University of Essex

School of Mathematics, Statistics
and Actuarial Science

# Predicting Chronic Diseases Using Machine Learning

**Vishal Manoj Bachal**

# Contents

# List of Figures

# List of Tables

CHAPTER 1

# Abstract

The study focuses on predicting chronic diseases like asthma, diabetes, and brain stroke by applying several machine learning algorithms for diagnosis and interpretation. Advanced algorithms used in this study include 'Random Forest', 'XGBoost', and 'Neural Networks' because of their robustness in modelling complex patterns. Feature scaling, balancing using SMOTE, and feature engineering were the preprocessing techniques applied to these datasets to improve the models' performance.

Each algorithm was fine-tuned to optimise accuracy, precision, and recall. Results show the performance of these models, where 'Random Forest' and 'XGBoost' performed very well across most datasets in terms of precision and recall, while 'Neural Networks' effectively captured the intricate relationships among features. Compared with earlier studies, this research demonstrated significant improvements in prediction accuracy and identified key risk factors such as age, BMI, and glucose levels. This highlights the reliability of these models and their practical applicability in healthcare.

The models performed well, but future work could look into using more advanced designs, a wider variety of datasets, and real-time applications to expand the research further. Overall, this study offers a solid framework for using machine learning to address important healthcare issues.

CHAPTER 2

# Introduction

Asthma, diabetes, and brain stroke are examples of chronic diseases that rank among the leading causes of death and disability in the world. This sets a high burden on health systems and affects many lives. The World Health Organization reports that chronic diseases are responsible for about 74% of global deaths [41]. They are generally characterized by interrelated acted interaction involving genetic, environmental, and lifestyle factors. These diseases, if managed appropriately and detected early, can be effectively dealt with; this again explains why there has been growing research interest in developing predictive models for these diseases.

Asthma, brain stroke, and diabetes were chosen for this study because they form diseases that are common and whose outcomes can be altered if early causative treatment is offered on time. Asthma is a form of chronic disease affecting hundreds of millions worldwide, and if not well managed, it becomes a severe health complication. This can help practitioners prevent high-risk cases before one can develop symptoms or trigger an emergency incident and make long-term care possible[42]. Among the most common and abrupt causes of grave disability and death, brain stroke-from here on referred to as stroke-requires the necessity of applying primary preventive strategies that can be empowered with machine learning-this latter technique being potentially able to manage many risk factors ranging from blood pressure to lifestyle habits[8]. Diabetes is a metabolic disorder that affects hundreds of millions of people worldwide. The complications arising from diabetes, such as heart disease and kidney failure, may be life-threatening conditions. Predictive models may enable the early adoption of lifestyle

modifications or treatments to avoid the onset of a more serious disease and, therefore, support the disease with less burden for patients and healthcare systems[1].

Although traditional methods for the diagnosis of chronic diseases are useful, they remain very time and resource consuming. They tend to be overly dependent on the specialists. Due to the ageing population, a greater number of people suffer from chronic diseases; similarly, people living in urban areas are more prone to chronic diseases due to their lifestyles, so there is a dire need for quick diagnostic tools. Hence, machine learning and artificial intelligence are close to becoming the good solution. They can work with huge volumes of data at an unprecedented speed, find hidden trends and patterns, and forecast health outcomes with a degree of precision to which no human meshed group could aspire[11]. It is particularly effective when considering relatively complicated data from electronic health records and wearable devices data.

Certain procedures of machine learning apply well in the field of healthcare. Random Forest, XGBoost, and Neural Networks-three of the best classification methods-have the properties of prediction for a patient having a disease based on medial data. Random Forest is an ensemble of decision trees whose predictions are averaged out to get a more accurate and stable prediction. It is appropriate for this type of chronic disease prediction because it easily handles heterogeneous and complex data[14]. XGBoost also enjoys efficiency and superior performance on large data sets, characteristic of medical environments. It is a boosting-based method, considering the idea of several models in sequence with each successive model attempting to correct errors made by the last. It is particularly useful for the accurate estimation of risk [9]. Neural Networks, patterned after the human brain, are very effective in finding patterns in complex data. Deep learning has proven to be a class of neural networks that itself has revolutionized the field of healthcare by providing accuracy in disease prediction through modelling non-linear and complex relationships within the data[25].

This study tries to investigate the prediction capability of Random Forest, XGboost, and Neural Networks on the disease. I will further use accuracy, precision, recall, and F1-score metrics to evaluate the performance of these algorithms. This method plays an important role in dealing with data imbalance, feature selection, and improving the accuracy of models in order to find the most effective model toward each of the diseases.

CHAPTER 3

# Literature Review

Machine learning (ML) offers powerful tools for the prediction and early diagnosis of chronic diseases, which significantly burden healthcare systems and impact patient quality of life. In particular, ML methods have shown promising results in predicting asthma, brain stroke, and diabetes, allowing for proactive management and timely interventions. This review evaluates key studies on ML applications for each of these conditions, exploring the methodologies used, primary findings, and existing gaps in the literature. Each disease is discussed individually, followed by a cross-disease analysis and an exploration of the challenges and limitations common across ML applications in chronic disease prediction.

## 3.1 Asthma

In this study [16], applied machine learning to telemonitoring data in order to predict asthma exacerbations. To this dataset, including 7,001 records of self-reported symptoms and medication on a daily basis by asthma patients, several algorithms were applied: Naïve Bayes, Adaptive Bayesian Networks, and Support Vector Machines. Sensitivities were 80%, 100%, and 84%, respectively. The study concluded that the temporal window was one of the most limiting aspects, which limited predictive accuracy of the classes to 7 days, thus proving that stratified sampling was effective since it aided in solving class imbalance problems by increasing sensitivity and specificity. The authors have also discussed how the integration of other databases, such as genomic, clinical

data, and environmental factors, would be important in further enhancing predictive accuracy. Whereas temporal patterns were helpful in their model, the study outlined some limitations regarding data diversity and the impacts of temporal autocorrelation. Their results also give a good example of how machine learning could be useful in the management of chronic diseases, providing a basis for more complete predictive frameworks.

This study [6], applied various machine learning algorithms to real-world asthma diagnosis data involving 3,000 subjects. They tested ten different machine learning models of which the key ones pertained to Random Forest, GBM, and C5.0-as based on their performance metrics comprising accuracy, precision, and recall. Among these, the Random Forest and C5.0 algorithms produced the highest classification accuracy of 94.75% on testing data. Thus, the above two algorithms are efficient in classifying asthma cases. FEV3, FVC, and FEV1 were identified as the important variables of pulmonary function in this analysis and thus were considered the top influential features of asthma diagnosis, giving emphasis on the importance of lung function tests. The authors further explained that feature importance evaluation is crucial to understand the contribution of each variable in the predictions made by the models, with further support for clinical decisions. They also stressed that proper proper data preprocessing and hyperparameter tuning can be done for the performance optimization of models using techniques such as cross-validation and grid search. It points out the potential of ensemble methods, like Random Forest and GBM for health analytics, thus giving a useful reference to future research related to the field of asthma and other respiratory diseases.

Machine Learning techniques are applied in this work by [38], for the prediction of asthma and COPD, using clinical data from 132 patients. They used various algorithms in order to ascertain what will give the best model for each disease, such as Random Forest, Logistic Regression, Neural Networks, and Support Vector Machines. Random Forest had the highest asthma precision with 80.3%, where MEF2575 was identified as the most important feature in the prediction, followed by other important features of age, smoke, and wheeze, pointing toward spirometry and symptoms in diagnosis. The study focused on identifying the most critical features and tried to enhance both the optimization of the prediction and the interpretability. Techniques of hyperparameter tuning and cross-validation were applied for performance improvement of the models.

This fact shows capability in the machine learning field to detect respiratory diseases using only spirometry data, focusing on variables such as MEF2575. Results such as this could form a basis on which such models should be integrated into a clinical decision support system.

This research focuses on [22], examined certain machine learning models for predicting childhood asthma based on a sample size of 202 children. The proposed algorithms included Random Forest, Logistic Regression, and Support Vector Machines, among which the best performing was Random Forest with an accuracy of 84.9%, followed by Logistic Regression at 82.57% and SVM at 82.5%. Maternal atopy, cesarean delivery, cold air, and dust mites were some of the major factors that once again identified the importance of environmental and maternal characteristics during the prediction of asthma. Chi-squared testing feature selection showed that 19 out of 36 variables with significant associations consisted of maternal health, breastfeeding, and childhood exposures. Although quite strong, the authors did note limitations to this study with regard to sample size and population diversity; stated differently, larger and more diverse datasets could enhance generalizability.

## 3.2 Diabetes

In this work [34], have proposed the prediction of diabetes using three machine learning classification algorithms, namely Naïve Bayes, Support Vector Machine, and Decision Tree. The experiments have been conducted on the Pima Indians Diabetes Database containing 768 instances and eight attributes like glucose level, BMI, and age. The models were evaluated for accuracy, precision, recall, and F-measure. Among these, Naïve Bayes gave the best accuracy value of 76.30%, closely followed by the Decision Tree with an accuracy of 73.82%, while SVM had an accuracy of 65.10%. It was also clear that the Naïve Bayes handled the incomplete data best and was robust to unbalanced datasets since it assumes features probabilistically. This paper also underlined feature selection and preprocessing with standardization and cross-validation. Comparing different algorithms, among them, Naïve Bayes is more suitable to work with datasets concerning efficiency and reliability in classification. These results further support the potential of machine learning in the prediction of diabetes and the ability of machine

learning to simplify early diagnosis in medical fields.

In this study [37],applied six machine learning algorithms, Random Forest, Gradient Boosting, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), on the Pima Indians Diabetes Dataset comprising a total of 768 samples to predict diabetes. Glucose level, BMI, and age were among the dataset features. Handling missing values and normalization were among the preprocessing steps. Among those, the best performance was obtained with Random Forest, which had 77% and stated its capability in handling big datasets and interactions among features. Then comes Gradient Boosting, which also had a very similar performance, hence highlighting its strength in handling the complexity introduced by diverse data. It proposes the use of ensemble techniques such as bagging and boosting on different algorithms for better predictive accuracy. The most informative attributes identified were glucose and BMI. Obtained results from multiple algorithms offered greater reliability as compared to the ones from single models. The findings brought out the potential of the ensemble methods in improving diagnostic accuracy and, thereby, possible early detection of chronic diseases like diabetes.

This research focuses on [23] to compared the performances of different machine learning algorithms like Logistic Regression, SVM, Random Forest, and Neural Networks using the Pima Indians Diabetes Dataset. Performances were as follows: Neural Networks with 88.6%, Logistic Regression with 78%, and SVM with 78%. Glucose and BMI are thus selected as the major predictors through feature selection. These findings underlined the results of the feasibility of Neural Networks for medical prediction by showing how much preprocessing and well-tuned models guarantee better results. Findings like these really encourage deep learning to be a promising tool in managing diabetes.

This research [12], used machine learning to predict complications among T2DM patients. In their paper, data from 1,000 patients were used to run Logistic Regression, Random Forest, and SVM to predict outcomes such as retinopathy and nephropathy at 3, 5, and 7 years. The best performance was obtained by Logistic Regression with an accuracy of 83.8%. Strong predictors included HbA1c, BMI, and hypertension. In this study, missing data were imputed, and oversampling was done to handle the problem of class imbalance. Although effective, some biomarkers were lacking in this

study, hence limiting the scope. This research therefore manifests the potential usage of machine learning in the early complication detection for diabetes care.

## 3.3 Brain Stroke

In this work [3], machine learning algorithms were used to develop a system for predicting brain stroke. Using a dataset of 5,110 records from Kaggle, comparisons of three classifiers (Random Forest, Support Vector Machine, Decision Tree) on the basis of prediction accuracy were done. The dataset consisted of 11 input features such as age, hypertension, BMI, smoking, and average glucose level, against a single output feature regarding stroke. Random Forest gave the best performance with an accuracy of 95.30%. This signifies its high supremacy in performance at classification while SVM and DT gave accuracies of 91.98% and 89.53%, respectively. In the study, the data were preprocessed to include standardization and imputation to enhance quality. Some of the performance measures used to achieve the model sensitivity were: sensitivity, specificity, and false positive rates. Comparisons between different models were done based on the performance measures. The best performance turned out to be from the Random Forest, as it can interactively handle features effectively and make robust predictions. The current study underlines the use of ensemble methods such as RF in the analysis of medical data and presents how well RF performs in early stroke detection, thus supporting clinical decision-making.

The author [15] have proposed a stroke prediction with weighted voting classifier using separate machine learning algorithms. The dataset consists of 5,110 records collected from Bangladeshi hospitals containing features that are age, BMI, hypertension, heart, glucose level, and smoking status. The model was trained and tested with ten classifiers, such as Logistic Regression, XGBoost, Gradient Boosting, and Decision Tree. Among these, the weighted voting classifier achieved the best performance with an accuracy of 97%, beating results for all individual classifiers. Closing in on it is the algorithm of XGBoost, followed by Gradient Boosting at an accuracy of 96% each. The combination of predictions in an attempt to improve accuracy and reduce error rates was effective, according to the study. AUC, among other main performance metrics like precision, recall, and F1-score, was higher for the weighted voting classifier.

This approach also gave more insight into the contributing factors of strokes and thus illustrated the application of ensembles in medical diagnosis. These findings would, therefore, suggest that use of such models in a clinical setting could potentially boost the prospects for early diagnosis of stroke, thereby enhancing outcomes.

This author [39], proposed a stroke risk prediction by employing six machine learning algorithms: SVM, Random Forest, and KNN. Using a publicly available dataset that included features such as age, BMI, and hypertension, among others, the best performance was realized by the SVM algorithm with an accuracy of 94.6% and AUC of 99%. Preprocessing included cleaning for outliers, encoding, and normalization. This work showed that the combination of cross-validation and metrics like precision and recall increases the reliability of the predictions. The correlation features analysis allowed the exclusion of less important attributes, while methods such as SVM and ensemble became an important tool for early diagnosis in stroke conditions.

In this paper [30], developed a machine learning model for predicting brain stroke risk. The authors used Random Forest and AdaBoost on patient data that had demographic presentation, past conditions, and lifestyle factors. Random Forest showed an accuracy of 95%, while AdaBoost literally followed it with an accuracy of 94%. The reason for this choice is their strength in working with high-dimensional data and elaborating complex relationships between variables. The study confirmed the strength of Random Forest in making robust predictions by means of ensemble learning and the iterative emphasis of AdaBoost on improving misclassifications. While most of them were focused on the theoretical part, this research gave more emphasis to its practical application by integrating the models outputs into an interactive and user-friendly system. The results showed that the system could take input test data from the users and provide the predictions, thus proving quite useful in healthcare applications in real time. It indicated how advanced machine learning models can be used to support health professionals in the early detection and prevention of stroke.

The author[28]propose a Stacking Ensemble model for predicting brain stroke by integrating five base learners of machine-learning algorithms: Random Forest, KNN, Logistic Regression, SVM,SVC, Naïve Bayes and Stacking, while taking Random Forest as the meta-learner. On the basis of 5,110 records of such features as age, BMI, glucose, and hypertension, the stacking model obtained an accuracy of 97% with an MCC of 94%.

Preprocessing consisted of handling missing values, encoding labels, standardization, and SMOTE to handle data imbalances. This study mentioned the better performance of various stacking methods, ensemble multiple algorithms into improving the prediction. The approach highlighted how ensemble methods can be used to enhance the diagnostic accuracy of such complex conditions as brain stroke by making the decision more reliable and flexible for clinical applications.

## 3.4 Cross-Disease Analysis

Among asthma, brain stroke, and diabetes, Random Forest appeared to be one of the most powerful methods for prediction. Large Neural Networks follow next in the list, when more than once it turned out to perform best in diseases like stroke or diabetes with complex nonlinear interactions between the data samples. The reviewed studies show that the diversity of integrated types of data improves accuracy, but at the price of sophisticated preprocessing and quality control.

## 3.5 Research Questions

1. Which machine learning algorithms work best for predicting asthma, brain stroke, and diabetes? How do these algorithms stand in comparison to each other on the basis of accuracy and reliability?

2. What are the main challenges researchers face when using ML to predict chronic diseases? In other words, does the small size of datasets or data quality regarding the analysis affect the results?

3. How can the ML models be designed to provide specific, actionable insights with which healthcare providers can make timely personalized decisions for patients?

4. Where are the gaps in current research, and what improvements are needed to make these models more accurate and reliable across different groups of people?

The iterature shows that machine learning, especially ensemble methods like Random Forest and Neural Networks, holds very promising results in the prediction of chronic diseases such as asthma, brain stroke, and diabetes. Generally, Random Forest

models excel in handling diverse data types, while Neural Networks prove to be effective in capturing complex nonlinear interactions that are most evident in stroke and diabetes predictions. Most of the literature reviewed emphasizes how the integration of genetic, clinical, and environmental information enhances accuracy and enables personalized health approaches. However, most of these studies relied on small datasets, thus limiting generalizability and possibly overfitting, especially for deep learning models. Future studies should aim at more extended and diversified datasets that guarantee the robustness, reliability, and applicability of the ML models to wider population groups.

CHAPTER 4

# Objective and challenges of project

**Purpose of the Study:**

This research targets the application of ML in predicting and diagnosing chronic diseases like asthma, brain stroke, and diabetes. Such conditions not only affect a long-term health situation but also press hard on healthcare systems. With ML, this study contributes to early detection and accurate diagnosis so that timely and precise interventions can be made by the healthcare providers.

One of the key strengths of this study is its use of an appreciably large dataset than previous research, which enhances the reliability and generality of our predictions, thereby furthering the understanding of factors that influence disease development. This work carefully examines different data types for the most powerful predictors that applies to ML algorithms: medical history, genetic profile, and environmental exposures.

The study, therefore, tries to incorporate a larger and diversified dataset in the process of developing a model with the purpose of providing personalized predictions according to the evaluation of some case over the possibility of a patient having a particular disease given his or her particular risk factors. The paper embraces some common challenges concerning the application of ML to healthcare, unusual assurance of data quality, and effective integration for diversified data types. Addressing such challenges will create robust and flexible ML models instrumental in producing early and accurate diagnoses, thereby advancing the management of chronic diseases.

**Challenges and Limitations:**

The aim of this work requires emphasizing the problems and limitations that exist in the application of machine learning in chronic disease prediction. One of the first major challenges involves the quality of data, since loss or inconsistency of values in health data may affect model performance and hence requires careful processing. Also, complex models, like Neural Networks, have the risk of overfitting; this means that the model could learn concrete patterns in training data, which may not generalize well to new, unseen cases. Not to mention ethical considerations: these are fairness, absence of bias, and equal opportunities in providing access to predictive tools, so the advances benefit all patient groups. Although this survey is based on a higher number than in many other studies, generalization across diverse populations remains a challenge. The development of machine learning models as a means for reliable, fair, and adaptive practitioners in different healthcare settings faces a raft of issues that need to be targeted.

## 4.1 Supervised Classification

In supervised classification, which is one variation of machine learning, models can be trained using labeled data such that they identify patterns and, thus, predict specific outcomes based on new data. The application of supervised classification in predicting chronic diseases will enable us to classify patients into haves and have-nots regarding their disease status, using their medical and demographic information. This concept depends on the historical data with known outcomes-labeled data that helps the model in learning relationships between different health features such as age, genetic markers, medical history, environmental factors, and the target variable, which in this case is disease presence or absence[36].

In health care, this classification turns out to be appropriate because risk prediction of chronic diseases may allow early intervention and precautionary care, hence reducing its impact on patients and health systems alike. Using previous data, a supervised classification model can learn how to make accurate predictions for new case identification that will be helpful in identifying high-risk individuals to support healthcare providers in making timely, informed decisions[36]. The section below develops reasons for choosing supervised classification as the main approach in the study, the chosen algorithms, and metrics to assess their effectiveness in predicting chronic diseases.

## 4.2 Why Supervised Classification is Suitable for This Study

Supervised classification is applicable in this study, given the nature of our datasets and the objectives of chronic disease prediction. There exist labeled examples in each dataset where the target variable "disease" clearly states any of the chronic diseases such as asthma, brain stroke, or diabetes. By training on this labeled data, the model learns from past cases and generalizes its predictions to new, unseen data to improve its capabilities in accurately identifying disease risk.

This predictive capability is important in healthcare, where any form of early diagnosis leads to better patient outcomes by way of enabling preventive measures before the symptoms worsen. Using supervised classification, it will identify patterns within patient data that provide insight into risk factors that might not be immediately apparent through traditional analysis. This approach brings us back to our objective of accurate and actionable predictions upholding early intervention and personalized healthcare, and thus calls for the best choice of supervised classification in this study.

## 4.3 Algorithms Used in This Study

In this work, we adopt three supervised classification algorithms for their strong points in handling healthcare data: Random Forest, XGBoost, and Neural Networks. These algorithms are selected accordingly based on their ability to deliver reliable predictions and because their effectiveness has been proved in applications related to healthcare. Further, each algorithm is overviewed in sections presenting a discussion of its main characteristics and specific application in this research. Specific methodologies will be discussed in further detail throughout successive chapters with a view to explaining how those algorithms work and why they were chosen for chronic disease prediction.

### 4.3.1 Random Forest

Random Forest is also an ensemble-based algorithm that takes a combination of the outputs from multiple decision trees in trying to present improved accuracy and stability. It is highly effective in handling datasets with diverse feature types and interactions,

making it a reliable choice for classification tasks. It reduces the risk of overfitting because the model averages the outputs across many trees, resulting in robust and interpretable results. Among its many desirable features, Random Forest especially has the ability to provide feature importance rankings to understand which variables most influence our predictions.[14]

### 4.3.2 XGBoost

XGBoost is stands for Extreme Gradient Boosting, a powerful boosting algorithm, ideated over structured data. It builds decision trees one at a time, with every tree correcting errors from previously built trees. This iterative process is the way XGBoost will refine predictions and weak patterns in data. It is well-liked for its speed and efficiency. It is quite adaptable and handles high-dimensional data very well.

It is also built into the algorithm that regularization techniques prevent overfitting and making it as a reliable tool in coming up with precise results[10].

### 4.3.3 Neural Networks (NN)

Neural networks represent a family of deep learning models that enable the modeling of sophisticated nonlinear relationships between data. The neural network is composed of layers of interconnected neurons, with each learning higher levels of abstraction of features. Neural networks are naturally good for data that has complex patterns that involve more than the regular number of features. This flexibility enables the models to capture hidden relationships that might be hard to discern by using other, more traditional models[33].

## 4.4 Evaluation Metrics for Supervised Classification

To evaluate the performance of these algorithms, we used several key metrics in predicting chronic diseases. Each metric provides information on different aspects of model performance that guarantees the precision, reliability, and clinical relevance of the models predictions.

- **Accuracy**:The metric calculates the overall percentage of the correct predictions out

of all the made predictions. Accuracy provides a bird's view of model performance and gives a general sense of correctness.

Accuracy is calculated using the formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

[19]

This can also be written as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

[19]

Where:

- **TP (True Positives)**: Cases where the model correctly predicted the positive class.
- **TN (True Negatives)**: Cases where the model correctly predicted the negative class.
- **FP (False Positives)**: Cases where the model incorrectly predicted the positive class.
- **FN (False Negatives)**: Cases where the model incorrectly predicted the negative class.

This formula calculates the ratio of correctly classified instances (both positive and negative) to the total number of instances, offering a clear and concise evaluation of the model's overall performance.

- **Precision and Recall**: Precision is the exactitude of the model's selection of those true positive cases tuned to minimize false positives. Recall measures the effectiveness of classifying all actual positive cases correctly, emphasizing minimization of false negatives. In healthcare, this will be very important, as these metrics can avoid misclassifications which might lead to wrong diagnosis.

  Precision calculated using the following formula:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

[19]

Where:

- **TP (True Positives)**: Cases where the model correctly predicted the positive class.
- **FP (False Positives)**: Cases where the model incorrectly predicted the positive class.

Precision measures how accurate positive predictions are and is calculated using a specific formula. This formula focuses on reducing false positives, which is especially important in fields like healthcare. In such cases, a wrong diagnosis could lead to unnecessary treatments or procedures, potentially causing harm or extra costs.

Recall is calculated using the following formula:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

[19]

Where:

- **TP (True Positives)**: Cases where the model correctly predicted the positive class.
- **FN (False Negatives)**: Cases where the model failed to predict the positive class.

This above formula focuses on reducing false negatives, making sure no positive cases are overlooked, which is crucial in diagnosing diseases.

- **F1-score**:This is a single measure that combines precision and recall. Many times, one needs to handle an imbalanced dataset in which the number of disease cases is usually less compared to the non-disease cases. There, one may consider F1 Score in health care.

F1-score is calculated using the following formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

[19]

Where:

- **Precision**: The proportion of correctly predicted positive cases out of all predicted positive cases.
- **Recall**: The proportion of correctly predicted positive cases out of all actual positive cases.

This formula is especially helpful when dealing with datasets where one class is much smaller than the other. It balances the advantages of both precision and recall into one metric, making it easier to evaluate the model's performance.

CHAPTER 5

# Description of Dataset

## 5.1 Asthma Dataset

**Description**- The asthma dataset provides a comprehensive look into factors associated with asthma diagnosis and management. The dataset contains various demographic, environmental, and clinical variables, thus finding a good ground on which predictive modeling may be used. Key variables expected within the dataset include demographics like age and gender, lifestyle factors-smoking and physical activity, and environmental exposure to pollen and dust. It allows the feature of carrying out a detailed analysis of asthma risk and enables modeling for disease occurrence and severity. This dataset was downloaded from Kaggle, and it is available on the Kaggle website.

**Link**: https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset/data

The dataset's organized structure makes it perfect for drawing insights and boosting prediction accuracy in asthma-related health studies.[24]

**Dimensions:** The dataset, **asthma_disease_data.csv**, contains **2,392 rows** and **29 columns**. Each row represents data for an individual patient, while each column expresses specific attributes related to asthma. It uniquely identifies every patient with a Patient ID and includes highly elaborative informationacross various categories like demographics, lifestyle factors, environmental exposures, medical history, clinical measurements, symptoms, and asthma diagnosis status.Below table represents important

features of dataset such as:

**Column Descriptions:**

| Column Name | Description |
|---|---|
| PatientID | Unique identifier for each patient, ranging from 5034 to 7425. |
| Age | Ranges from 5 to 80 years. |
| Gender | 0 = Male, 1 = Female. |
| Ethnicity | Coded as 0 = Caucasian, 1 = African American, 2 = Asian, 3 = Other. |
| Education Level | 0 = None, 1 = High School, 2 = Bachelor's, 3 = Higher. |
| BMI | ranges from 15 to 40. |
| Smoking | 0 = No, 1 = Yes. |
| Physical Activity | Weekly hours, ranges from 0 to 10. |
| Diet Quality | Score from 0 to 10. |
| Sleep Quality | Score from 4 to 10. |
| Pollution Exposure | Scored from 0 to 10. |
| Pollen Exposure | Scored from 0 to 10. |
| Dust Exposure | Scored from 0 to 10. |
| Pet Allergy | 0 = No, 1 = Yes. |
| Family History of Asthma | 0 = No, 1 = Yes. |
| History of Allergies | 0 = No, 1 = Yes. |
| Eczema | 0 = No, 1 = Yes. |
| Hay Fever | 0 = No, 1 = Yes. |
| Gastroesophageal Reflux | 0 = No, 1 = Yes. |
| Lung Function (FEV1) | Ranges from 1.0 to 4.0 liters. |
| Lung Function (FVC) | Ranges from 1.5 to 6.0 liters. |
| Wheezing | 0 = No, 1 = Yes. |
| Shortness of Breath | 0 = No, 1 = Yes. |
| Chest Tightness | 0 = No, 1 = Yes. |
| Coughing | 0 = No, 1 = Yes. |
| Nighttime Symptoms | 0 = No, 1 = Yes. |
| Exercise-Induced Symptoms | 0 = No, 1 = Yes. |
| Diagnosis | 0 = No asthma, 1 = Diagnosed with asthma. |
| Doctor In Charge | Listed as "Dr_Confid" for all patients for privacy. |

Table 5.1: Descriptions of Asthma Dataset Columns

## 5.2 Diabetes Health Indicators Dataset

**Description:** The diabetes health indicators dataset focuses on diabetes prediction by incorporating a balanced set of health indicators collected from survey data.Each record represents a unique individual, with features that help predict diabetes risk and enable the model to generalize well to new cases.This dataset was downloaded from Kaggle and is available on the Kaggle website.[7]

**Link**: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

Below table represents important features of dataset such as:

| Variable | Description |
|---|---|
| Diabetes_binary | Indicates diabetes status (1 = Yes, 0 = No). |
| HighBP, HighChol | Binary indicators for high blood pressure and high cholesterol (1 = Yes, 0 = No). |
| CholCheck | Shows if cholesterol was recently checked (1 = Yes, 0 = No). |
| BMI | Ranges from 12 to 98 |
| Smoker | Smoking status (1 = Yes, 0 = No). |
| Stroke, HeartDiseaseorAttack | Indicate stroke and heart disease history (1 = Yes, 0 = No). |
| PhysActivity | Physical activity status (1 = Active, 0 = Not Active). |
| Fruits, Veggies | Regular consumption indicators for fruits and vegetables (1 = Yes, 0 = No). |
| HvyAlcoholConsump | Indicates heavy alcohol consumption (1 = Yes, 0 = No). |
| AnyHealthcare, NoDocbcCost | Shows healthcare access and cost barrier (1 = Yes, 0 = No). |
| GenHlth | Self-rated health on a scale from 1 (Excellent) to 5 (Poor). |
| MentHlth, PhysHlth | Days in the past month with poor mental or physical health. |
| DiffWalk | Difficulty walking or climbing stairs (1 = Yes, 0 = No). |
| Sex | Gender (0 = Female, 1 = Male). |
| Age | Age group, categorized by intervals. |
| Income | Income level on a scale from 1 to 8, with 1 = Less than $10,000, 5 = Less than $35,000, and 8 = $75,000 or more. |
| Education | Education level on a scale from 1 to 6, where 1 = Never attended school, 6 = College graduate. |

Table 5.2: Descriptions of Diabetes dataset columns

**Dimensions**: The dataset is named **diabetes_binary_5050split_health_indicators _BRFSS2015.csv**, containing a total of **70,692 rows** and **22 columns**. Every row in the dataset presents a profile of the health condition of a person.

## 5.3 Brain Stroke Dataset

**Description**: This dataset includes a broad set of demographic, medical, and lifestyle factors that could be used in predictive analytics of stroke risk. It covers important health conditions which it provides a person's life choices-smoking status hence, very relevant to building predictive models for stroke. This dataset combines those features with deep analysis and model training, enabling studies that achieve early identification of high-risk individuals.This dataset was downloaded from Kaggle and is available on the Kaggle website[35].

Link:https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset/data

Below table represents important features of dataset such as:

| **Variable** | **Description** |
|---|---|
| Gender | Patient's gender, with values "Male," "Female," or "Other." |
| Age | Age of the patient. |
| Hypertension | Indicates hypertension status; 0 if absent, 1 if present. |
| Heart Disease | Binary indicator for heart disease; 0 if absent, 1 if present. |
| Ever Married | Marital status; "No" or "Yes." |
| Work Type | Type of employment; options include "Children," "Govt job," "Never worked," "Private," or "Self-employed." |
| Residence Type | Categorized as "Rural"(49%) or "Urban"(51%) |
| Avg Glucose Level | Average glucose level in the blood. |
| BMI | Range from 14 to 48.9 |
| Smoking Status | Smoking history; categorized as "Formerly smoked," "Never smoked," "Smokes," or "Unknown." |
| Stroke | Target variable indicating stroke occurrence; 1 if the patient had a stroke, 0 if not. |

Table 5.3: Descriptions of Brain Stroke Dataset Columns

**Dimensions**:The dataset,named **brain_stroke.csv**, contains **4,337 rows** and **11 columns**. Each row is a unique patient entry. while each column captures information related to stroke risk, including both medical history and one on lifestyle factors of the patient.

In summary, the analysis datasets are complete, carrying all the attributes required for building predictive models on chronic diseases. Each of these subject datasets is demographically, lifestyle-wise, and clinically varied to capture appropriate factors associated with the risk related to asthma, stroke, and diabetes. Clean and pre-processed in structure, these datasets do not contain missing values and, therefore, are ready for exploration and analysis.

CHAPTER 6

# Methodology

## 6.1 Introduction to Methodology

This chapter explains the methodology for the prediction of asthma, diabetes, and brain stroke is represented through the flowchart.

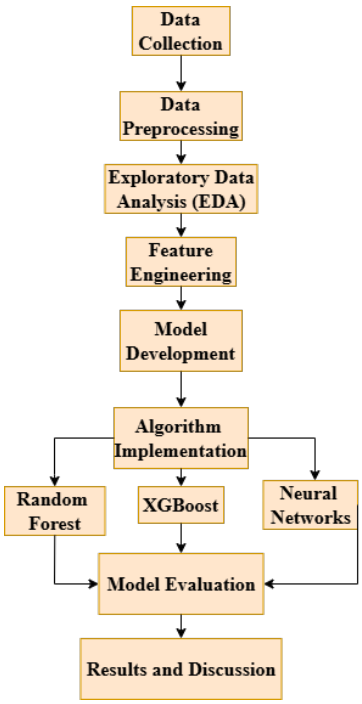


Figure 6.1: Flowchart of the Methodology

The steps include data collection, followed by preprocessing and feature engineering. Then, machine learning algorithms such as Random Forest, XGBoost, and Neural

Networks are used for developing predictive models. Model evaluation and discussing results to draw meaningful insights are the last steps.

## 6.2 Data Preprocessing

It gives an overview of preprocessing in asthma, diabetes, and brain stroke datasets. In general, this section describes cleaning, ensuring the consistency, and preparing the data for machine learning modeling.Proper preprocessing in health datasets has variability in data types, missing values, and outliers, which may affect model prediction performance. We discuss detailed handling of missed values, cleaning the data, encoding categorical variables, and scaling and normalization.

### 6.2.1 Handling Missing Values

One of the most important first steps in data preprocessing involves the handling of missing values. This is because, in any health care dataset, a single gap in data leads to biased predictions, hence reduced reliability in the model. Given the critical nature of the asthma, diabetes, and brain stroke datasets in this study, ensuring data completeness was a priority. By running .isnull().sum() on each dataset, it was able to confirm that no imputation or advanced handling strategies were necessary.

- **Asthma Dataset:** The Asthma dataset consists of 2,392 rows and 28 columns of various patient profile aspects: demographic, clinical, lifestyle, environmental, and symptomatic data. We checked the completeness of the data using **asthma_data.isnull().sum()**; the output for all columns was zero. It simply meant that each attribute, such as 'Age', 'BMI', 'PollutionExposure', and 'Diagnosis' , had a complete set of data entries. Since this dataset was complete, there was no need to apply imputation techniques, such as filling missing values with means or modes. This subsequently simplified the pre-processing steps in code, whereby we moved directly to other transformations like encoding and scaling without necessarily taking extra measures on missing data. This also gave a guarantee that the predictive model would work on the actual distribution of data, hence increasing the reliability of each feature concerning asthma risk predictions.

- **Diabetes Dataset:** The Diabetes dataset was larger: 70,692 records and 22 columns, including a wide range of health indicators like 'HighBP' and 'HighChol', which are indications of high blood pressure and cholesterol, 'BMI', respectively; and 'lifestyle' indicators of Smoking and 'PhysActivity'. We checked for missing values with "**diabetes_data.isnull().sum()**" on the dataset. This indeed confirmed zero missing values across all columns. Since all values were intact, no imputation code was needed. Neither were handling techniques necessary.With the features available, the model could interpret each health indicator completely. This is very important in any kind of chronic disease prediction model.

- **Brain Stroke Dataset:** The stroke dataset consists of 4,337 rows and 11 columns of data, each representing a certain attribute of a stroke. Some of the attributes of this dataset include 'Age', 'Gender', 'Hypertension', 'Heart Disease present', and 'Avg Glucose Level'. To verify if any values were missing in the columns of this dataset, the code **data.isnull().sum()** was run. Just like the asthma and diabetes datasets, the result indicated that there were no missing entries in all the columns of this dataset. No imputation was required since all the attributes were fully populated, and we could maintain the structure and values of the dataset intact. Generally, since we checked through code that there were no missing values in all three datasets using ".isnull().sum()",passing was achievable over the imputation step without data integrity compromise.

**This has a number of implications:**

1. **Streamlined Code:** With no need for missing value handling functions, our preprocessing steps are much simpler and more efficient to focus on other important transformations like encoding and scaling.

2. **Better Predictive Power:**Completeness in each dataset lets every feature contribute to its full predictive power without changes which may offset results.

3. **Better Reliability of Models:**Complete data ensures that models learn from actual real-world values. In this way, these models are prepared for developing reliable and generalizable predictions over asthma, diabetes, and brain stroke cases that are incident.

This preprocessing phase prepares the dataset in its original form without omitting gaps to create a robust foundation for the preparation and further analysis of data, which in turn enables the training of models that guarantee reliable results in the form of predictions related to chronic diseases.

### 6.2.2 Data Cleaning

- **Asthma Dataset:**The asthma dataset had very detailed patient information, comprising a total of 28 columns. We first cleaned up the data by removing all the irrelevant identifiers. For instance, the column named 'PatientID' was just an identifier with absolutely no predictive power. To clean up the dataset, it was dropped using asthma_data.drop(['PatientID'], axis=1). Checking for uniqueness in data: A method used here is 'duplicated().sum()'. Duplicate entries were removed so that each row represented a unique patient profile and maintained data integrity. Continuous variables such as 'BMI', 'PollutionExposure', and 'PollenExposure' were checked to find out if any of these values had been out of the expected range- a potential outlier. Fortunately, these values reflected a good correspondence with expected ranges in the population; hence, no further treatment for outliers would be considered at this stage. Up to this point, the asthma dataset had been preprocessed and standardized in a way to be ready for encoding and scaling.

- **Diabetes Dataset:**On the Diabetes dataset, with more than 70,000 rows, there is barely cleaning to be performed. As done in the asthma data, duplication was checked using the code '.duplicated().sum()'just to double-check that any redundancies would not affect a result. The most relevant indicators, such as 'BMI', 'Age', and 'PhysActivity', were verified by using '.describe()', with a regard for outliers, which all seemed within reason in a healthcare context. No extreme values were noticed; therefore, no further cleaning was necessary. Hence, this dataset needed very minimal changes other than those from initial checks. Since it was full and of high quality, we did more in the sense of encoding and scaling to ensure every record made unique contributions to model training.

- **Brain Stroke Dataset:**The brain stroke dataset, containing 4,337 rows, and it was more variable in terms of numerical data compared to others at columns such

as Age, Avg Glucose Level, and BMI. To ensure those values would not have an impact on performance, we resorted to the IQR method to handle outliers accordingly. Here is the IQR approach:

Each variable was computed the **Q1** and **Q3**, after which we calculated the IQR.

Any values falling were removed outside **Q1 - 1.5*IQR and Q3 + 1.5*IQR** to reduce the effect of extreme values on the model, especially for sensitive health indicator variables like glucose level and BMI.

Now the dataset was cleaned of outliers and duplicates, it was ready for encoding and scaling.

### 6.2.3 Encoding Categorical Variables

Encoding of categorical variables plays a very important role in machine learning, as it turns non-numeric data into a format that models can understand. Now, let's look into how encoding was performed on each dataset to make the models compatible and interpretable.

- **Asthma Dataset:** The asthma dataset included a few categorical variables such as 'Gender', 'Ethnicity', and 'EducationLevel', which were already encoded numerically. However, the 'DoctorInCharge' column required one-hot encoding due to its categorical nature, representing different doctors without implying any order. Here's how it was managed:

  1. **One-Hot Encoding:** Using **OneHotEncoder()**, we transformed 'DoctorInCharge' into multiple binary columns, each representing a unique doctor.
  2. **Combining Encoded Data:**These new columns (e.g., DoctorInCharge_0, DoctorInCharge_1) were merged back into the main dataset, while dropping the original 'DoctorInCharge' column to avoid redundancy.

  This approach preserved differences among doctors but without imposing any ordinal relationship such that it conformed to machine learning algorithms.

- **Diabetes Dataset:**In the diabetes dataset, categorical variables such as Sex, Education, and Income needed to be encoded differently:

1. **Ordinal Encoding:** Ordinal encoding has been used for variables that possess an inherent ordering, such as Education and Income, to reflect their hierarchy from lowest to highest.
2. **Binary Encoding:** Binary encoding was employed for the Sex variable, using 0 for Male and 1 for Female, for clarity and compatibility with ML models.

This careful encoding strategy was done so that the interpretation of demographic and socioeconomic data would be preserved while the information could be exploited by the algorithms.

- **Brain Stroke Dataset:**Encoding in categorical format was necessary for several features in the dataset brain stroke, namely 'Gender', 'Ever Married', 'Work Type', 'Residence Type', and 'Smoking Status'.We applied **LabelEncoder()** to convert each of these categories into numerical values. Here's a breakdown of the encoding scheme:

  1. **Gender:**Male (0), Female (1)
  2. **Work Type:** Children (0), Govt_job (1), Never worked (2), Private (3), Self-employed (4)
  3. **Smoking Status:**Never smoked (0), Former smoker (1), Smokes (2), Unknown (3)

  This encoding method allows the model to learn the differences in each category, hence capturing meaningful information in categorical variables of lifestyle and risk factors.

### 6.2.4 Scaling and Normalisation

Scaling and normalisation are important parts of the preparation of data for a machine learning model, where features are numerical in nature and span across different magnitudes. This will ensure that all the features around the model contribute equally to the performance of the model, without any large values from dominating the learning process of the model. The following section explains how the scaling and normalisation were carried out across these datasets.

- **Asthma Dataset:** First, continuous features in the asthma dataset were identified like 'Age', 'BMI', and 'Physical Activity'-measured on different scales. Using **StandardScaler()**,these features were transformed to have a mean of zero and a standard deviation of one.Scaling ensures that all features are comparable and do not disproportionately influence the model by default. This standardization of these variables helped in improving the stability of algorithms such as Random Forest, XGBoost, and Neural Networks, which need standardized input for better performance.

- **Diabetes Dataset:**In the diabetes dataset, features such as 'BMI', 'Age', 'Physical Health', and 'Mental Health' were identified as continuous variables requiring scaling. Initially, these features were standardized using **StandardScaler()** to bring them to a uniform scale. For the neural network model, an additional step of normalization using **MinMaxScaler()** was performed to map values into the 0-1 range. As a result, this normalization is very important in neural networks-it helps to make gradients stable during training and deficits smoother model convergence.

- **Brain Stroke Dataset:**In the case of the Brain Stroke Dataset, scaling on numerical variables such as 'Age', 'Average Glucose Level', and 'BMI' was performed by using **StandardScaler()**. These variables are highly variable and glucose levels and 'BMI' hold much importance in predictions related to stroke. For this, Standardization makes sure that this variation concerning scales does not affect distance-based models or calculation of gradient in algorithms related to XGBoost and Neural Networks. Additionally, it helped in scaling the model on diversified patient profiles.

## 6.3 Exploratory Data Analysis (EDA)

### 6.3.1 Descriptive Statistics

Descriptive statistics give a summary of datasets, highlighting key values such as the mean, standard deviation, and percentiles. They help in understanding the data's distribution and behavior, making it easier to identify patterns or outliers before moving

on to feature engineering and modeling.

- **Asthma Dataset:**The Asthma dataset includes 2,392 records, each capturing essential information about patient demographics, lifestyle, environmental exposure, and asthma symptoms. Below table focuses on variables most relevant to asthma prediction, providing a clear view of patient demographics, lifestyle factors, and symptom presence.

| Variable | Count | Mean | Std Dev | Min | 25th Percentile | Median | 75th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| Age | 2392 | 42.14 | 21.61 | 5.00 | 23.00 | 42.00 | 61.00 | 79.00 |
| Gender | 2392 | 0.49 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Ethnicity | 2392 | 0.67 | 0.99 | 0.00 | 0.00 | 0.00 | 1.00 | 3.00 |
| Education Level | 2392 | 1.31 | 0.90 | 0.00 | 1.00 | 1.00 | 2.00 | 3.00 |
| BMI | 2392 | 27.24 | 7.20 | 15.03 | 20.97 | 27.05 | 33.56 | 39.99 |
| Smoking | 2392 | 0.14 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Physical Activity | 2392 | 5.05 | 2.90 | 0.00 | 2.58 | 5.02 | 7.54 | 9.99 |
| Pollution Exposure | 2392 | 5.01 | 2.94 | 0.00 | 2.43 | 5.04 | 7.63 | 9.99 |
| Wheezing | 2392 | 0.60 | 0.49 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |

Table 6.1: Descriptive Statistics of Asthma Dataset

Here are some key takeaways from the descriptive statistics:

  - **Age:**The mean age of patients in this dataset is 42 years, with the minimum being 5 and the maximum being 79. This median value is quite close to the mean, which implies that this population of asthma patients tends to have a well-balanced age distribution.
  - **Gender:**The distribution of gender is almost equal, with the mean close to 0.5, where 0 most probably represents females and 1 represents males.
  - **BMI:**The mean of this variable is 27.24, ranging from 15 to almost 40. The shape of this distribution is telling that the asthma patients might have high values of BMI, which agrees with the general view that obesity may worsen asthma.
  - **Environmental Exposures (Pollution, Pollen, Dust):**These variables are all scaled with a mean of about 5 and range up to 10. The standardized

scale and relatively symmetric distribution suggest that exposure to these environmental factors could play a role in asthma severity.

– **Symptom-Related Variables (Wheezing, Shortness of Breath, Chest Tightness):**These are all binary variables and have an approximate mean of 0.5, indicating that about half of the patients report these symptoms.That is a good balance, since for modeling, the data becomes useful as there is a fair representation of symptomatic and asymptomatic patients.

Overall, these data support the hypothesis that factors related to Age, BMI, and Exposures to Environmental causes, as well as specific symptoms, are important in understanding and assessing asthma severity and prevalence.

- **Diabetes Dataset:**The Diabetes dataset is significantly larger, with 70,692 entries, and includes a wide range of health and lifestyle indicators. The values within the table below show some critical health indicators for diabetes, including lifestyle factors such as physical activity and general health ratings that contribute to understanding diabetes risk.

| Variable | Count | Mean | Std Dev | Min | 25th Percentile | Median | 75th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| Diabetes (Binary) | 70692 | 0.50 | 0.50 | 0.00 | 0.00 | 0.50 | 1.00 | 1.00 |
| High BP | 70692 | 0.56 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| High Cholesterol | 70692 | 0.53 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| BMI | 70692 | 29.86 | 7.11 | 12.00 | 25.00 | 29.00 | 33.00 | 98.00 |
| Physical Activity | 70692 | 0.70 | 0.46 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| General Health Rating | 70692 | 2.84 | 1.11 | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 |
| Mental Health | 70692 | 3.75 | 8.16 | 0.00 | 0.00 | 0.00 | 2.00 | 30.00 |
| Physical Health | 70692 | 5.81 | 10.06 | 0.00 | 0.00 | 0.00 | 6.00 | 30.00 |

Table 6.2: Descriptive Statistics of Diabetes Dataset

Here's what stands out in the descriptive statistics:

– **Diabetes (binary):**This is the target variable that shows the presence of diabetes. The mean value is 0.5, indicating there is an equal split between diabetic and nondiabetic. That is very good in terms of constructing a classifier; this does not suffer from class imbalance.

- **Blood Pressure (HighBP) and Cholesterol (HighChol):** Both variables have means above 0.5, suggesting that a significant portion of the population has hypertension and high cholesterol these two common risk factors for diabetes.
- **BMI:** The mean for BMI is approximately 29.85, ranging between a minimum of 13.35 and an extreme of 98. This broad range covers a wide variety of body weights, including cases of severe obesity, which can heavily impact diabetes risk and model predictions.
- **Lifestyle Factors (Physical Activity, Smoking, Fruit and Vegetable Intake):**The mean is 0.7, relatively high to express that most of the people are doing some kind of exercise. Almost half of the population reported a mean smoking history of 0.47, which complicates diabetes outcomes.
- **General Health:** This is a variable ranging between 1-excellent and 5-poor whose mean is roughly 2.8. Most people have rated health average to poor. This might perhaps be taken to bear some relation to the risk for diabetes and perhaps might prove a good predictor for it.

These pieces of information point toward BMI, lifestyle habits, blood pressure, cholesterol, and general health as major criteria to predict diabetes. Diversity in these variables in the dataset will definitely help in building a robust model.

- **Brain Stroke Dataset:** This Brain Stroke dataset contains 4,337 entries for health factors related to stroke risk. The table below confines to the variables of interest regarding stroke conditions, including age, hypertension, heart disease, BMI, and glucose level-all being major determinants of stroke conditions.

| Variable | Count | Mean | Std Dev | Min | 25th Percentile | Median | 75th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| Age | 4337 | 41.13 | 22.50 | 0.08 | 23.00 | 42.00 | 58.00 | 82.00 |
| Hypertension | 4337 | 0.07 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Heart Disease | 4337 | 0.04 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| BMI | 4337 | 27.80 | 6.46 | 14.00 | 23.20 | 27.60 | 31.90 | 45.20 |
| Avg Glucose Level | 4337 | 91.45 | 22.60 | 55.12 | 75.08 | 88.10 | 104.00 | 168.68 |
| Smoking Status | 4337 | 1.35 | 1.08 | 0.00 | 0.00 | 2.00 | 2.00 | 3.00 |
| Stroke | 4337 | 0.04 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

Table 6.3: Descriptive Statistics of Brain Stroke Dataset

Here's what we can learn from the descriptive stats:

- **Age:** The average age is around 41, with values ranging from as young as 0.08 years (possibly representing infants) to 82 years old. This is a wide range and it gives us the opportunity to study stroke risk across groups of age, from very small children up to elderly subjects.
- **Hypertension and Heart Disease:** Both are binary variables, with relatively low means of 0.07 for hypertension and 0.04 for heart disease. Although these conditions are less common in the dataset, they are known to be significant risk factors for stroke.
- **BMI and Glucose:** The average for BMI is about 27.8, ranging from 14 to 45.2; hence, this population covers a wide range of body types. Glucose has an average value of about 91.4, although the dispersion is remarkably high. High glucose level is one of the most critical risk factors for stroke, this variability could provide valuable insights.
- **Smoking Status:** With a mean of 1.35, this variable might represent the intensity or frequency of smoking. Smoking is a major risk factor for stroke, and understanding its role across different smoking levels is essential for our analysis.

### 6.3.2 Visualization Techniques

This section highlights various ways in which the Asthma, Diabetes, and Brain Stroke dataset visualizations were used to bring out insights.These are selected carefully so as to feature critical aspects of the data to be used in developing machine learning models.

- **Asthma Dataset: Key Visual Insights**

  a) Bar Plots of Age and BMI Distribution

  - **Bar Plot of Age**

    This age distribution plot shows the age of asthma patients, ranging from about 5 to 80 years. Each bar represents the number of patients falling within a certain age group, and most of these age groups contain over 100 patients. It can be observed that the bracket of 60-70 age group contains a

little over 130 to 140 patients, which may indicate that more aged adults suffer from asthma or symptoms of asthma. Inclusion of younger and older age groups points to the occurrence of asthma at any age. This breakdown will help us understand how asthma cases compare across different ages, which is helpful for health insights and support that are focused on age.

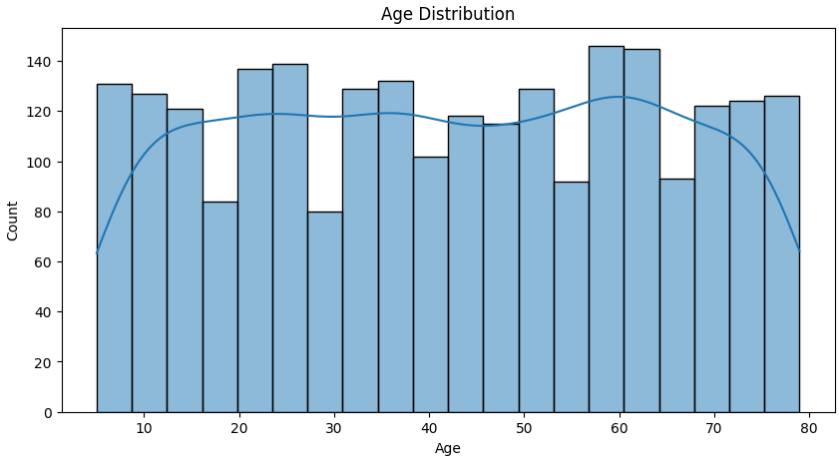


Figure 6.2: Distribution of Age

– **Bar Plot of BMI**

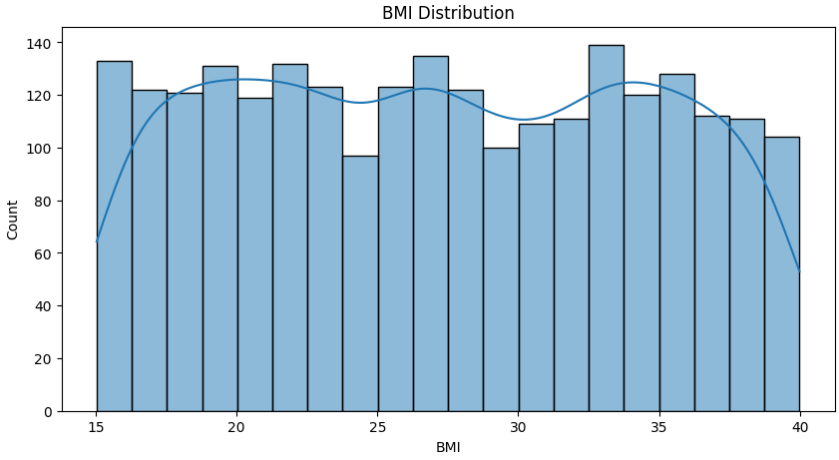


Figure 6.3: Distribution of BMI

The distribution plot of 'BMI' describes the dispersion of Body Mass Index for asthma patients present in this dataset. Values of 'BMI' span from approximately 15 to 40, indicating a wide variation in the range of body types. Bars show specific ranges of BMI and the height of each

bar represents the count of patients within that specific range. Most of the 'BMI' groups have more than 100 persons, and slight peaks are seen around the ranges of 25-30 and 35-40. Therefore, these levels of 'BMI' were the most common in asthma patients in this dataset. This distribution helps identify how body weight-activities related to 'BMI', may relate to symptoms or risk factors of asthma across diverse individuals. Understanding the pattern of 'BMI' may help in investigating the relationship between weight management and asthma management.

b) **Bar Plots for Smoking Status Across Age Groups**

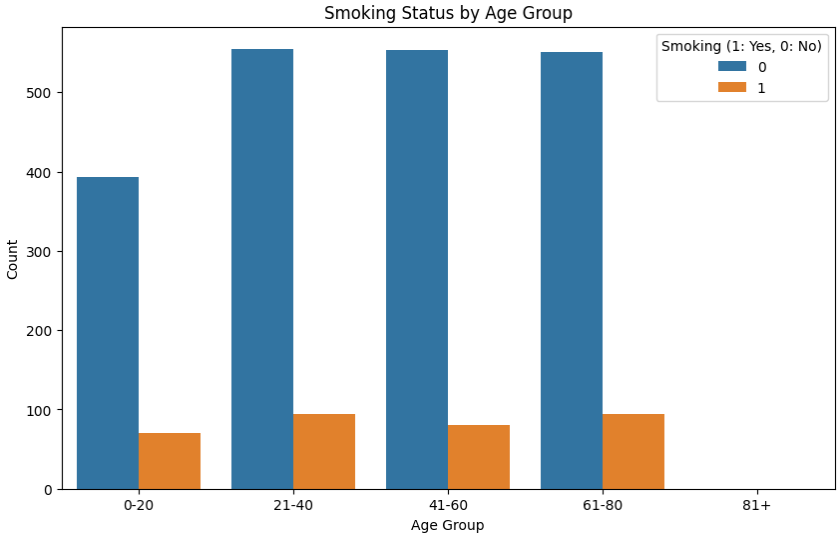


Figure 6.4: Smoking Status by Age Group

This graph describes the smoking status of asthma patients in different age groups. Blue bars indicate non-smokers, while smokers are represented by orange-colored bars. In all the age groups considered, the number of non-smokers is very high compared to that of smokers. For example, in the age groups 21-40 and 41-60, over 500 are non-smokers, whereas smokers are less than 100. This shows that smoking among asthma patients here is rare, which may indicate that other factors, such as environmental ones or genetic pre-disposition, may be more influential in these cases This understanding can assist healthcare providers in prioritizing non-smoking-related triggers when advising on asthma management strategies.

c) **Pie Charts for Gender Distribution in Asthma Dataset**

This pie chart represents the distribution of gender in the asthma dataset, in which 50.7% are male, while 49.3% are female The near-equal split between genders suggests that asthma affects both men and women at comparable rates in this group. By investigating such an even representation, we are able to review gender-related patterns in asthma symptoms and triggers in order to present a well-rounded perspective on how asthma presents itself across male and female patients. This is important in terms of showing both genders in balance to show how each gender may be similarly or differently affected concerning symptoms or lifestyle impacts.

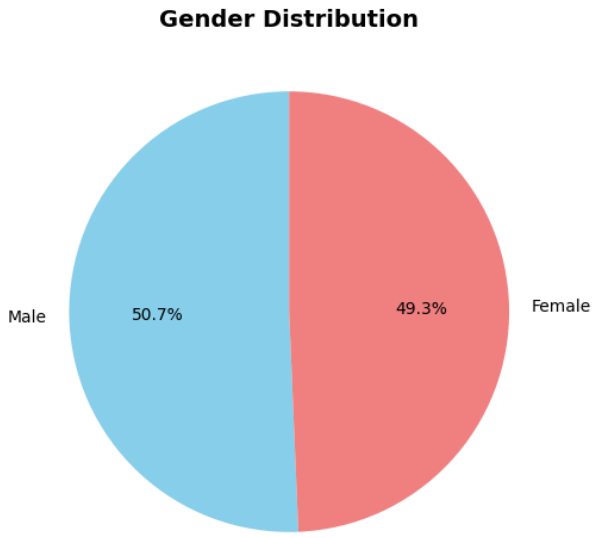


Figure 6.5: Gender Distribution in Asthma Dataset

d) **Box Plot for Pollution Exposure by Asthma Diagnosis**

This box plot shows that the subjects with asthma, indicated by "1", and without asthma, indicated by "0", have similar levels of exposure to pollution, coming to a median score of approximately 5 in both groups. It reflects a consistent range of exposures in this dataset, most of them falling between 2 and 8, indicating even distribution among the population. In fact, there is no difference in the exposure of pollution between asthmatic and non-asthmatic subjects. This may indicate that while both cases are indeed exposed to pollution, it is not distinctly higher in asthma patients. This plot gives us insight into the distribution of pollution exposure across different health

statuses, enabling a better understanding of how widespread pollution levels impact individuals in this dataset.

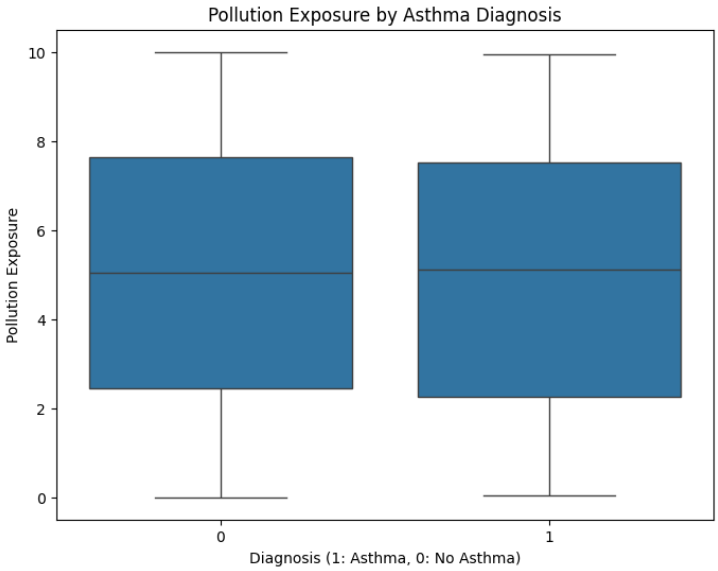


Figure 6.6: Pollution Exposure by Asthma Diagnosis

e) **Pie Chart Analysis of Asthma-Related Symptoms**

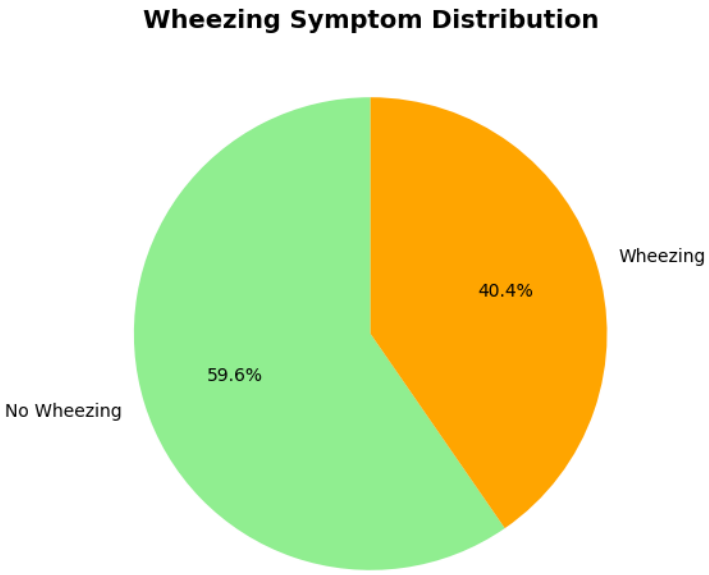


Figure 6.7: Wheezing Symptom Distribution

The resulting dataset contains a total of 2,392 entries, where 40.4% of the people wheeze and 59.6% do not. Wheezing is one of the more common

respiratory symptoms associated with asthma, and it is seen herein to affect a majority of the subjects under study, which goes a long way in describing just how prevalent this symptom will be across the group. This may suggest that wheezing is a cardinal symptom of respiratory issues possibly related to asthma, in order for us to understand the symptom's prevalence and its value as an identifier of persons with potential risk.

- **Diabetes Dataset: Key Visual Insight**

  a) **Heatmap Analysis of Health Correlations**

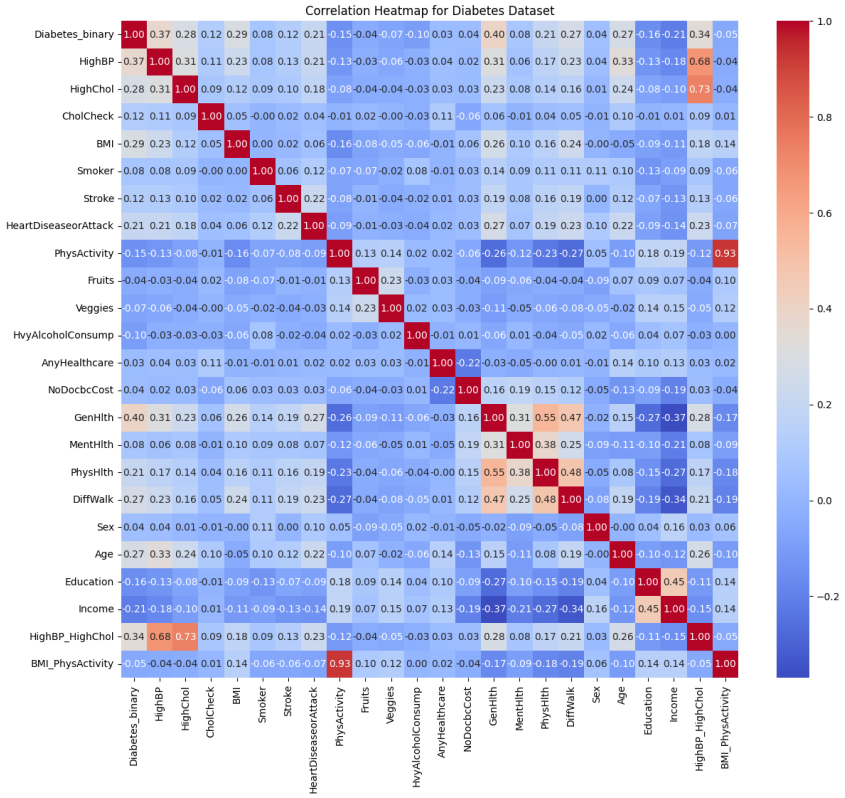


Figure 6.8: Correlation Heatmap for Diabetes Dataset

This is a correlation heatmap of different health and lifestyle factors. Each square calculates the relation between two variables; color varies from dark red for strong positive correlation to dark blue for strong negative correlation. For example, there's a high positive correlation between 'HighBP' and 'HighChol', meaning people with high blood pressure often have high cholesterol as well. Similarly, 'PhysHlth' and 'MentHlth' are moderately correlated, which could suggest that poorer physical health is associated with poorer

mental health. This visual helps in finding out how some factors of health, such as blood pressure and cholesterol, may commonly appear together and provide insight into the relationship of various health variables.

b) **Distribution of Key Health Indicators**

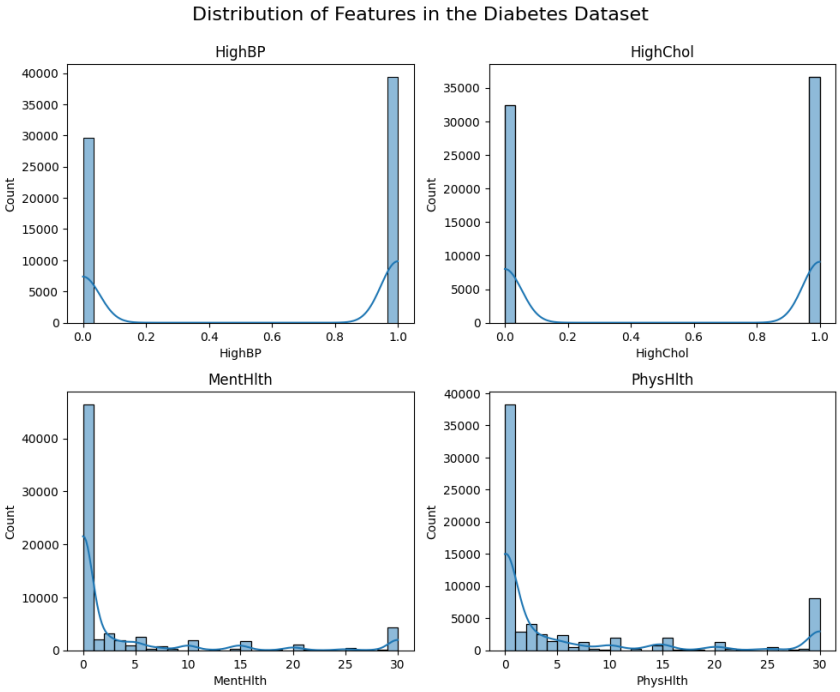


Figure 6.9: Distribution of HighBP, HighChol, MentHlth, and PhysHlth

These plots outline the distribution of four health features: 'HighBP'-high blood pressure, 'HighChol'-high cholesterol, 'MentHlth'-mental health, and 'PhysHlth'-physical health. HighBP and HighChol are binary, showing that individuals are either within normal or high levels, with a significant portion having elevated levels. In the 'MentHlth' and 'PhysHlth' plots, most people report few days of poor mental or physical health, but there are some who experience up to 30 days, indicating chronic health challenges. These distributions taken together suggest both stable and recurring health issues within the population.

c) **Age Distribution by Diabetes Status**

This plot represents the distribution of age amongst persons with and without diabetes. The x-axis is age, and the y-axis is density. The orange line on the chart represents persons with diabetes, while the blue line represents persons

without diabetes. From this we could infer that there is a higher prevalence of diabetes for middle aged and older citizens; it is apparent from the graph that the orange line has a very high peak between ages 8 to 12 on this scale, or ages 40 to 60 in real terms. The density for non-diabetics, meanwhile, is flatter, with smaller peaks for a greater range of ages. This may reflect that diabetes is generally in the middle-aged to older age group, which gives a clearer view of the effects of age on the status of diabetes within the population.

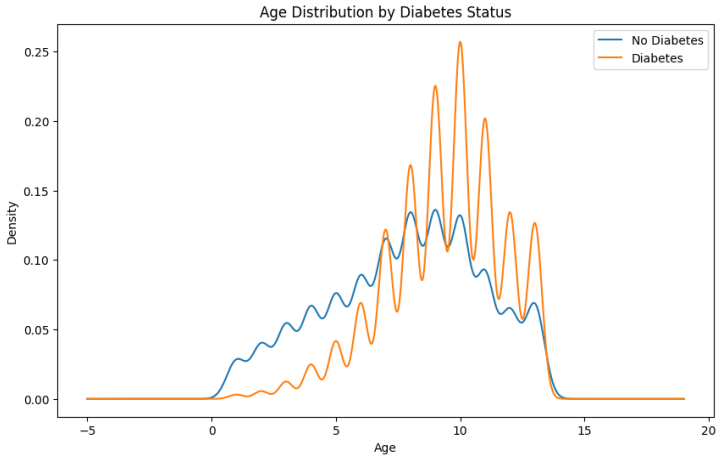


Figure 6.10: Age Distribution by Diabetes Status

d) **Physical Activity Levels Among Individuals with Diabetes**

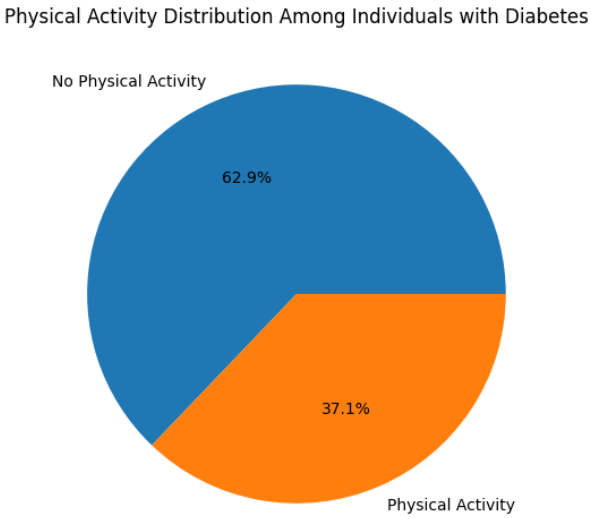


Figure 6.11: Physical Activity Distribution Among Individuals with Diabetes

This pie chart depicts the percentage of diabetic respondents who are physi-

cally active and those who are not; 62.9% reported not to be physically active, and 37.1% reported being physically active. That means that the majority in this population are inactive, hence the need to understand the role of lifestyle behavior such as exercise in maintaining good health with diabetes.

e) **Diabetes Prevalence by General Health Rating**

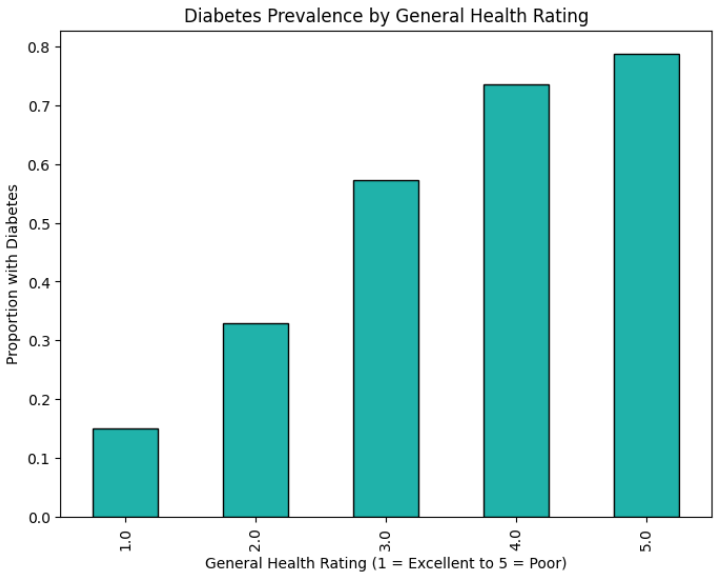


Figure 6.12: Diabetes Prevalence by General Health Rating

This bar chart illustrates the relationship between general health ratings and diabetes prevalence. The health rating scale ranges from 1 (excellent) to 5 (poor), with each bar showing the proportion of people with diabetes at each rating level. As general health ratings decline, the prevalence of diabetes increases significantly. Individuals with an excellent health rating (1) have a relatively low rate of diabetes, while those with poorer health ratings (4 and 5) show a much higher prevalence. This pattern highlights how lower self-rated health is closely linked with a higher likelihood of diabetes, suggesting that general health perception may reflect underlying risks associated with diabetes.

- **Brain Stroke Dataset: Key Visual Insight**

  a) **Heatmap Analysis of Feature Correlations**

  Correlation heatmap that visualizes the relationship between various features of brain stroke. Each cell in this heatmap has information about the

strength and direction of the relationship between two variables. Darker red represents a high positive correlation, whereas darker shades of blue depict negative correlation. Such a good number of positive correlations exist from age to variables such as "ever_married" of (around 0.68), which could mean increased likelihood for older persons to have been married. Also, 'Age' and 'Hypertension' are moderately positively-correlated, supporting the fact that with increasing age, a tendency toward hypertension, indeed, does occur. Understanding these correlations is essential for identifying risk factors associated with stroke, as factors like age, hypertension, and lifestyle choices may influence the chances of having a stroke

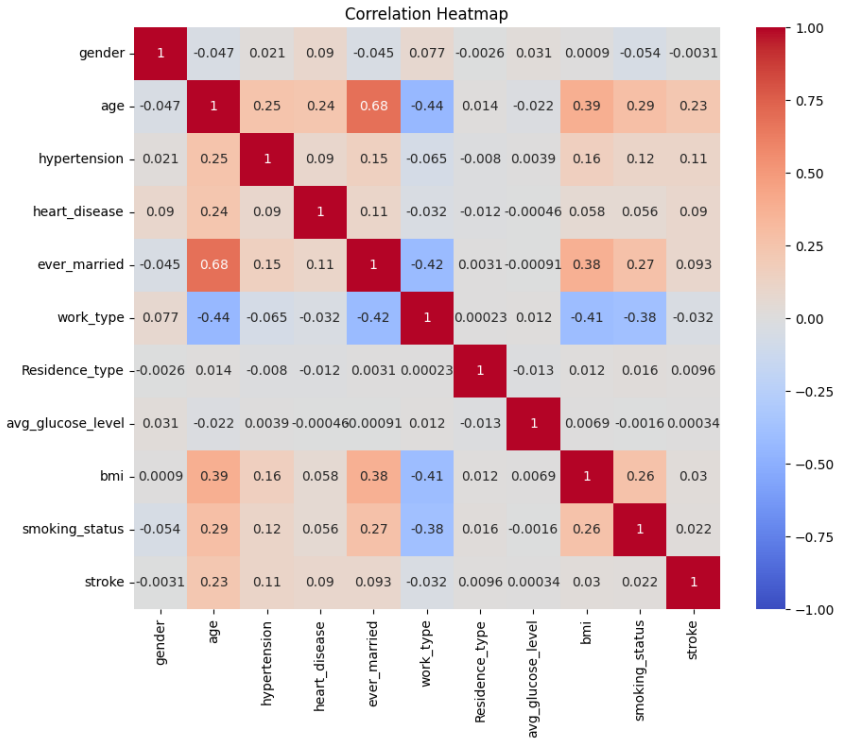


Figure 6.13: Correlation Heatmap for Brain Stroke Features

b) **Distribution of Key Health Features**

This figure shows the distribution of key features relevant to health and stroke risk. The age plot centers around ages 50-70, again reflecting increased stroke risk in older adults. Hypertension and heart disease have a small presence but are significant risk factors, as people with these conditions are more prone to stroke, giving the model a clear pattern to recognize in predictions. The status of smoking is divided into four categories: 0- never

smoked, 1- formerly smoked, 2- occasional smoker, and 3- regular smoker. This breakdown helps assess different smoking levels impact on stroke risk. Besides this, the distribution of average of glucose and BMI will be able to depict how high results lead to stroke; high glucose and high BMI usually correspond to diseases like diabetes and obesity. Based on such trends, the model can more correctly identify those people who are at risk due to a combination of these factors.

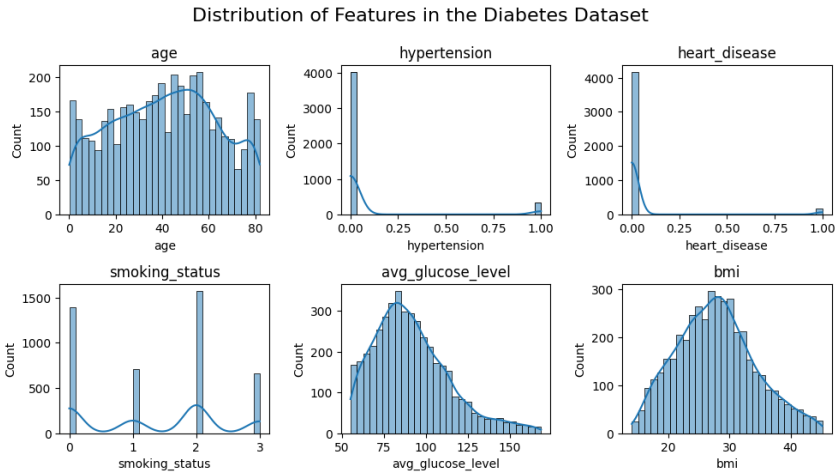


Figure 6.14: Distribution of Health-Related Features by Counts

c) **Bar Plots for Work Type Distribution**

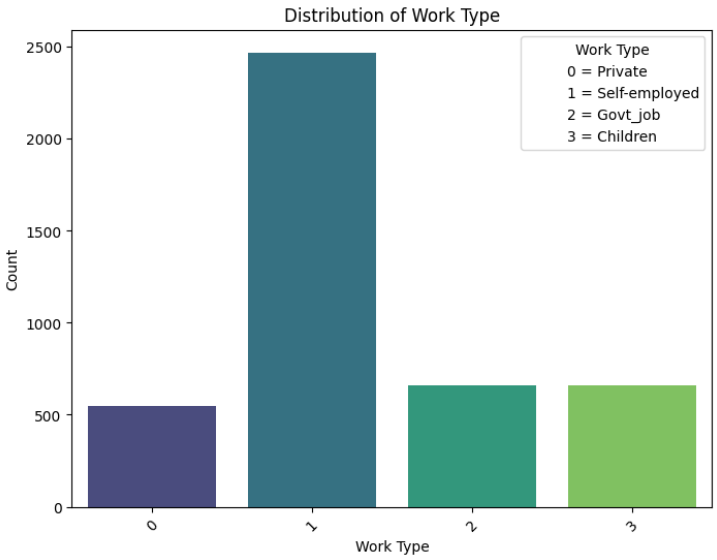


Figure 6.15: Distribution of Work Type Among Individuals

This bar plot shows the distribution of work types among individuals, cate-

gorized as private (0), self-employed (1), government job (2), and children (3). This breakdown provides a view of different occupational backgrounds and can help examine if any specific work type correlates with stroke risk. For example, related job factors, such as stress or the particular lifestyle habits of a certain job, could be associated with stroke prevalence in certain groups.

d) **Pie Chart for Residence Type Distribution**

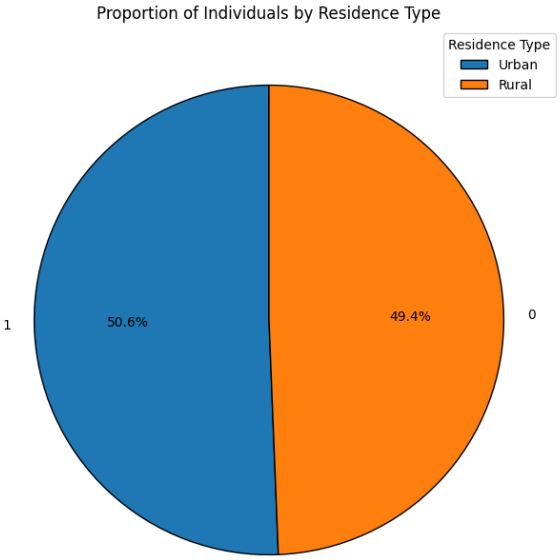


Figure 6.16: Proportion of Individuals by Residence Type

The graph shows that there is a split, between people residing in urban and rural areas. With 50.6% in urban areas and 49.4% in rural areas. Indicating a fair representation of both settings. This helps in bringing out the analysis on how these living conditions might affect the brain stroke related factors. Both contain equal numbers for each type of residence that will put us in a better position to find the differences in stroke and health outcomes between urban and rural living.

e) **Box Plot Analysis for BMI and Stroke Status**

The box plot compares Body Mass Index (BMI) between individuals with and without a stroke history, where "0" refers to no stroke and "1" refers to subjects who have had a stroke. For no stroke, the values of the 'BMI' range from about 15 to above 45 with a median around 27. Those with a stroke

history have a slightly narrower 'BMI' range, generally between 20 and 40, with a median close to 30. We can see some few outliers in the stroke group, where 'BMI' values are above the typical range. This plot helps us to see that higher values of 'BMI' are common in both groups but are somewhat more concentrated around the median for stroke patients. This might suggest that the 'BMI' can play a part in understanding stroke risk, at least in so far as high levels of 'BMI' are often linked to health conditions which contribute to stroke risk.

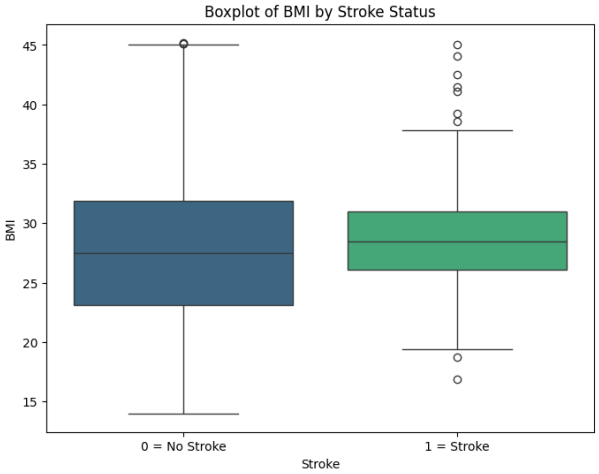


Figure 6.17: Boxplot of BMI by Stroke Status

## 6.4 Feature Engineering

In this section, we explore various feature engineering techniques implemented to improve model performance. Each dataset (Asthma, Diabetes, and Brain Stroke) required tailored transformations and enhancements to handle class imbalances, scale features, and refine the data.

### 6.4.1 Handling Class Imbalance with SMOTE

**How SMOTE Works:** The SMOTE algorithm works by interpolating new instances between existing instances in the minority class. It takes a given instance in the minority class, selects several of its nearest neighbors, and creates synthetic data points along the line segment joining the instance to its neighbors. In such a manner, the representation

of the minority class is enhanced without simply duplication of the instances, and thus, the model learns from a more balanced distribution of outcomes.[29]

The major challenge in the prediction of disease outcomes is dealing with imbalanced classes. This can occur under conditions where the target variable contains a large number of instances falling into one class compared with another, such as non-diabetic versus diabetic.

**Implementation in Our Datasets:**

1. Different splitting of X- features and y- target variable have been done using different splitting options for the Asthma, Diabetes, and Brain Stroke datasets, and then SMOTE was applied on the same target variables such as Diagnosis, Diabetes_binary, and stroke respectively in the mentioned datasets.

2. SMOTE balances the dataset, hence it ensures that our models are exposed to relatively equal spreads of positive and negative cases, which further reduces the chances of any bias toward the majority class. This enhances model accuracy and makes model predictions more reliable considering both classes.

### 6.4.2 Dimensionality Reduction with PCA

In some datasets, especially after encoding categorical features, we have a very large number of features.To handle this high dimensionality and prevent overfitting, **Principal Component Analysis (PCA)** was used in the feature engineering pipeline for some of the models. PCA decreases the number of features by mapping features into a new feature space while retaining the most important variance. This step of dimensionality reduction simplifies the model and enhances the efficiency of computation significantly[18].

In the case of the Diabetes dataset, the Random Forest pipeline used PCA to retain 95% of the variance, thus effectively reducing the feature space while retaining most of the information. This dimensional reduction not only prevents overfitting but also makes models more interpretable by focusing on the most significant parts of the data.

### 6.4.3 Feature Scaling and the Use of Pipelines

Feature scaling is one of the most important processes in data preparation, especially for those algorithms that are sensitive to the magnitude of features such as Neural

Networks, XGBoost, and Random Forest. We standardized the data using Standard Scaling by transforming each feature with mean zero and standard deviation of one.

To streamline the feature engineering process and ensure consistent application of transformations across training and testing data, we implemented Pipelines A pipeline allows us to combine of many steps, which include scaling, dimensionality reduction, and model training into one predictable and reproducible process, so as to emphasize how every similar makeover can be easily applied to both training and testing datasets[2].

Pipelines were created for the Diabetes and Brain Stroke data sets that involved feature scaling with PCA and model training. This helps in minimizing errors by including all the proper transformations consistently for the testing and training data, hence making the models more stable and accurate.

### 6.4.4 Disease-Specific Feature Engineering

For each dataset, some feature engineering was done to enhance the model's performance and give more reliable predictions. We used one-hot encoding on the variable 'DoctorInCharge' within the **Asthma dataset**; each unique doctor would become a separate binary feature, allowing the model to understand this categorical information without suggesting any particular order among the doctors. Also, standardized scaling techniques were used for continuous features like 'Age', 'BMI', 'PhysicalActivity'; these were on the same scale to make sure that each feature has an equal contribution to the model.

In the **Diabetes dataset**, we created interaction terms to capture combined health effects, such as 'HighBP_HighChol' (combining high blood pressure and high cholesterol) and 'BMI_PhysActivity' (combining BMI and physical activity). These new features help the model understand how these factors might work together to influence diabetes risk. Additionally, standard scaling applied to features like 'BMI', 'GenHlth', and 'PhysHlth' because scaling is especially helpful for models like neural networks and XGBoost, which work better with standardized data.

On the **Brain Stroke dataset**, we performed outlier treatment. To remove extreme values on features 'age','avg_glucose_level', and 'bmi' using the IQR method to avoid skewing the model with abnormal values. After balancing the data with SMOTE,

continuous features scaled to maintain them all in the same page, helping the model perform better.

## 6.5 Development and Training of Models

### 6.5.1 Train-Test Split of Dataset

The training set is used to help the model learn patterns from the data, while the testing set evaluates the model's ability to make accurate predictions on unseen data. Training data for Asthma, Diabetes, and Brain Stroke consists of an 80-20 split for all three disease datasets. This means that for this purpose, 80% of the data is needed to train the model, while the remaining data will be used to test the performance unbiasedly.

The splitting process involved three key steps. First, we separated the features(X) from the target variable(y). Next, applied SMOTE on the target variable to balance the class distribution in cases where there was significant class imbalance. Finally, the dataset split into training and testing sets using the '**train_test_split**' function with a random seed for reproducibility.

### 6.5.2 Deployment of Machine Learning Algorithms

The data was divided into training and testing sets, three machine learning models were trained and evaluated for each dataset: Random Forest, XGBoost, and Neural Networks. The models identified key features that were most important for predictions in each disease. For **asthma**, factors like 'pollution exposure' and 'family history' were identified as significant predictors. For **diabetes**, 'BMI', 'physical activity', and 'high blood pressure' played critical roles, while for **brain stroke**, 'age' and 'glucose levels' stood out as the most influential features.

1. **Random Forest Classifier** Random Forest is a powerful machine learning algorithm that improves prediction accuracy by combining the results of multiple decision trees. It is particularly good at handling large datasets, dealing with missing data, and avoiding overfitting. The way it works is by creating many decision trees during training. Each tree is built using a random sample of the data (a process called bootstrap sampling) and by randomly selecting a subset of

features to consider when making splits in the tree.[20]

When making predictions, each tree in the Random Forest gives its own output, essentially 'voting' for a class. The final prediction is based on '**majority voting**', meaning the class that most trees agree on becomes the final result. This process helps improve accuracy and stability because it combines the strengths of multiple trees. The diagram visually represents how each tree makes its prediction, and these individual predictions are combined to decide the final class.

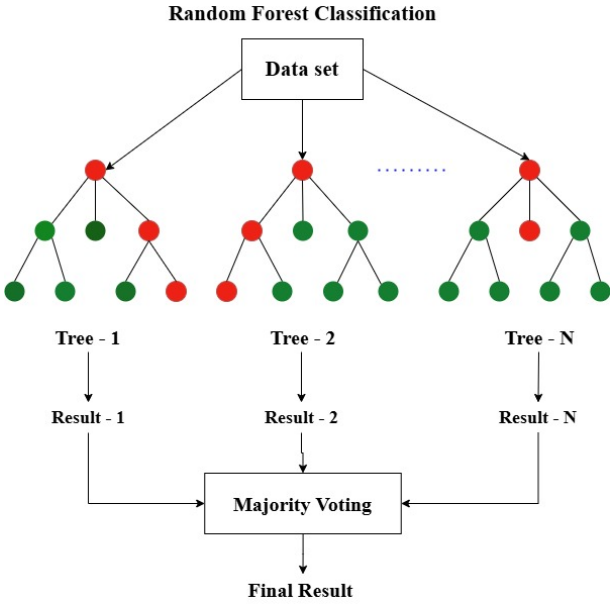


Figure 6.18: Random Forest Decision Tree Workflow

[17]

**Gini impurity** is a metric used in Random Forest to measure how 'pure' a node is during the tree-building process. It calculates the likelihood of incorrectly classifying a randomly chosen element if its class label is assigned based on the distribution of classes in that node. A Gini score of '0' means the node is perfectly pure, meaning all elements in the node belong to the same class.[40]

The **Gini Index** is employed to measure node impurity. It is calculated using the formula:

$$G = 1 - \sum_{i=1}^{c} p_i^2$$

[32]

Here, $p_i$ represents the proportion of samples in class 'i' , and 'c' is the total number of classes. A lower Gini Index indicates purer splits, meaning the node predominantly contains samples from a single class. This helps the model make better decisions as it grows the trees.

Random Forest was implemented using the **RandomForestClassifier** from the sklearn library to predict asthma, diabetes, and brain stroke.To optimize its performance, key parameters were adjusted, such as setting the number of trees (n_estimators) to 100 for a balance between accuracy and computational efficiency. The maximum tree depth **(max_depth)** was limited to prevent overfitting while capturing essential patterns. Gini impurity was used to decide the best splits, ensuring purer nodes. Parameters like **min_samples_split** and **min_samples_leaf** were tuned to refine the decision-making process, controlling how many samples were required for splits or leaf formation. These optimizations helped the model perform reliably and accurately.

2. **XGBoost for Classifier**

   XGBoost, which stands for Extreme Gradient Boosting, is a powerful algorithm widely used for classification tasks because of its speed and high performance. It works by building multiple decision trees one after another, where each tree tries to fix the mistakes made by the previous ones. This approach, called **boosting**, allows XGBoost to handle complex and noisy data effectively. Its ability to process high-dimensional datasets quickly makes it a favorite choice for many machine learning tasks.

   XGBoost starts with a simple prediction, like the average for regression or initial probabilities for classification. It then calculates the errors (called residuals) between these predictions and the actual values. A new tree is built to learn from these errors and improve the predictions. This process is repeated, with each tree focusing on minimizing the mistakes of the earlier trees. XGBoost also uses techniques like regularization to avoid overfitting, handles missing values seamlessly, and supports custom loss functions, making it both flexible and reliable for various datasets.[5]

   **XGBoost Classification Formulas:**

a) **Objective Function:**

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t(x_i) + \Omega(f_t)$$

[27]

Here's what each term represents:

- $L^{(t)}$: Represents the overall loss at the $t^{\text{th}}$ iteration.
- $\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right)$: This term calculates the sum of the loss function $l$ (e.g., log loss or mean squared error) between the true label $y_i$ and the predicted value $\hat{y}_i^{(t-1)}$ from the previous iteration.
- $f_t(x_i)$: Represents the prediction from the current tree for instance $i$. This is the incremental improvement added by the new tree.
- $\Omega(f_t)$: This regularization term controls the complexity of the tree $f_t$. It prevents overfitting by penalizing overly complex trees.

b) **Logistic Loss Function (for binary classification):**

$$l(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

[31]

Here:

- $y_i$ **:** The true label for the i-th instance (either 0 or 1 for binary classification).
- $\hat{y}_i$ **:** The predicted probability for the i-th instance belonging to class 1.
- $n$ **:** The total number of instances in the dataset.
- $\log(\hat{y}_i)$ **:** The natural logarithm of the predicted probability for the positive class.
- $1 - y_i$ **:** Represents the complement of the true label (used for the negative class).
- $1 - \hat{y}_i$ **:** Represents the complement of the predicted probability.

This function calculates the error between the predicted probabilities ($\hat{y}$ ) and the actual labels ($y$), penalizing incorrect predictions. Smaller values of l(y,$\hat{y}$)

indicate better alignment between the model's predictions and the actual outcomes.

c) **Weight Update Rule:**

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

[27]

Here:

- $\tilde{L}^{(t)}(q)$: This represents the total loss at the current iteration $t$, considering the structure of the tree and the weights assigned to its leaf nodes.
- $\sum_{j=1}^{T}$: This is the sum over all leaf nodes in the tree, where $T$ is the total number of leaves.
- $\sum_{i\in I_j} g_i$: This is the total gradient for all the data points that fall into a specific leaf $j$. It shows how much the model's prediction for that leaf needs to change to reduce the error.
- $\sum_{i\in I_j} h_i + \lambda$: This is the total second-order derivative (Hessian) for the points in leaf $j$, plus a regularization parameter $\lambda$. Adding $\lambda$ helps prevent making overly large adjustments to the leaf's prediction and keeps the model stable.
- $\gamma T$: This is a penalty term based on the total number of leaves $T$. It discourages the tree from growing too many leaves, which helps reduce the risk of overfitting.

This objective function combines the loss function that ensures the model predicts as close to the actual target and the regularization term to discourage overfitting by not letting overly complex tree structures set in. In the case of classification, logistic loss function is used very much to find a gap between probabilities predicted and actual classification. The weight update rule refines predictions coming from each tree using gradients and Hessians, which by their shape show how adjustments should be fine-tuned in the light of the shape of the loss function.

3. **Neural Networks (NN) for Classification :**

Neural Network architecture and functionality are inspired by the human brain, where several interconnected neurons are involved in the processing and transmission of information. It consists of layers of nodes (neurons) that include an input layer, one or more hidden layers, and an output layer that generally work together to transform inputs into meaningful predictions. Neural Networks are very good at modeling intricate patterns and relationships within data, thus being very appropriate for classification tasks[4].

**How Neural Networks Work:**A neural network is composed of layers: an input layer, one or more-hidden layers, and an output layer. Each of these layers is made up of nodes or neurons that are interconnected with each other to form a network. These interconnections between the nodes have weights, adjusted during training through an optimization algorithm that minimizes the prediction error, hence the network learns from this and is able to predict more accurately.

- **Forward Propagation:** The information moves from the input layer through the hidden layers to the output layer. Each neuron will process the data through multiplication by weights, adding a bias term, and then applying some activation function like ReLU or sigmoid. This provides the network with the capability of transforming input into meaningful output by capturing complex relationships[21].

- **Backward Propagation:** It is the step where weights and biases of the network get updated to minimize the loss. The direction and amount of weight adjustment depend on the gradients, calculated through algorithms like gradient descent. It is an iterative process, ensuring that the model learns from it with each training cycle[21].

Below is a visual representation of a simple Architecture of Artificial Neural Network classification:

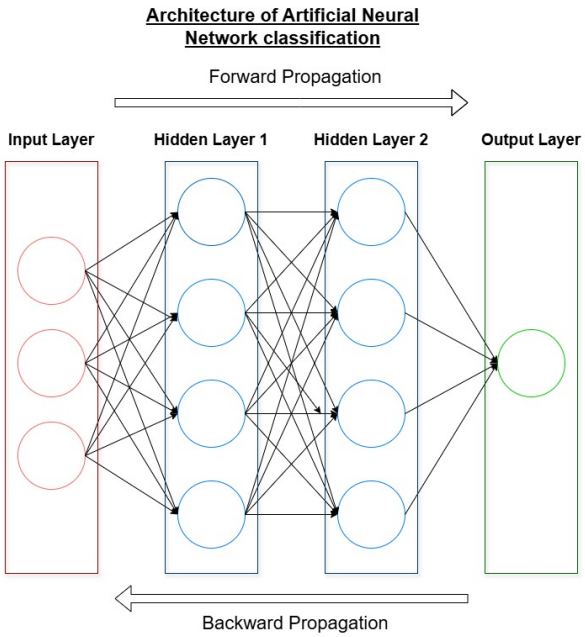


Figure 6.19: Architecture of Artificial Neural Network classification

In a neuron, the first step is calculating the weighted sum of its inputs, represented as:

$$z = \sum (w_i \times x_i) + b$$

[13]

Where $w_i$ are the weights, $x_i$ are the inputs, and $b$ is the bias. This value $z$ is then passed through an activation function to introduce non-linearity and produce the neuron's output. In our project, we used two key activation functions: ReLU and sigmoid, depending on the layer's purpose.

- **ReLU (Rectified Linear Unit):**

  $$f(z) = \max(0, z)$$

  [26]

  outputs the input $z$ directly if it's positive and zero otherwise. ReLU is computationally efficient and avoids the vanishing gradient problem, which often hampers deep networks. In our project, ReLU was applied in the hidden layers, allowing the network to learn complex, non-linear relationships between

features such as 'BMI', 'glucose levels', 'age', 'physical activity', and 'family history'. This function ensures that the network can model intricate patterns without saturating gradients, leading to effective learning.

- **Sigmoid Activation Function:**

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

[26]

maps the input $z$ to a range between 0 and 1, effectively representing probabilities. This was used in the output layer of our network because our task involved binary classification (e.g., predicting the presence or absence of a disease). The sigmoid function enables the output to be interpreted as a probability, making it suitable for decision-making by thresholding outputs at 0.5.

CHAPTER 7

# Discussion and Results

Performance evaluation of Random Forest, XGBoost, and Neural Networks for the prediction of asthma, diabetes, and brain stroke is done in this chapter. Precision-Recall, f1-score, and Accuracy are considered as metrics for strength-weakness analysis on each model across the datasets.

In every dataset,the target variable is binary, where 0 = Does not have Disease and 1 = Have Disease

## 7.1 Comparative Analysis of Machine Learning

a) **Results of Asthma Prediction:** Random Forest (RF) has achieves the best performance of 92.62% accuracy, showing outstanding results for both class 0 and class 1. For class 0, the precision was 94.99%, and the recall was 89.64%, which indicates that most patients with no asthma are correctly identified by it with a good level of low false positives. For class 1, RF achieves a precision of 90.59% and a recall of 95.47%. This means that the model effectively classifies patients with asthma with high accuracy.

XGBoost has achieves an accuracy of 87.89%, showing competitive performance for both classes. It provides a precision of 90.91% and recall of 85.59%, hence effectively identifying patients without asthma at class 0.For class 1, it showed a recall of 90.44% and a precision of 84.90, performing well in recognizing patients

with asthma.

The closest competing algorithm to RF is NN, which maintains an accuracy of 91.74%. Also, the performance of both classes is fairly good: for class 0, precision and recall are 92.08% and 92.28%, respectively. These ensure that it classifies correctly patients with no asthma, while in the case of class 1, with a precision of 91.36% and recall of 91.14%, it comes out as a reliable classifier for the class that has asthma.

| **Algorithms** | **Precision (0)** | **Precision (1)** | **Recall (0)** | **Recall (1)** | **F1-Score (0)** | **F1-Score (1)** | **Accuracy** |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.9499 | 0.9059 | 0.8964 | 0.9547 | 0.9224 | 0.9297 | 92.62% |
| XGBoost | 0.9091 | 0.8490 | 0.8559 | 0.9044 | 0.8817 | 0.8758 | 87.89% |
| Neural Network | 0.9208 | 0.9136 | 0.9228 | 0.9114 | 0.9218 | 0.9125 | 91.74% |

Table 7.1: Results of Asthma Predicitions

In summary, the Random Forest algorithm showed the best results for the classification task of predicting asthma and proved to be more effective in the determination process for both classes. The performance of neural networks was found to be strong and balanced.The results obtained in this work are more accurate, therefore with better F1-score values, than those seen in the previous studies. Though many previous studies are based on smaller datasets, the larger dataset used in the current study, along with using more advanced techniques such as SMOTE, has allowed for balanced predictions, especially in identifying key factors related to pollution exposure and family history.

b) **Results of Diabetes Prediction:** Random Forest (RF) follows with an accuracy of 73.82%, reflecting quite reliable performance in both classes. On the other hand, in class 0, RF performs with 75.77% precision and 69.77% recall, revealing that RF is capable of capturing the majority of the patients who are non-diabetic. For class 1, RF gives a precision of 72.16% and recall of 77.84%, showing very good results for patients with diabetes.

XGBoost performs well for both class 0 and class 1, with a highest accuracy of 76.31%. With an accuracy of 78.35% and a recall of 72.48% for class 0, it successfully detects people without diabetes. Its ability to accurately categorize individuals with diabetes has been shown by its 74.56% accuracy and 80.11% recall for class 1.

The accuracy for the neural networks is very strong at 75.60%, and this neural network performance was relatively consistent on both classes. NN gives the precision of class 0 as 78.03%, with a recall of 71.04%, making a good determination of patients with no diabetes. On class 1, it is with a precision of 73.59% and a recall of 80.14%, able to classify well patients that have diabetes.

| Algorithms | Precision (0) | Precision (1) | Recall (0) | Recall (1) | F1-Score (0) | F1-Score (1) | Accuracy |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.7577 | 0.7216 | 0.6977 | 0.7784 | 0.7264 | 0.7489 | 73.82% |
| XGBoost | 0.7835 | 0.7456 | 0.7248 | 0.8011 | 0.7530 | 0.7723 | 76.31% |
| Neural Network | 0.7803 | 0.7359 | 0.7104 | 0.8014 | 0.7437 | 0.7672 | 75.60% |

Table 7.2: Results of Diabetes Prediction

In short, XGBoost is the best performance in predicting diabetes, doing well in both classes.These results have shown better accuracy,precision and recall than previous studies, especially for XGBoost. Most previous studies, usually based on smaller datasets, suffered from the problem of data imbalance, while the larger dataset used and careful preprocessing in this work enabled the model to make more accurate predictions based on critical features such as BMI, physical activity, and blood pressure.

c) **Results of Brain Stroke Prediction:** The maximum accuracy by Random Forest is 95.09%, performing exceptionally well in both classes 0 and 1. It achieves a precision of 96.41% and recall of 93.63% in class 0, hence identifying the patients who have not had a stroke. RF yields an accuracy of 93.85% with a recall of 96.54% for class 1, thus turning out to be highly effective in identifying patients who have had a stroke.

XGBoost follows closely with an accuracy of 94.91%. Its performance is also very close to that of the winning model for both classes. This algorithm yields a precision of 96.75% and a recall of 92.91% for class 0, showing a strong identification of patients who have not had a stroke. For class 1, it achieves a precision of 93.23% and a recall of 96.90%, hence accurately classifying stroke cases.

Neural Networks is 87.19% and performs well, though a little lower than those from RF and XGBoost. Class 0: The precision of NN reaches 88.92%, and recall is

84.86%, which shows that it performs pretty well in identifying patients without stroke. Class 1: NN yields a precision of 85.62% and recall of 89.50%, which shows that NN performs well in finding stroke patients.

| Algorithms | Precision (0) | Precision (1) | Recall (0) | Recall (1) | F1-Score (0) | F1-Score (1) | Accuracy |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.9641 | 0.9385 | 0.9363 | 0.9654 | 0.9500 | 0.9518 | 95.09% |
| XGBoost | 0.9675 | 0.9323 | 0.9291 | 0.9690 | 0.9479 | 0.9503 | 94.91% |
| Neural Network | 0.8892 | 0.8562 | 0.8486 | 0.8950 | 0.8684 | 0.8751 | 87.19% |

Table 7.3: Results of Brain Stroke

Among all, Random Forest yields the highest performance for the stroke prediction problem, with consistently high precision and recall in both classes, while XGBoost is also very close to it.Higher accuracy and recall than in previous studies.It has correctly predicted the significance of certain variables that are to be predicted, such as age and glucose levels. While studies restricted by their dataset might not be accurately generalisable with respect to predicting stroke or not, applying a larger-scale dataset provided the study with appropriate insights regarding reliability into stroke prediction.

## 7.2 Performance Comparison Through ROC Curves

a) **ROC Curve Analysis for Asthma Prediction:**

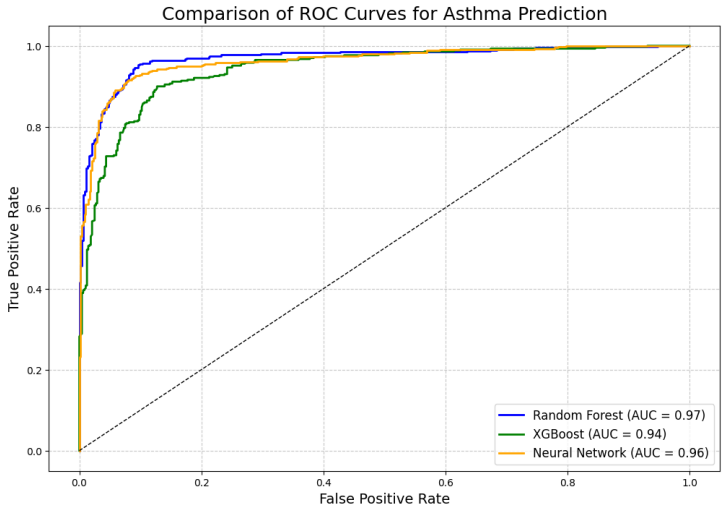


Figure 7.1: Comparison of ROC Curves for Asthma Prediction

The ROC curves below compares the performance of the three machine learning algorithms: Random Forest, XGBoost, and Neural Networks, on the classification of asthma, where 0 indicates that the patient does not have asthma and 1 represents a patient with asthma. The curve plots the balance between the true positive rate (sensitivity) and the false positive rate, thus offering a complete view of the classification capability of each model.

The Random Forest is very capable, with an AUC of 0.97, of differentiating between classes 0 and 1, and it is very steep to the top left; its curve means a high true positive rate while keeping the false positive rate under effective control and makes it the most efficient algorithm.

Neural Networks have an AUC of 0.96, showing high performance, closely after. It has a smooth curve throughout, indicating a great balance in the classification of classes 0 and 1. It stands a little behind Random Forest but still very effective. The XGBoost, having a pretty good performance, provides an AUC of about 0.94, which works fine to classify the classes well; still, in this one, the curve is pretty flat as compared to the rest, showing some increase in false positives. Overall, the Random Forest excels by AUC and reliable form of the ROC curve. Neural Networks and XGBoost also report quite good results, all of them making three models viable for asthma prediction.

b) **ROC Curve Analysis for Diabetes Prediction :**

The highest AUC value is achieved by XGBoost, which is 0.84, reflecting a very strong capability in distinguishing between 0 and 1. Its ROC curve increases consistently to the top-left corner, which means that there is a good balance between the true positive rate and the false positive rate. Thus, XGBoost can be trusted for the prediction of diabetes.

Neural Networks follow closely at 0.83, which demonstrates very robust results. The model's ROC is quite stable and therefore represents a high class balance between the two classes. Though not as strong as XGBoost, the Neural Networks results are very dependable over a wide range of thresholds.

Random Forest gives an AUC of 0.81, performing well but a bit lower than the other models. Its curve shows balanced performance but with a slightly higher

false positive rate compared to XGBoost and Neural Networks.

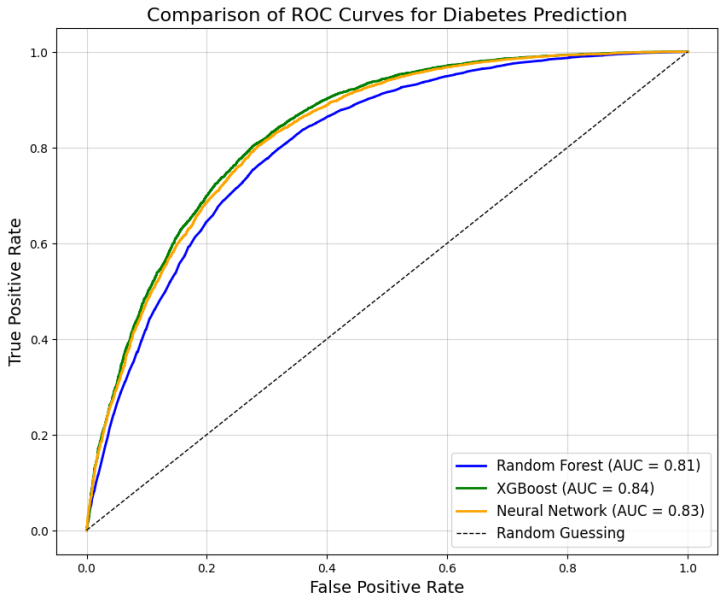


Figure 7.2: Comparison of ROC Curves for Diabetes Prediction

It can be concluded that XGBoost had the best performance in predicting diabetes, as reflected by its AUC value and the position of the curve. The Neural Networks also present quite consistent and reliable results, whereas the Random Forest still remains a good alternative, though a bit lower in its classification capability.

c) **ROC Curve Analysis for Brain Stroke Prediction :**

Random Forest and XGBoost have the same best AUC value, 0.99, which is quite excellent to differentiate between classes 0 and 1. The ROC curves of these models are steep and near the top-left corner of the plot, showing that these models have a very high true positive rate and a minimal false positive rate. This goes to prove the dependability of both algorithms in the prediction of brain stroke.

The Neural Networks have an AUC of 0.94, which is very strong in terms of classification performance. Although a little lower than that of Random Forest and XGBoost, the ROC curve is consistent and classifies 0 and 1 effectively. The slight deviation from the other two curves suggests a marginally higher false positive rate but still reflects robust overall performance.

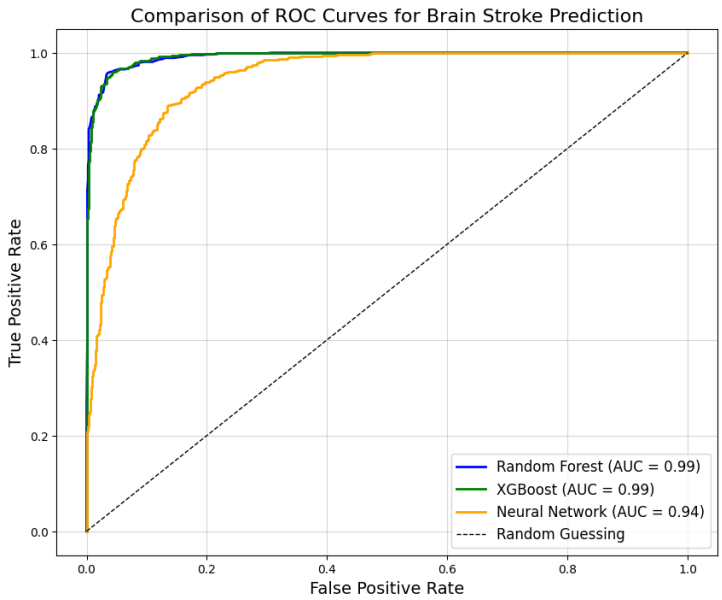


Figure 7.3: Comparison of ROC Curves for Brain Stroke Prediction

In summary, Random Forest and XGBoost turn out to be the best models for the prediction of brain stroke, considering their high values of AUC and regular ROC curves. Neural Networks are a little worse but also show quite reliable results, making all three models suitable for this task.

CHAPTER 8

# Conclusion and Future Scope

## 8.1 Conclusion

This study has presented a prediction of Asthma, Diabetes, and Brain Stroke chronic diseases using different machine learning algorithms.Each dataset was scaled, SMOTE for class balancing was performed, and feature selection was carried out to present each model with unbiased and consistent input for training.

The results describe very well that machine learning based on key risk factors proves to be a good forecast for chronic diseases. Based on the asthma disease result, the best performance went to Random Forest, focusing upon pollution exposure and family histories in view of crucial predictors. Diseases caused by diabetes are best under-forecasted by XGBoost, emphasizing BMI or physical activity and blood pressure too. In the case of brain stroke, the outcome value was highest for Random Forest at age and glucose levels most accordingly important. These findings agree with what is already found in existing literature, hence reinforcing machine learning in health while advancing a comparative advantage in aspects of accuracy and interpretability.

Overall, though, the models fared pretty well. Due to its much larger size, training and tuning parameters of models took a great deal of computational resources for the diabetes dataset, hence limiting the ability to explore even more complex architectures. In addition, SMOTE increased the class balance of all the datasets but at a slight introduction of noise into the data. Advanced techniques in feature sampling could be a

limiting factor, or developing an ensemble strategy to combine different algorithms for better results might also form a good future direction for these limitations.

This finally led to the study has achieved better performance compared to previous studies by using better ML algorithms that have improved in performance and show high accuracy and reliability in all datasets. It further enhances the performance using comprehensive approaches along with robust methodologies to present the most effective model compared to previous studies for the prediction of chronic diseases.

## 8.2 Future Scope

The outcomes of this study have proven that the machine learning models work effectively for the prediction of asthma, diabetes, and brain stroke. Yet, there is considerable scope for research and development in this domain to improve model performance and applicability in a real-world environment.

Future studies can also look at the integration of larger and more diverse data from multiple regions and demographics for broader generalizability. The integration of longitudinal data, in which patients are tracked over time, may provide deeper insights into the course of a disease and improve predictive accuracy. Furthermore, advanced feature engineering, including automated feature extraction using deep learning, may unravel hidden patterns and improve model interpretability.

These could be further enhanced by ensemble methods, such as stacking or hybrid models, that combine the strengths of various algorithms. Further, the use of explainable AI to understand the 'why' of a prediction can help in gaining better trust and, hence, adoption of these models in healthcare. Real-time data integration, including wearable device metrics, may also offer a promising avenue for personalized disease prediction and early intervention.

Deployment of such models into clinical settings finally needs to focus on aspects related to robustness, data privacy, and ethical concerns of fairness for various groups of patients. How to Address the Challenges: Let us see how we could overcome the above-mentioned challenges so that machine learning makes reasonable contributions to healthcare in the near future.

# Bibliography


[1] E. Dale Abel, Anna L. Gloyn, Carmella Evans-Molina, Joshua J. Joseph, Shivani Misra, Utpal B. Pajvani, Judith Simcox, Katalin Susztak, and Daniel J. Drucker. Diabetes mellitus—progress and opportunities in the evolving epidemic. *Cell*, 187(15):3789–3820, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0092867424007037, doi:10.1016/j.cell.2024.06.029.

[2] M Aiswarya Raj, Jan Bosch, Helena Holmström Olsson, and Anders Jansson. On the impact of ml use cases on industrial data pipelines. In *2021 28th Asia-Pacific Software Engineering Conference (APSEC)*, pages 463–472, 2021. doi:10.1109/APSEC53868.2021.00053.

[3] Bonna Akter, Aditya Rajbongshi, Sadia Sazzad, Rashiduzzaman Shakil, Jahanur Biswas, and Umme Sara. A machine learning approach to detect the brain stroke disease. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 897–901. IEEE, 2022.

[4] Aytan. Neural Network For Classification with Tensorflow, 10 2024. URL: https://www.analyticsvidhya.com/blog/2021/11/neural-network-for-classification-with-tensorflow/.

[5] Guest Blog. What is XGBoost Algorithm?, 11 2024. URL: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

[6] Erkut Bolat, Hasan Yildirim, Sedat Altin, and Eray Yurtseven. A comprehensive comparison of machine learning algorithms on diagnosing asthma disease and copd. *International Journal of Sciences and Research*, 76(3), 2020.

[7] CDC. Diabetes Health Indicators Dataset, 11 2021. URL: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv.

[8] CDC :CDC.gov. About stroke, 10 2024. URL: https://www.cdc.gov/stroke/about/index.html.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Jyotsna Choudhary. Mastering XGBOOST: A Technical guide for machine learning practitioners. *Medium*, 5 2024. URL: https://medium.com/@jyotsna.a.choudhary/mastering-xgboost-a-technical-guide-for-intermediate-machine-learning-practitioners-f7ad167c6865.

[11] William C. Cockerham, Bryant W. Hamby, and Gabriela R. Oates. The social determinants of chronic disease. *American Journal of Preventive Medicine*, 52(1):S5–S12, 12 2016. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5328595/, doi:10.1016/j.amepre.2016.09.010.

[12] Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti, Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, 12(2):295–302, 2018.

[13] Aman Dalal. Standard Neural Network - Analytics Vidhya - Medium. *Standard Neural Network*, 8 2020. URL: https://medium.com/analytics-vidhya/standard-neural-network-78174d7608f2.

[14] Niklas Donges. Random Forest: A complete guide for machine learning, 11 2024. URL: https://builtin.com/data-science/random-forest-algorithm.

[15] Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, Md Mahfujur Rahman, M Shamim Al Mamun, and M Shamim Kaiser. Performance analysis

of machine learning approaches in stroke prediction. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*, pages 1464–1469. IEEE, 2020.

[16] Joseph Finkelstein and In Cheol Jeong. Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences*, 1387(1):153–165, 2017.

[17] GeeksforGeeks. Random Forest Classifier using Scikitlearn, 1 2024. URL: https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/.

[18] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30, 2021.

[19] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.

[20] Ibm. Random Forest, 10 2024. URL: https://www.ibm.com/topics/random-forest.

[21] Jainvidip. Forward and backward propagation in multilayered neural networks: a deep dive. *Forward and Backward Propagation in Multilayered Neural Networks: A Deep Dive*, 7 2024. URL: https://medium.com/@jainvidip/forward-and-backward-propagation-in-multilayered-neural-networks-a-deep-dive-d596e875dedf.

[22] Zineb Jeddi, Ihsane Gryech, Mounir Ghogho, Maryame El Hammoumi, and Chafiq Mahraoui. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. *Healthcare*, 9(11):1464, 10 2021. doi:10.3390/healthcare9111464.

[23] Jobeda Jamal Khanam and Simon Y Foo. A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, 7(4):432–439, 2021.

[24] Rabie El Kharoua. Asthma disease dataset, 2024. URL: https://www.kaggle.com/dsv/8669080, doi:10.34740/KAGGLE/DSV/8669080.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[26] Johannes Lederer. Activation functions in artificial neural networks: A systematic overview, 2021. URL: https://arxiv.org/abs/2101.09957, arXiv:2101.09957.

[27] Dimitris Leventis. XGBOost Mathematics Explained - Dimitris Leventis - Medium. *XGBoost Mathematics Explained*, 11 2018. URL: https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a.

[28] Samaa A Mostafa, Doaa S Elzanfaly, and Ahmed E Yakoub. A machine learning ensemble classifier for prediction of brain strokes. *International Journal of Advanced Computer Science and Applications*, 13(12), 2022.

[29] Gede Angga Pradipta, Retantyo Wardoyo, Aina Musdholifah, I Nyoman Hariyasa Sanjaya, and Muhammad Ismail. Smote for handling imbalanced data problem: A review. In *2021 sixth international conference on informatics and computing (ICIC)*, pages 1–8. IEEE, 2021.

[30] MYLAPALLI KANTHI REKHA and I PHANI KUMAR. Brain stroke prediction using random forest and adaboost algorithm. *no*, 6:84–94, 2023.

[31] Shipra Saxena. Binary Cross Entropy/Log loss for binary classification, 11 2024. URL: https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/.

[32] Madison Schott. Random Forest Algorithm for Machine Learning - Capital One Tech - Medium. *Random Forest Algorithm for Machine Learning*, 12 2021. URL: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb.

[33] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. A review of deep learning with

special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, 194:105596, 2020.

[34] Deepti Sisodia and Dilip Singh Sisodia. Prediction of diabetes using classification algorithms. *Procedia computer science*, 132:1578–1585, 2018.

[35] Jillani Soft Tech. Brain Stroke Dataset, 8 2022. URL: https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset/data.

[36] Fatemeh Sogandi. Identifying diseases symptoms and general rules using supervised and unsupervised machine learning. *Scientific Reports*, 14(1):17956, 2024.

[37] Mitushi Soni and Sunita Varma. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, 9(09):2278–0181, 2020.

[38] Dimitris Spathis and Panayiotis Vlamos. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health informatics journal*, 25(3):811–827, 2019.

[39] Venkata Sravan Telu, Vinay Padimi, and Devarani Devi Ningombam. Optimizing predictions of brain stroke using machine learning. *Journal of Neutrosophic and Fuzzy Systems (JNFS)*, 2(2):31–43, 2022.

[40] Analytics Vidhya. Splitting Decision Trees with Gini Impurity - Analytics Vidhya, 11 2024. URL: https://www.analyticsvidhya.com/articles/gini-impurity/#:~:text=It%20ranges%20from%200%20to,an%20equal%20distribution%20of%20classes).

[41] World Health Organization: WHO. Noncommunicable diseases, 9 2023. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[42] World Health Organization: WHO and World Health Organization: WHO. Asthma, 5 2024. URL: https://www.who.int/news-room/fact-sheets/detail/asthma.