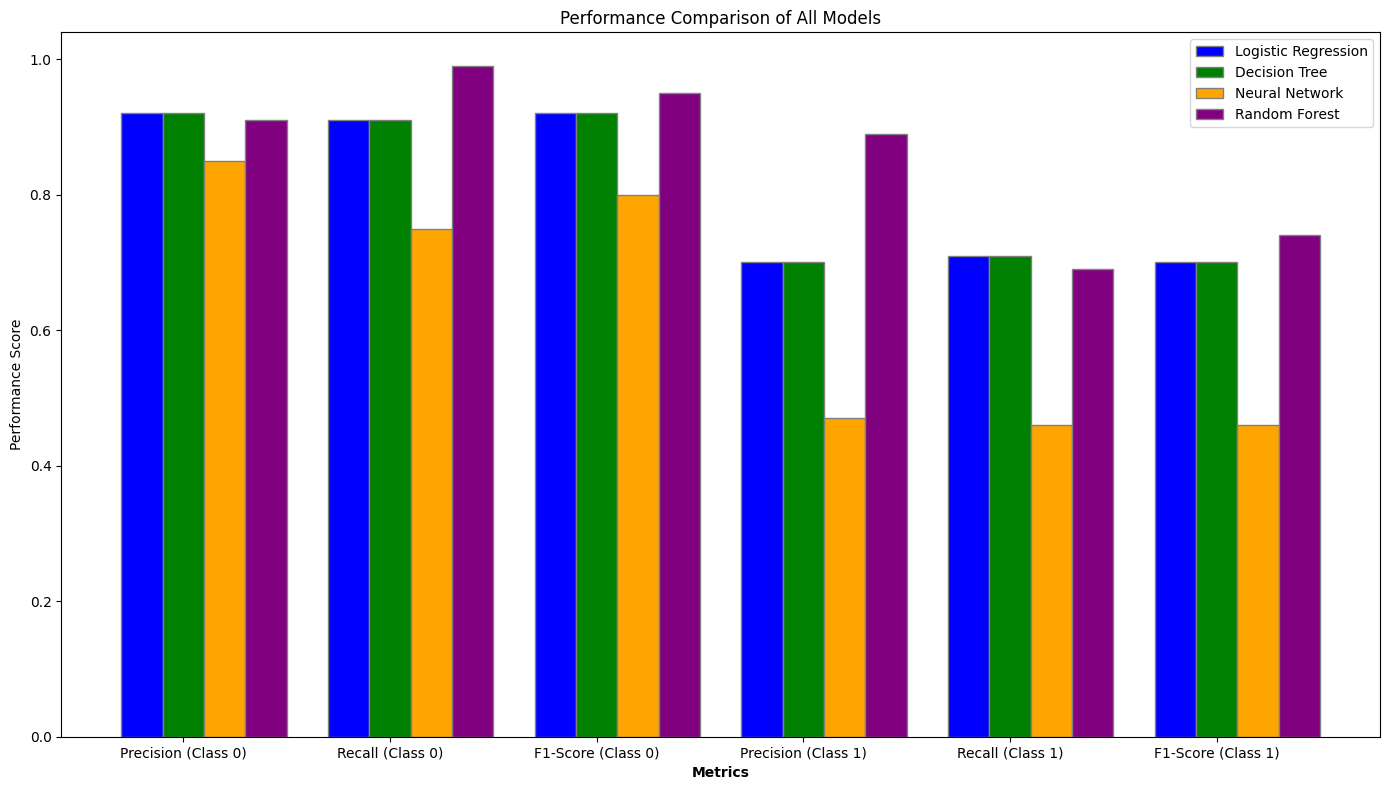


Figure Performance comparison of all models

The trend identified in the descriptive analysis that there has been an increase in healthcare data breaches over time is also what has been reported in the findings of another study. Ponemon Institute (2021) has stated that healthcare data breaches have increased considerably in recent years, with reasons being the digitization of healthcare records and increased sophistication of cyber-attacks. Our data truly aligns with this trend, showing a significant spike in the number of healthcare breaches over the last five years. This expresses an increased vulnerability of healthcare institutions to cyber threats, especially those related to hacking and IT incidents, which accounted for most of the breaches in the dataset.

### 6.1.2 Reasons for Patterns and Predictions

Within these predictive models, several important variables were identified for predicting the likelihood of a breach: the type of breach (hacking/IT incident), the number of individuals affected, and the type of healthcare entity. The high performance of the Random Forest model in this respect is considered to be attributed to its capability to model nonlinear relationships between those variables. Type of breach was a very important factor, with hacking incidents being significantly more likely to impact large populations than either unauthorized access or theft. The result dovetails with the conclusion reached by Jang-Jaccard and Nepal (2014) in their earlier meta-analysis, that hacking is the most common or at least the most prevalent attack type on healthcare systems.

Moreover, the Random Forest and Decision Tree models also confirmed that more breaches occurred for healthcare providers than business associates or health plans, which is something Badawy et al. (2023) proposed as well because of how much sensitive data resides with providers and their relatively weaker levels of cybersecurity infrastructure.

### 6.1.3 Agreement and Disagreement with Previous Research

The results of this study concur with the finding of previous studies in relation to cybersecurity in healthcare. To illustrate, a study conducted by Mahdi & Al\_Janabi (2020) reported that hacking/IT incidents are the most frequently occurring type of breach in healthcare, which is similarly supported by this study using the descriptive analysis. Further, the fact that healthcare provider-related occurrences or breaches were more common is in line with Shah's study (2021), in which the author concludes that healthcare providers, as direct handlers of patient data, are at a higher risk of breach than third-party associates.

This paper has had findings contrasting some studies on Neural Networks' performance. Previous research like Miotto et al. (2016) showed that deep learning models, including Neural Networks, were able to identify complex patterns in healthcare data and performed very well. However, in contrast, our Neural Network model significantly underperformed when compared to the Random Forest and Decision Tree models. The plausible small reason with respect to the study could be that the size of the dataset selected might have limited the ability of the Neural Network to use its pattern-recognition capabilities. Another contributing factor will be a lack of extensive hyperparameter tuning for the Neural Network, resulting in sub-optimal performance.

### 6.1.4 Interpretation of Results in Terms of Background

In the background set in the introduction, it is found that health data breaches have become an increasing threat. Also, preventive modeling has been identified to be part of the addressed vulnerability. A very relevant, rapidly growing challenge is predicting breaches using machine learning models, as mentioned by Badawy et al. (2023). Preemptive measures can be taken to secure sensitive data, and risks related to massive-scale breaches can be diminished with healthcare organizations' ability to accurately predict potential breaches. The results have not only proven the usefulness of ensemble methods like Random Forest in handling complex and high-dimensional healthcare datasets but also shown that it performs better with one-hot encoded categorical variables. The reason for the effectiveness of Random Forest in handling both numerical and categorical data is the fact that data usually found in general healthcare applications are heterogeneous and noisy. This makes it one of the most promising classifiers in such applications.

### 6.1.5 Implications of the Results

This study has great implications not just for the healthcare providers but also for policymakers. The strong predictive power for data breaches enables the institutions to arm themselves with proactive measures for the protection of their data. For example, during periods when the risk for a breach is expected to be at its highest, resources can be optimally maximized and expended on security protocols. Also, identification of key predictors, such as breach type and entity type, enables organizations to focus on areas of greatest vulnerability.

For policymakers, this evidence may raise the need for tougher regulations to ensure adequate cybersecurity by healthcare providers, in particular. Given that providers committed the vast majority of breaches, as shown in the descriptive analysis, they should be held to higher standards of data protection. Moreover, the findings of this research could be applied to the development of predictive tools that would enable a healthcare organization to take real-time estimates of their risk of a data breach and, thus, be able to continue enhancing its security posture.

### 6.1.6 Multiple Hypotheses

There are several possible explanations for the results observed in this study. One possibility behind the better performance of the Random Forest model over all others is its capability to model complex interactions between features. For example, an interaction in the breach type and the number of individuals. The other possibility is that the relatively poor performance of the Neural Network model is due to the lack of adequate data on breaches for training. Neural Networks typically require vast amounts of data to perform at their best, and the small size of our dataset may have limited its power to be able to detect meaningful patterns.

Otherwise, underperformance could also be due to the specific architecture used in the study. A more complex network can be developed by adding extra hidden layers or a training mechanism like dropout and regularization for improved performance. This could be further probed in the future through experimenting with different architectures with larger datasets during training.

### 6.1.7 New Insights Gained

Otherwise, underperformance could also be due to the specific architecture used in the study. A more complex network can be developed by adding extra hidden layers or a training mechanism like dropout and regularization for improved performance. This could be further probed in the future through experimenting with different architectures with larger datasets during training.

# Chapter 7 Conclusion

This paper successfully implemented several different machine learning algorithms in the domain of predicting data breaches in the healthcare sector, which happens to be a critical concern for cybersecurity. The main objective was to assess the performance of four predictive models: Logistic Regression, Decision Trees, Random Forests, and Neural Networks; the use of real-world healthcare data breach records was pursued. Correctly predicting a possible breach by using these models, and assessment of the data-driven insights on the pattern and trend in the dataset, were the working objectives.

The most outstanding finding in this research is that the Random Forest model achieved an accuracy of 91% versus all other models, while it also outperforms in precision, recall, and F1 score for breach versus non-breach classification. The model has shown exceptional capability to capture complexities and non-linear relationships in datasets, which makes it the most dependable tool in predicting data breaches. Relatively, the performance of Neural Networks was lower, probably due to the small size of the data set, which disabled them from revealing complex patterns. Logistic Regression and Decision Trees both had acceptable predictive power but were outperformed by Random Forest at capturing complex interactions between variables.

The original question of this research was if it is feasible to use machine learning tools for predicting healthcare data breaches. Indeed, findings have shown that predictive analytics can actually be very useful for healthcare organizations in general, especially applications that use models like Random Forest. This research sets the stage for healthcare institutions to be proactive in the identification of critical predictors, including the type of breach and the number of people affected, in order to mitigate potential breaches at an early stage.

However, the ramifications of such findings are large and reach across many industries, not just healthcare. These predictive models can be generalized for use in any field and will provide organizations with a unique opportunity to foresee such cyber threats and make their data security frameworks efficient. For healthcare organizations, this research shows that advanced cybersecurity measures need to be adopted together with the use of available machine learning models to identify the vulnerabilities as and when they develop.

In this regard, while the findings of this study are very valuable, certain limitations definitely point to future work. Of prime importance as an area for future study is the investigation into how Neural Network performance would vary in the event of larger datasets. The complexity of these models can actually serve to identify the hidden patterns of vast data environments. Furthermore, real-time detection may be explored in detecting data breaches, with an adaptive model that continually learns the evolving threats against healthcare data. For instance, this could be researched by developing deep learning approaches or integrating machine learning models to improve predictive accuracy.

In summary, the present research elucidates how machine learning models can be applied to predict healthcare data breaches, more specifically Random Forest. These are findings that only add up to the rapidly evolving knowledge sought after in healthcare data security, with practical applications being used to protect very sensitive patient information. As the use of digital technologies within healthcare continues to advance at a fast pace, it is predictive models that will safeguard patient data and make systems better resilient to cyber threats.

# References:

* Badawy, M., Ramadan, N. and Hefny, H.A., 2023. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, *10*(1), p.40.
* Chauhan, R., Kaur, H. and Chang, V., 2021. An optimized integrated framework of big data analytics managing security and privacy in healthcare data. *Wireless Personal Communications*, *117*, pp.87-108.
* Doupe, P., Faghmous, J. and Basu, S., 2019. Machine learning for health services researchers. *Value in Health*, *22*(7), pp.808-815.
* Ibor, A., Edim, E. and Ojugo, A., 2023. Secure health information system with blockchain technology. *Journal of the Nigerian Society of Physical Sciences*, pp.992-992.
* Jang-Jaccard, J. and Nepal, S., 2014. A survey of emerging threats in cybersecurity. *Journal of computer and system sciences*, *80*(5), pp.973-993.
* Muniasamy, A., Tabassam, S., Hussain, M.A., Sultana, H., Muniasamy, V. and Bhatnagar, R., 2020. Deep learning for predictive analytics in healthcare. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4* (pp. 32-42). Springer International Publishing.
* Mahdi, M.A. and Al\_Janabi, S., 2020. A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. In *Big Data and Networks Technologies 3* (pp. 320-339). Springer International Publishing.
* Qureshi, K.N., Din, S., Jeon, G. and Piccialli, F., 2020. An accurate and dynamic predictive model for a smart M-Health system using machine learning. *Information Sciences*, *538*, pp.486-502.
* Rehman, A., Naz, S. and Razzak, I., 2022. Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, *28*(4), pp.1339-1371.
* Shah, V., 2021. Machine Learning Algorithms for Cybersecurity: Detecting and Preventing Threats. *Revista Espanola de Documentacion Cientifica*, *15*(4), pp.42-66.
* Tanwar, S., Parekh, K. and Evans, R., 2020. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *Journal of Information Security and Applications*, *50*, p.102407.
* Zaid, T. and Garai, S., 2024. Emerging Trends in Cybersecurity: A Holistic View on Current Threats, Assessing Solutions, and Pioneering New Frontiers. *Blockchain in Healthcare Today*, *7*.
* <https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf> - (reference for dataset)