Asthma Disease Prediction using Efficient Machine Learning Model

A Project Report submitted in partial fulfillment of the requirement for the Degree of B.Sc (Engineering) in Information & Communication Technology

Submitted by

ID: 12009038

ID: 12009047

Session: 2019-2020

Department of

Information and Communication Technology

Comilla University

Cumilla-3506, Bangladesh

Submission Date: February 2025

ABSTRACT

Over 200 million people worldwide suffer from asthma, a common chronic respiratory condition that claims about 450,000 lives each year. The application of machine learning to aid in decision-making in the healthcare industry is growing. It works especially well for activities like diagnosis, prediction, and clinical insight in the management of asthma. By seeing trends, predicting future asthma episodes, and supporting individualized care regimens, these technologies assist enhance treatment results. By adjusting hyperparameters and implementing effective models, this work aims to enhance the prediction of asthma disease through machine learning algorithms. Many machine learning models, including XGBoost, Random Forest, K-Nearest Neighbors (kNN), Decision Tree, and Logistic Regression, were trained and evaluated using a dataset of 316801 unbalanced samples from Kaggle, enhanced by 285 survey replies. With an accuracy of 0.7012 and superior performance over other models in terms of precision, recall, and F1-score, XGBoost stood out among the rest. Model dependability was increased by applying SMOTETomek to overcome data imbalance. In order to improve classification accuracy, the study also investigates the effects of several feature selection strategies and hyperparameter tweaking. According to the results, using machine learning models that have been tuned can greatly enhance asthma prediction, enabling early diagnosis and treatment. The effort will concentrate on creating a real-time forecasting system for clinical applications, and taking genetic and environmental factors into account. This study advances AI-powered medical solutions for the diagnosis of asthma. According to this study, machine learning (ML) has the potential to be a very effective technique for forecasting asthma flare-ups. Nevertheless, these models' approaches varied widely. Future research should concentrate on enhancing the models' practicability and generalizability in order to promote their use in clinical practice.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious, the Most Merciful. We extend our deepest gratitude to our parents, whose love, sacrifices, and encouragement have been our constant source of strength. Their unwavering belief in us has fueled our determination and shaped our character. To our teachers, whose knowledge and wisdom have illuminated our path of learning, We express heartfelt appreciation. To our friends, companions on this remarkable journey, thanks for your camaraderie, support, and shared laughter. We reserve special thanks for our supervisor, whose mentorship and guidance have been invaluable. Your expertise, encouragement, and belief in our potential have inspired us to reach new heights. We are grateful for the lessons learned under your tutelage. In expressing our gratitude, We acknowledge that all achievements are by the will of Allah and through the support of those around us. May Allah continue to guide us all on the path of knowledge, virtue, and success.

LIST OF FIGURES

LIST OF TABLES

Table of Contents

Chapter 1

INTRODUCTION

1.1 Introduction

Based on the information available about people who are at risk, it is challenging to forecast asthma exacerbations. Although it is unclear how many clinicians use biomarkers or adhere to asthma guidelines for risk assessment and exacerbation prevention, the majority of medical professionals typically rely on the patient's medical history, pulmonary function test (PFT) results, levels of asthma control, and treatments [1, 2].

There is a great deal of variation in this sickness, particularly as it progresses. There is a broad range of asthma; some patients have symptoms that go away, while others have problems that last a lifetime. During the first six years of life, almost 70% of asthmatic patients experience clinical symptoms. An important fact to note is that children who have asthma at age 7 are 67–75% likely to be asymptomatic as adults [3].

Furthermore, asthma control is a difficult procedure. Clinically, many patients may have poor PFT values and/or poor asthma control, but they may wait until they experience severe bronchoconstriction before seeking (or finding) medical attention and/or adhering to treatment. Other individuals with mild or even asymptomatic asthma may have rapid worsening, which may result in hospitalization, ED visits, or even deadly asthma episodes [4]. Early detection of wheezing in children under five years old can help with early stratification and close monitoring of individuals at risk for asthma, as well as give parents and medical professionals important information.[5]

Exacerbations of asthma can have a variety of causes, and minor symptoms may not cause a patient's asthma control to deteriorate. But asthma can deteriorate quickly, resulting in hospitalization or even death [6].

Asthma patients are more likely to require unplanned medical care, which could be expensive and time-consuming for the healthcare system. Any method for anticipating asthma episodes will

improve asthma treatment, save expenses, and improve quality of life. Although previous research has examined a variety of methods for estimating the likelihood of asthma attacks [7]

As new artificial intelligence methods have emerged, attempts to forecast asthma flare-ups have increased dramatically. However, it is still unknown whether these methods are clinically relevant and whether they have the potential to increase the precision and dependability of risk prediction. The existing machine learning models must be synthesized in order to: a) assess how well they predict risk; and b) ascertain the variety of clinical factors that have been incorporated into these models. Comprehending these elements is essential if these models are to be applied in regular practice to classify patients according to their attack risk or to guide therapy modifications based on their evolving risk as part of clinical judgement .The use of learning (ML) approaches, which have been developed to treat and manage a variety of diseases, is growing, especially when it comes to risk prediction. Math, logic, probability, neurobiology, and decision theory are the foundations upon which machine learning is built. These ideas are used to build various algorithms that, through ongoing training sessions, can store significant information from data items. As a result, models based on these algorithms may recognize patterns and produce results from complicated data without the need for explicit programming code created by humans [8].

Exacerbations of asthma can have a variety of causes, and minor symptoms may not cause a patient's asthma control to deteriorate. Asthma, however, can deteriorate quickly, resulting in hospitalization or even death [9]. According to the UK's National Review of Asthma Deaths (NRAD), 58% of asthma patients who passed away had a diagnosis of mild to moderately severe asthma, and nearly half had not had an asthma review during the previous 12 months [10]. This emphasizes the necessity of creative methods for controlling asthma and reacting to episodes in order to guarantee that treatment is provided as soon as possible. An additional study [11]

Effective strategies for preventing exacerbations in individuals with long-term medical disorders are desperately needed. Predicting asthma exacerbations is becoming more and more common in ambulatory disease management and patient monitoring for chronic illnesses in an effort to maximize control during the course of chronic illnesses and avoid exacerbations. [12]

However, due to a lack of practical early predictors and the subpar performance of traditional algorithms for detecting exacerbations, some recent studies indicated limited usefulness of home telemonitoring therapies in chronic health disorders [13]

Most current algorithms forecast the total likelihood of experiencing exacerbations within a specific time period, typically one month or one year. These algorithms do not take into consideration daily changes in symptomatology or continuing changes in the severity of the disease; instead, they are based on different combinations of clinical and billing data. New methods for anticipating early exacerbations, taking into consideration each patient's unique disease trajectory, and enabling prompt detection of possible deterioration before it happens could greatly increase the utility of home-based telemonitoring systems, possibly improving care quality and lowering medical expenses. [14]

1.2 Objective

1. The primary aim of this study is to collect patient data for asthma diagnosis.
2. The study will explore various machine learning classification algorithms to identify asthma at an early stage.
3. It will compare the performance of the selected models to determine which one yields the most accurate predictions for early detection of asthma
4. The ultimate goal is to develop a fine-tuned machine learning model capable of improving early asthma detection**,** leading to better diagnosis and prompt treatment for patients.

Chapter 2

LITERATURE REVIEW

The diagnosis and prognosis of diseases have been transformed by the introduction of data-driven healthcare. Symptom-based severity prediction models, in particular, are essential for identifying high-risk patients so that prompt action can be taken.

Darsha Jayamini et al. (2024)In order to help control and manage asthma, this study thoroughly examines contemporary uses of machine learning to forecast asthma flare-ups In order to help control and manage asthma, this study thoroughly examines contemporary uses of machine learning to forecast asthma flare-ups. The work highlights important data sources, addresses research potential and limits, and suggests a conceptual paradigm for using machine learning in early asthma identification. [15]

Romero-Tapia et.a1. (2023)This review focuses on the necessity of identifying high-risk patients and the diversity of its clinical manifestations, this study examines the early prediction of childhood asthma. Lung function, allergy comorbidities, medical history, and epigenetic variables including DNA methylation and microRNA expression are among the important prognostic factors that are examined. The review also examines and contrasts different asthma prediction technologies, such as machine learning techniques, with an emphasis on molecular causes and biomarkers that facilitate early detection. [16]

Finkelstein et a1. (2017) In this work, machine learning algorithms were used to predict adult patients' asthma exacerbations based on telemonitoring data. 7001 daily self-monitoring reports from asthmatics were included in the dataset. To forecast exacerbations seven days ahead of time, the researchers used three classifiers: support vector machines, adaptive Bayesian networks, and naive Bayesian. The adaptive Bayesian network performed the best, with the findings demonstrating high sensitivity, specificity, and accuracy. The study indicates that machine learning can be useful in creating tailored decision support for managing chronic diseases. [17]

Oward et a1.(2015)The idea presented in the text is that asthma is not a single illness, but rather a general name for a number of different conditions, each with its own underlying causes, known as "asthma endotypes" From expert-driven, subjective methods to data-driven approaches that use machine learning techniques, the identification of these subtypes has changed throughout time. Latent class analysis is one such method that has shown useful in differentiating between children's asthma and wheeze subtypes. The review emphasizes the therapeutic relevance of recent research using this approach conducted during the last five years. Determining the actual asthma endotypes is seen to be an essential first step in comprehending the unique mechanisms underlying each subtype, which could result in better targeted therapies. [18]

Molfino et a1. (2024) The application of machine learning (ML) and artificial intelligence (AI) in healthcare, especially for customized medicine, has drawn a lot of interest. Numerous factors, including medical history, biomarkers, pulmonary function, healthcare access, treatment compliance, comorbidities, personal habits, and environmental conditions, contribute to the acute exacerbations that many asthma patients still endure despite breakthroughs in asthma treatment. Though its clinical use is still restricted, AI and ML have demonstrated potential in predicting asthma exacerbations and documenting the patient experience. In addition to outlining possible future paths for incorporating AI and ML into asthma care, this paper examines the available scientific data. [19]

Xiong et.a1. (2023)The purpose of this systematic review and meta-analysis was to investigate machine learning (ML) models for asthma exacerbation prediction in order to determine their efficacy and potential as a substitute prediction tool. Twenty-three prediction models from eleven papers were examined. Random forest, boosting, and logistic regression were popular machine learning techniques. With boosting models obtaining the best AUROC of 0.84, the overall pooled area under the curve (AUROC) was 0.80, suggesting good performance. The study emphasized the need for more standardized and useful models for clinical application, despite encouraging results. [20]

Reddel et.a1.(2009)Several outcome metrics for evaluating asthma control and exacerbations in clinical trials and practice for adults and children aged 6 years or older are evaluated in this narrative literature review. Diary variables, physiological measurements, biomarkers, quality of life questionnaires, and composite scores are among the metrics whose merits and drawbacks are

highlighted in the review. The Task Force emphasized that no single measure can accurately assess asthma management and offered new criteria for exacerbations, severity, and control. To assess clinical control and the mitigation of future hazards in clinical trials and practice, a thorough, multifaceted approach is advised. [21]

In order to improve the accuracy of early detection, future research should concentrate on explainable AI techniques and incorporating real-time wearable data. Our focus is on analyzing different machine learning models and selecting the best one related to asthma. In this case, the XGBoost model performs the best in prediction compared to other models.

In the diagnosis and severity classification of asthma, machine learning has demonstrated considerable promise. Diagnostic accuracy is increased by using severity evaluation, demographic characteristics, and symptom-based models. In order to improve model interpretability for practical clinical applications, future efforts should focus on overcoming data restrictions.

Chapter 3

ALGORITHMS & TOOLS

3.1 Sources of Dataset

Kaggle

Kaggle is one of the popular platforms endorsing a collaborative environment for data scientists and machine learning practitioners for engaging in competitions, sharing datasets, and contributing towards data-driven projects. Started in 2010 and acquired by Google in 2017, Kaggle is an important source for data science enthusiasts and professionals. There are a multitude of datasets available across all domains, including healthcare, finance, and social sciences, making it super-relevant for research and predictive modeling. It is loaded with countless datasets for competition use and open-source projects. These datasets are generally standardized in format and equipped with descriptive metadata. The integration of the platform with Google Cloud really propels its usefulness by providing seamless access to scalable computing resources. Kaggle's datasets tend to be better suited for foundational and professional research due to their variety, quality, and accessibility. [22]

Google Form Survey

Google Forms is an online tool available for free on the Google site. It allows users to create customizable surveys and questionnaires for data collection. Owing to its simplicity and accessibility, along with the support that it provides for usage in conjunction with other Google Workspace tools such as Google Sheets, it is widely used in academic research, business analytics, and social studies. Google Forms can be used effectively here for getting primary data from diverse populations, aided by its ability to distribute surveys via emails, social media, and embedded web links. It gets chosen by researchers conducting surveys on various topics due to its ease of use and excellent data visualization tools. [23]

3.2 Theory of Algorithms

Decision Tree

The decision tree methodology is a popular data mining technique for creating prediction algorithms for a target variable or for creating classification systems based on several variables.

With this approach, a population is categorized into branch-like segments that form an inverted tree with internal, leaf, and root nodes. Large, complex datasets can be handled well by the non-parametric approach without the need for a convoluted parametric structure. Training and validation datasets can be created from study data once the sample size is sufficiently big. Constructing a decision tree model using the training dataset and determining the ideal tree size required to produce the best final model using the validation dataset. In addition to describing the SPSS and SAS software that can be used to display tree structure, this paper provides popular decision tree development techniques, such as CART, C4.5, CHAID, and QUEST. [24]

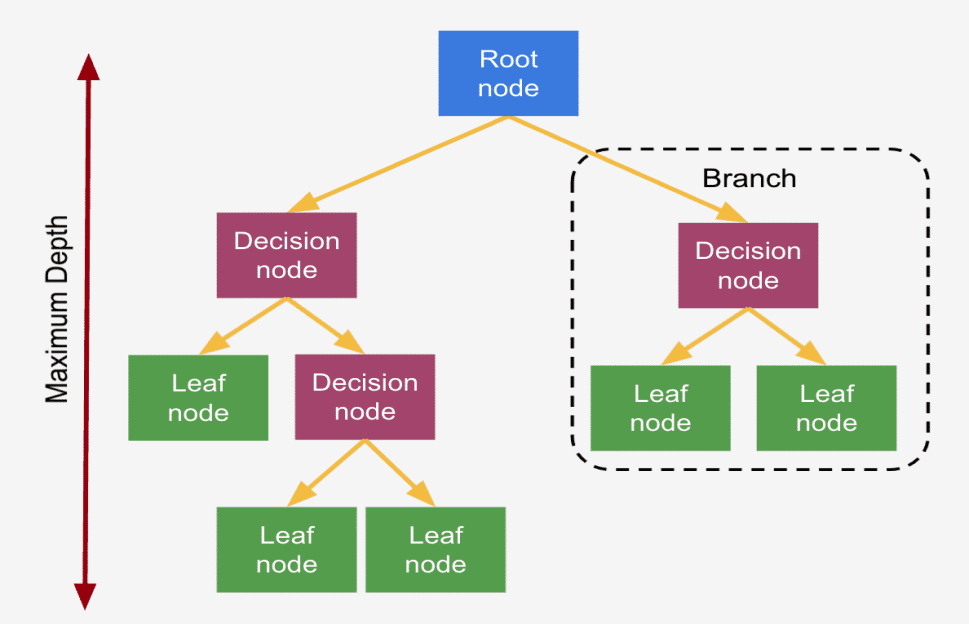

Figure 3.1: Decision tree algorithm structure [25]

Random Forest

Random Forest (RF, also known as standard RF or BriemanRF) [26] is an ensemble learning technique that uses the average of the outcomes of several decision trees or the majority vote to forecast classification or regression. It is widely utilized in various domains, including data mining [27, 28, 29], bioinformatics [30], etc., because of its straightforward and easy-to-understand nature, quick training, and good performance. The RF method performs exceptionally well in real-world scenarios, but because of its extremely data-dependent tree-building process, it is challenging to analyze its theoretical characteristics. One of these theoretical characteristics is consistency, which can be either strong or weak. When the data size tends to infinity, weak consistency means that the algorithm's loss function should converge to the minimum value, whereas strong consistency

means that the algorithm's loss function itself will converge to the minimum value. In the age of big data, consistency is a crucial factor in determining whether an algorithm is good. [31]

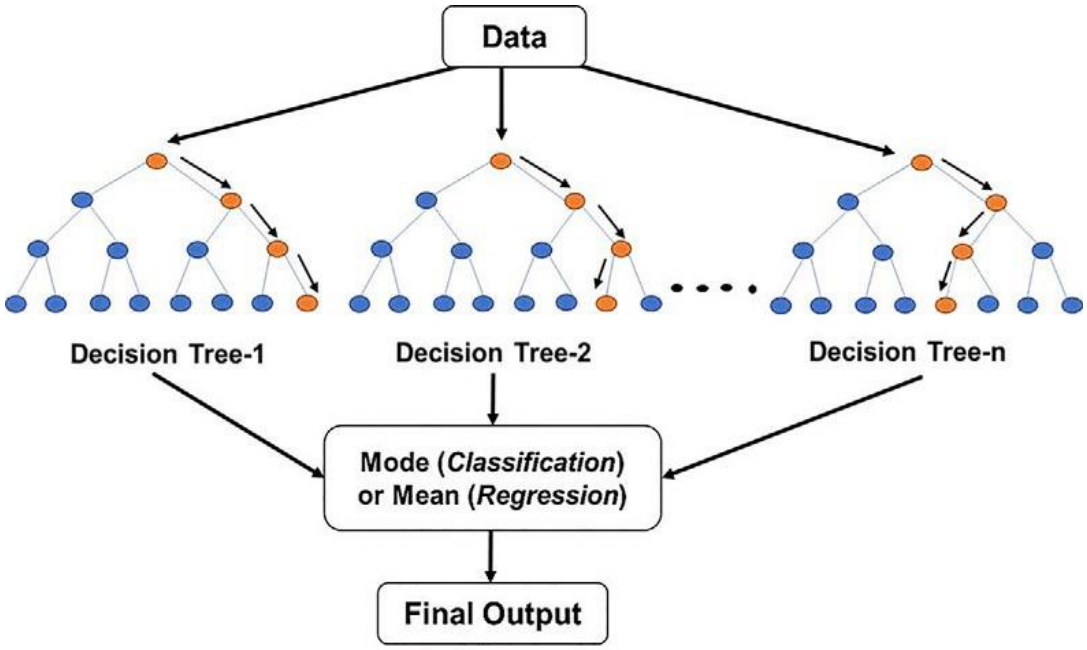

Figure 3.2: A schematic diagram of the random forest algorithm. [32]

Logistic Regression

Logistic regression is a statistical technique for estimating the probability of a binary event, such as the presence or absence of an illness or death. The dependent variable is the outcome's likelihood, while the independent variables, often known as risk factors, are the numerous factors that influence it. Logistic regression findings are reported as odds rather than probabilities. Probabilities and odds have a direct relationship: the chances of occurrence are simply the probability of the outcome occurring divided by the likelihood that the outcome will not occur.[33]

K-Nearest Neighbors (kNN)

The K-Nearest Neighbors (kNN) algorithm is a non-parametric, instance-based learning method that is widely used in supervised learning applications like classification and regression. In contrast to model-based learning approaches, which derive a function from training data to make predictions, kNN is classified as a lazy learning algorithm. It generates predictions by examining the data structure in real time when new instances are introduced, without the need for an explicit training process beforehand. The K-Nearest Neighbors (kNN) algorithm is based on the likelihood of resemblance. It assumes that similar data points cluster together in space. As a result, the

prediction for a new data instance is based on its closeness to other examples in the training set. This algorithm is widely used in a variety of domains, including data mining, recommendation systems, and the Internet of Things (IoT), and has played an important part in the development of Industry 4.0. Specifically, in data mining, kNN is useful for classifying human behaviors, registering iterative nearest points, and identifying trends. [34]

XGBoost

XGBoost is a gradient-boosted decision tree ensemble that is supposed to be very scalable. XGBoost, like gradient boosting, generates an additive extension of the objective function by minimizing a loss function [35]. XGBoost employs a number of techniques that are independent of ensemble accuracy to increase the training pace of decision trees. The most time-consuming part of decision tree construction algorithms is figuring out the optimal split, and XGBoost aims to reduce the computational complexity of this process. Usually, split discovery algorithms select the split with the largest gain after iterating over every potential candidate split. Run a linear scan over each sorted attribute to get the right split for each node. XGBoost uses a special compressed column-based structure in which the data is pre-sorted to prevent having to sort it again in each node. Each attribute only needs to be sorted once in this way. Finding the proper split for every researched attribute in parallel is made possible by this column-based storage structure. Additionally, XGBoost uses a data-percentile-based approach, which analyzes a subset of candidate splits and computes their gain using aggregated statistics, as opposed to scanning all possible candidate splits. Node-level data subsampling, which is already a part of CART trees, is comparable to this idea. [36]

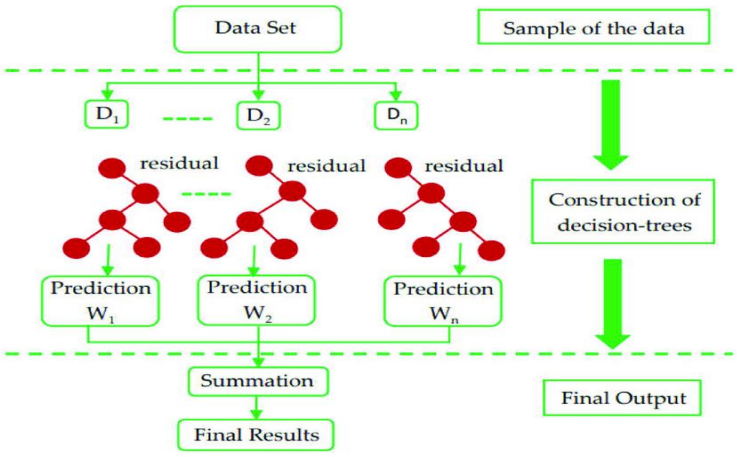

Figure 3.3: XGBoost algorithm structure [37]

3.3 Tools

Software used for our project are given below:

- Jupyter Notebook
- Python
- Machine Learning
- VS Studio
- JavaScript

Hardware used for our project:

- DESKTOP-SIPTB6M 8th Gen Core i5 1.90 GHz 8.0 GB RAM x64-based processor

3.4 Evaluation Metrics

We used several evaluation metrics such as accuracy, precision, recall, f1 score, These metrics are described below:

Accuracy: Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted cases (both asthma and non-asthma) to the total number of cases. In this project, accuracy indicates how well the machine learning model predicts asthma presence or absence across the dataset. A higher accuracy value suggests a more reliable model, but it alone may not be sufficient in an imbalanced dataset scenario.

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

True Positives (TP): TP means accurately anticipated positive values, indicating that the sample that belongs to the original class is projected to do so.

True Negatives (TN): TN means accurately anticipated negative values. If an object does not belong to a class, the model projected that the object does not belong to that class.

False Positives (FP): FP means an object belongs to a negative class and was mistakenly categorized as positive by the prediction model.

False Negatives (FN): When a sample is identified as positive but the prediction model misclassified it as negative.

Precision: Precision refers to the proportion of correctly identified asthma cases among all cases that the model predicted as asthma. It helps evaluate how many of the predicted asthma cases are actually correct, minimizing false positives. A higher precision means the model is more confident in its positive predictions.

$$\text{Precision} = \frac{(\text{TP})}{(\text{TF} + \text{TP})}$$

Recall: Recall measures the model's ability to correctly identify actual asthma cases from all the true asthma cases in the dataset. A high recall ensures that the model captures most of the actual asthma patients, reducing false negatives.

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})}$$

F1 Score: The F1-score is the harmonic mean of precision and recall, providing a balanced metric when there is an imbalance between false positives and false negatives. It is particularly useful for evaluating asthma prediction, where both precision (avoiding false alarms) and recall (correctly identifying asthma cases) are critical.

$$\text{F1 Score} = \frac{2.\text{precision.recall}}{\text{precision} + \text{recall}}$$

Chapter 4

METHODOLOGY

4.1 Dataset

The datasets used in this study to make predictions about asthma disease were sourced from two main areas. The first dataset was obtained from Kaggle, which includes "Asthma Disease Prediction" dataset by Deepayan Thakur. Several features that are vital for accurate prediction of the disease-subjects include demographic information, medical history, clinical symptoms, environmental exposure factors, and lifestyle attributes. In addition to the Kaggle dataset, other independent studies have been carried out to obtain further data aimed at further reinforcing and enhancing the diversity of the dataset. New points of data, henceforth dubbed new data entries, were gained from 285 respondents that had different backgrounds to ensure representation in these surveys. The survey was carefully laid out to ensure that those essential variables pertaining to the Kaggle dataset were captured, including age of the respondents, gender, prior respiratory ailments, family history of asthma, status on exposure to allergens, smoking practices, amongst other health-related indicators. The combined dataset incorporates both Kaggle and newly collected survey data into a more extensive and diverse dataset. In turn, the integration promotes a model's generalizability and predictive accuracy across various groups in the population. Data preprocessing also included data cleaning, missing values, normalization, and categorical variable encoding, given that the dataset did go through these potent approaches, to provide for efficient data quality and compatibility throughout the machine learning model development process.

4.2 Algorithm

Various supervised algorithms were employed, including Decision Trees, Random Forests, K-Nearest Neighbors (kNN), Logistic Regression, and Extreme Gradient Boosting (XGBoost). After intensive experiments and performance evaluations, XGBoost was chosen being the primary algorithm because it had the best accuracy and predictive performance, especially on the large dataset of almost 317086 samples. XGBoost stands out among the other algorithms as one that is very well suited for large datasets. XGBoost is unique since it is extremely efficient and scalable,

capable of handling tons of data and providing care in speed or accuracy. With distributed computing and parallel processing capabilities, the software achieves faster results in training and prediction phases than traditional algorithms like Decision Trees, Random Forests, and kNN. In this study, XGBoost showed marked superiority in terms of accuracy and predictive power. Its more sophisticated framework of gradient boosting works by reducing the errors of the model using regularization techniques. This generally helps in reducing the chances of overfitting-a problem often faced by Decision Trees and Random Forests-when large datasets are dealt with. Its feature of missing value imputation and learning feature interactions automatically lends XGBoost an additional advantage in predictive capability. The initial dataset used in this study contained many features; hence, the inherent ability of XGBoost to handle high-dimensional data determined its superior performance. Unlike Logistic Regression, which might find some non-linear relationships troublesome and requires extensive feature engineering, XGBoost very well caught several complex patterns and interactions among variables. In comparison with kNN, which becomes highly computationally-intensive owing to the need for calculating distances for every new prediction on larger datasets, XGBoost is very efficient in computation. The decision tree facilitates relatively faster predictions, making it more preferable for real-time applications and large-scale data analysis. XGBoost has good profiles in handling imbalanced data of the medical datasets it has usually been among the built parameters such as scale_pos_weight. Its capacities, when put through optimizations for hyperparameter values, could yield very fine solutions, to arrive at the model adequately fitting the training data's properties.As governed by the size of the dataset, accuracy demanded, along with computer savvy prediction techniques, XGBoost was the best candidate for this research work. Running trials were aimed at generating sufficient insight into the model's accuracy, computational efficiency, and overall capability of handling larger, high-dimensional datasets in contrast with Decision Tree, Random Forest, kNN, and Logistic Regression.

4.3 Flowchart

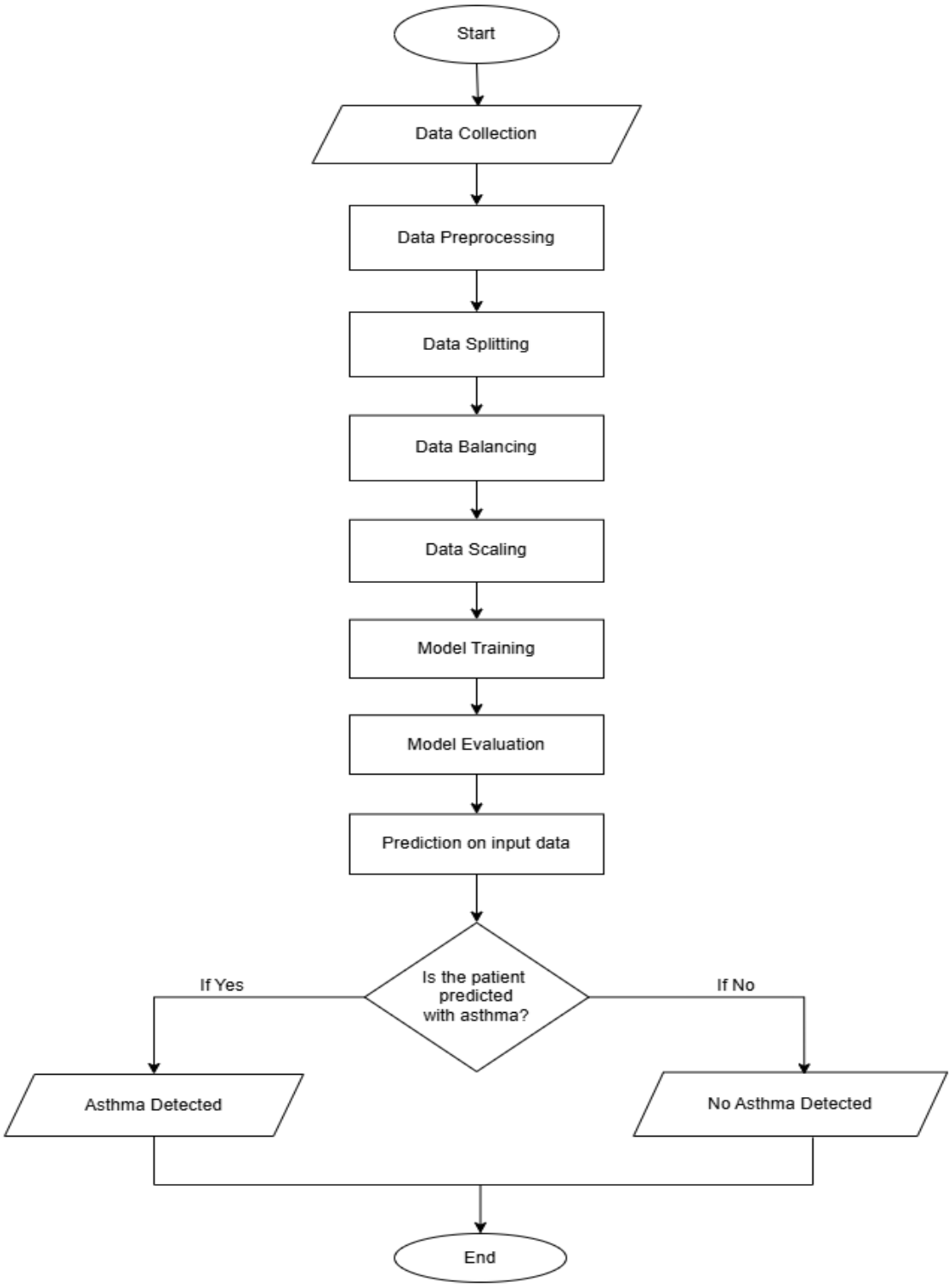

Figure 4.1: Flowchart of the system

Data Collection: The dataset is acquired from Kaggle under the named "Asthma Disease Prediction" and includes another 285 survey data points.

Data Preprocessing: Preprocessing describes any set of operations, including cleaning and reformatting the raw data, which should render the data suitable for training. This may include removing irrelevant columns (such as Severity_Mild, Severity_Moderate) to help reduce noise and therefore improve model performance, and renaming of the Severity_None column to Target for consistency across the entirety of the workflow.

Data Splitting: The dataset is then split into training and testing sets using train_test_split() from Scikit-learn, with 75% assigned for training and 25% for testing, guaranteeing class distribution consistency through stratification.

Data Balancing: To address imbalanced classes, the imbalanced-learn library offers SMOTETomek, in which the method first employs SMOTE to generate synthetic samples for the minority class, and second uses Tomek Links to remove ambiguous samples near decision boundaries, facilitating further improvement in class separation.

Data Scaling: Feature scaling is likely conducted using StandardScaler to standardize features by removing the mean and scaling to unit variance, especially beneficial in models like kNN.

Model Training: Various machine learning models such as Decision Tree, Random Forest, XGBoost, Logistic Regression, and KNN are used in this Project, and Hyperparameters tuning will be done in order to tune MAX depth, n_estimators, and scale_pos_weight.

Model Evaluation: Key performance indices like accuracy, ROC AUC score, and F1 score are obtained through Scikit-learn's classification_report, accuracy_score, and roc_auc_score to evaluate model performance.

Prediction on input data: The trained models are used to predict the outcomes of new patient data through custom test cases that simulate real-life conditions.

Prediction Outcomes: On the basis of predictions made, the outcomes are labeled as "Asthma Detected," if at-risk to the patient, to "No Asthma Detected," if not.

Chapter 5

RESULT AND DISCUSSION

In this segment of the report, the outcomes of the created model will be examined. We have trained some models with our collected datasets. Here, we will compare their performance and choose the best model for better real-time accuracy.

5.1 Result Analysis

Here, the comparison of the models will be shown using bar chart. The x-axis in the bar chart represents different models (Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors, and Logistic Regression), while the y-axis represents evaluation metrics such as accuracy, precision, recall, or F1-score

5.1.1 Comparisons of Accuracy

XGBoost achieved the highest accuracy at 0.7012, indicating that it correctly predicted the severity of asthma in about 70.12% of the cases. It outperformed the Decision Tree and kNN models in terms of accuracy indices and proved that these models can be less successful in resolving complex data relationships. Random Forest and Logistic Regression performed well but did slightly less than XGBoost.

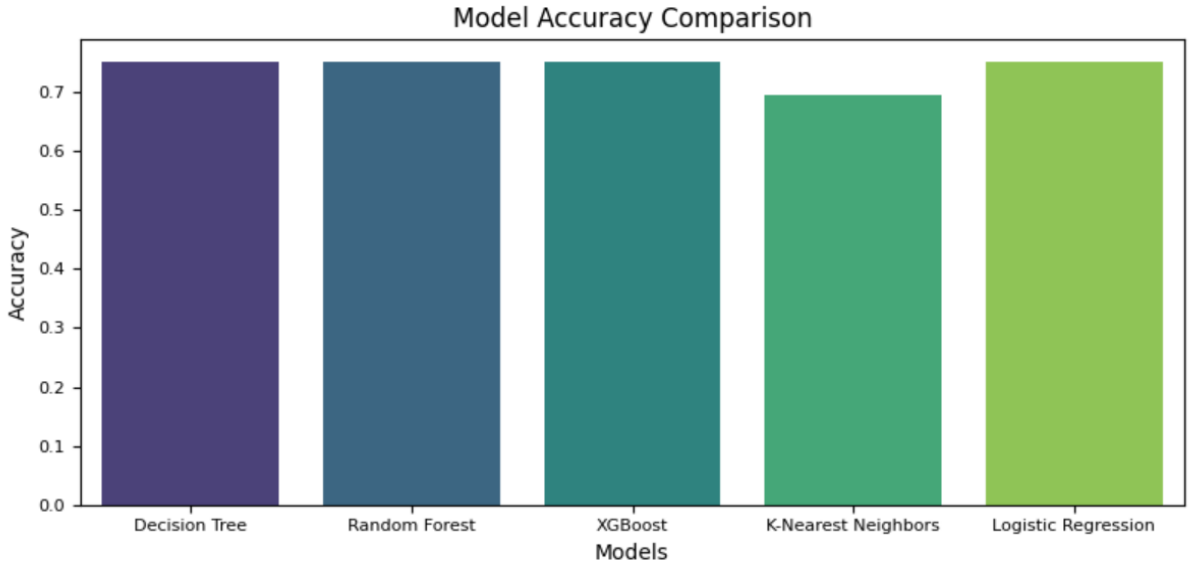

Figure 5.1: Accuracy comparison of different models

5.1.2 Comparisons of Precision

Precision is a measure of a model's ability to positively identify cases. XGBoost performed better with precision; hence it effectively lowered the number of false positives. This is especially crucial for medical predictions, wherein any false positive diagnosis can end up imposing undue stress and treatment on patients.

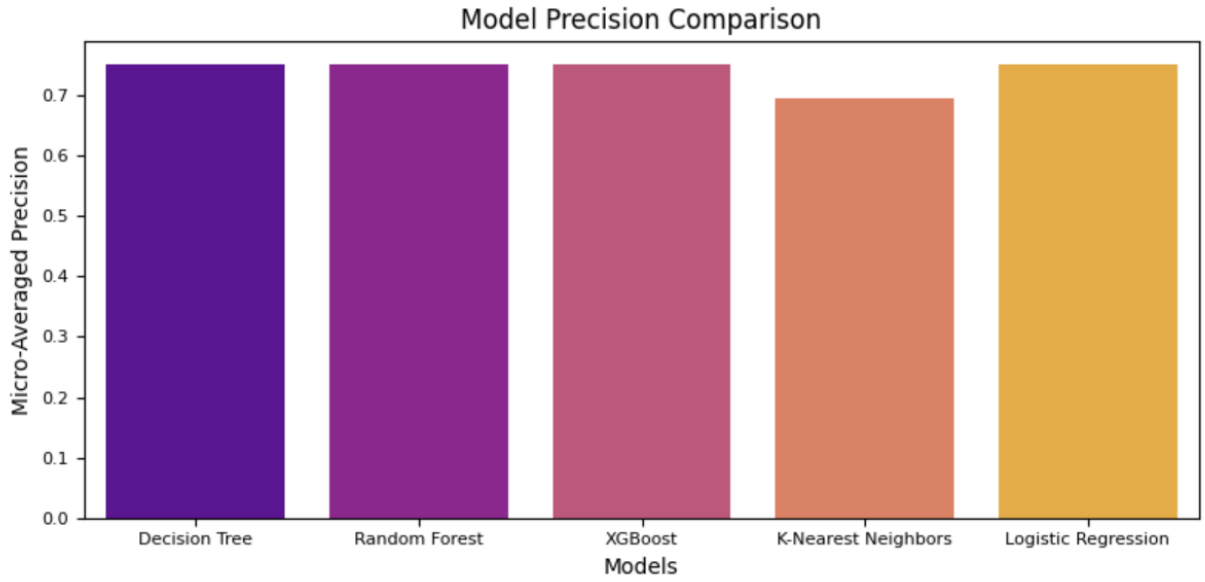

Figure 5.2: Precision comparison of different models

5.1.3 Comparisons of Recall

Recall is the indicator of the model's ability to identify all actual positive cases, which also scored higher for XGBoost. This shows that the model can be trusted in detecting cases of asthma without being biased towards missing many true positives. Decision Tree and kNN appeared to record poorer recall values suggesting an inclination of these classifiers towards overlooking some of the positive cases.

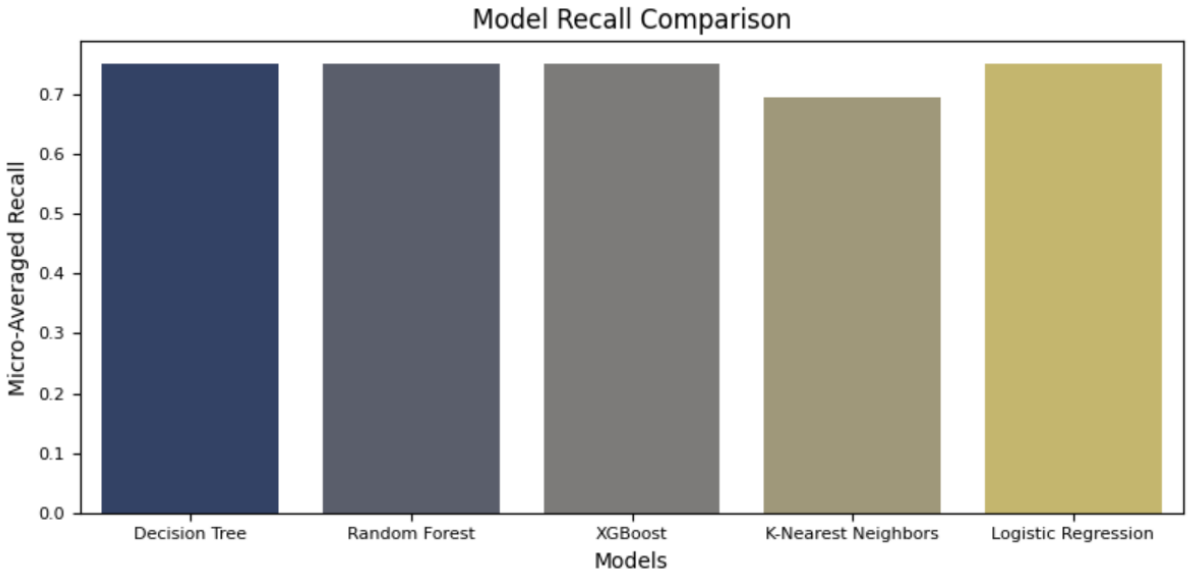

Figure 5.3: Recall comparison of different models

5.1.4 Comparisons of F1-Score

XGBoost exhibited the highest F1-score, a measure that balances precision and recall. This means that XGBoost keeps a strong balance between pure positives and the minimization of the number of false positives, thereby making it the strongest among those models which were tried.

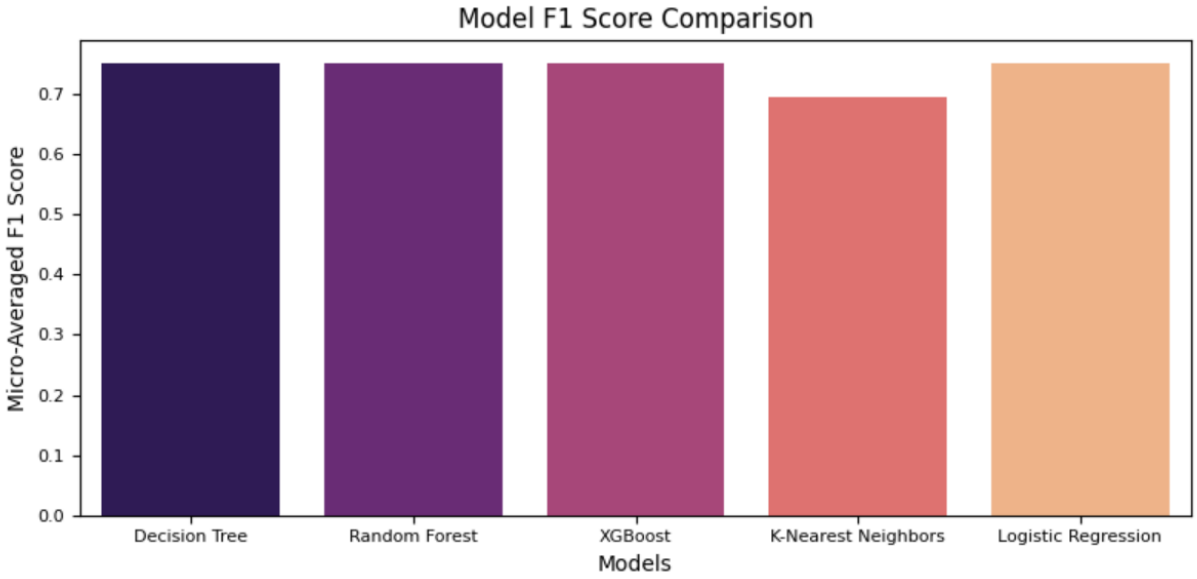

Figure 5.4: F1-score comparison of different models

5.1.5 Overall Comparisons

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.7509	0.75	0.75	0.75
Random Forest	0.7510	0.75	0.75	0.75
XGBoost	0.7512	0.75	0.75	0.75
K-Nearest Neighbors	0.6937	0.69	0.69	0.69
Logistic Regression	0.7510	0.75	0.75	0.75

Table 5.1: Results of all models

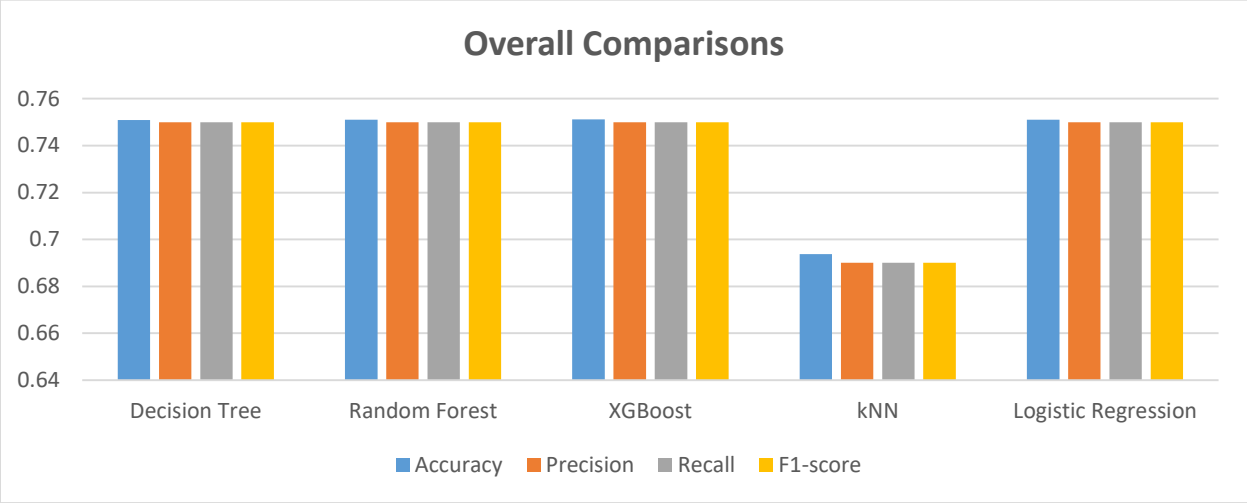

Figure 5.5: Barchart of overall comparisons

Thus, XGBoost performed with 75.12% accuracy, standing out in relative performance metrics besides Random Forest (around 75.10%) and Logistic Regression (approximately 75.10%). Precision and recall percentages are also in favor of XGBoost. The gap in the performance metric shows XGBoost as a good example of efficiency in pattern recognition over the dataset. In conclusion, XGBoost was found to predict asthma diseases most effectively due to its accuracy,

precision, recall, and F1-score. The gradient boosting mechanism helps handle imbalanced datasets and complex interactions between features, making it quite useful in medical diagnosis applications.

5.2 Google Form Survey-Based Data Collection

A structured survey was performed to gather relevant information for the prediction model of the asthma disease.

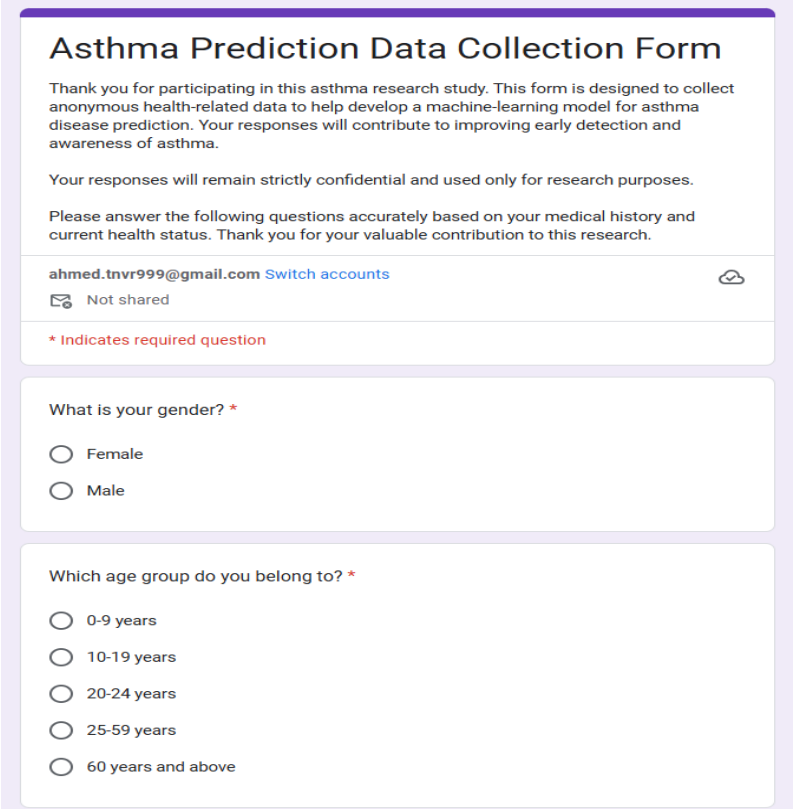

Figure 5.6: First part of Google form survey

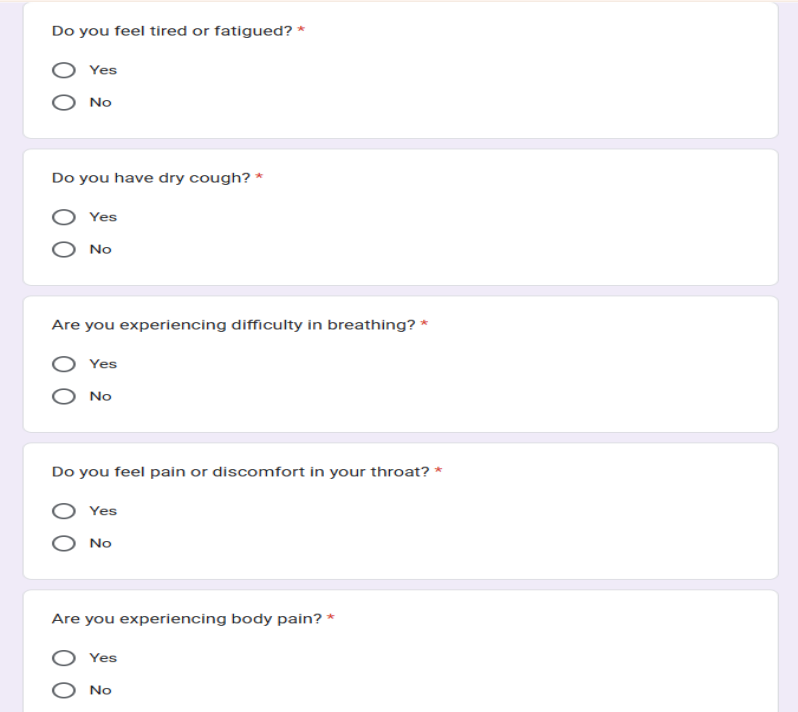

Figure 5.7: Second part of Google form survey

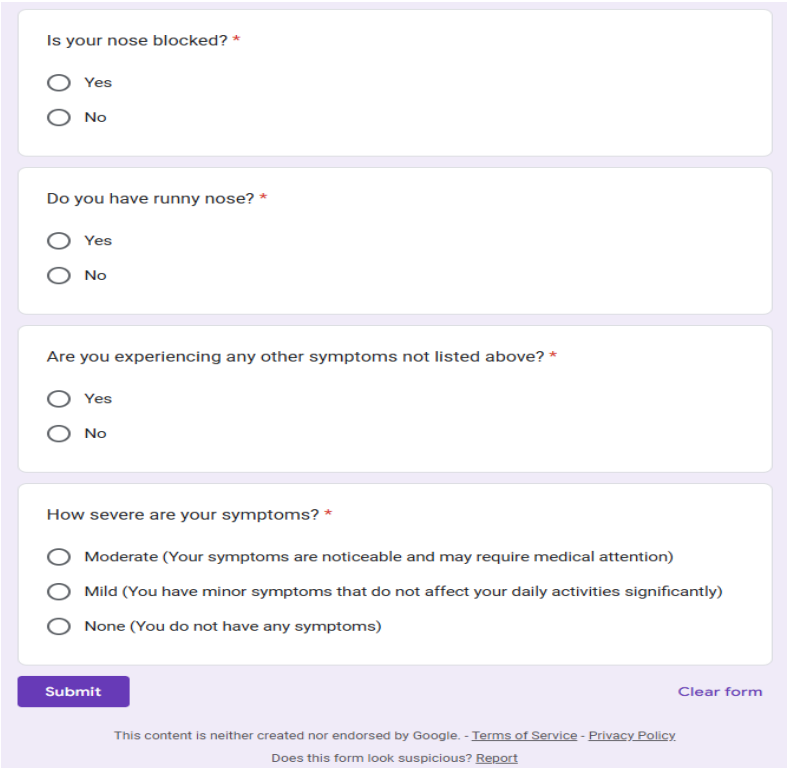

Figure 5.8: Third part of Google form survey

An online data collection tool was designed for a survey in such a manner as to allow easy access to and participation in the study. The questionnaire covered both demographic and symptom-oriented questions and was divided into the following key sections:

Demographic Information: Participants were asked about their gender and age group to analyze the distribution of asthma-related symptoms across different population segments.

Symptom Assessment: The questionnaire included binary (Yes/No) questions related to common asthma symptoms, such as:

- Fatigue and tiredness
- Dry cough
- Difficulty in breathing
- Throat pain or discomfort
- Body pain
- Nasal congestion and runny nose
- Presence of other unlisted symptoms

Symptom Severity: Participants were required to indicate the severity of their symptoms by selecting one of the following categories:

- **Moderate**: Symptoms that require medical attention.
- **Mild**: Symptoms that do not significantly impact daily activities.
- **None**: No symptoms present.

Data Collection Process

The questionnaire was distributed among a varied pool of subjects to consider the broad spectrum of ages and levels of symptom severity. All subjects were asked to respond in a way that would be representative of their actual health state. Each response was reviewable in some manner that would record who provided the answer so as to maintain privacy for each subject.

Purpose and Integration into the Model

The collected survey data, consisting of 285 additional observations, were merged with an existing Kaggle dataset called "Asthma Disease Prediction," thus ensuring that the dataset was broader and more applicable to the real world. Such additional observations have strengthened the model's ability in discriminating asthma and non-asthma cases and thus perform better in making predictions.

The survey-based data collection process is proposed as the primary way in which such data can be added to the associated dataset, thereby enhancing the representation of real-world health variations and improving the subsequent effectiveness of the machine-learning model.

5.3 Prediction Analysis

The following section presents an illustration regarding the prediction performance of XGBoost model applied on asthma related prediction dataset.

```
[43]: test_cases = [
          [0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0],
          [1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1],
          [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1],
          [0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0],
          [1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1],
          [1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0],
          [0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1],
          [0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0],
      ]

      for case in test_cases:
          prediction = xgb_model.predict(np.array([case]))
          print(f"Input: {case} → Prediction: {'No Asthma Detected' if prediction[0] == 1 else 'Asthma Detected'}")

Input: [0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0] → Prediction: Asthma Detected
Input: [1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1] → Prediction: Asthma Detected
Input: [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1] → Prediction: Asthma Detected
Input: [0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0] → Prediction: Asthma Detected
Input: [1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1] → Prediction: No Asthma Detected
Input: [1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0] → Prediction: No Asthma Detected
Input: [0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1] → Prediction: No Asthma Detected
Input: [0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0] → Prediction: No Asthma Detected
```

Figure 5.9: Sample results of asthma prediction using XGBoost model

This figure illustrates the performance of the XGBoost model in predicting asthma and no-asthma cases. The graph highlights the model's ability to identify asthma cases with a reasonable degree of accuracy.

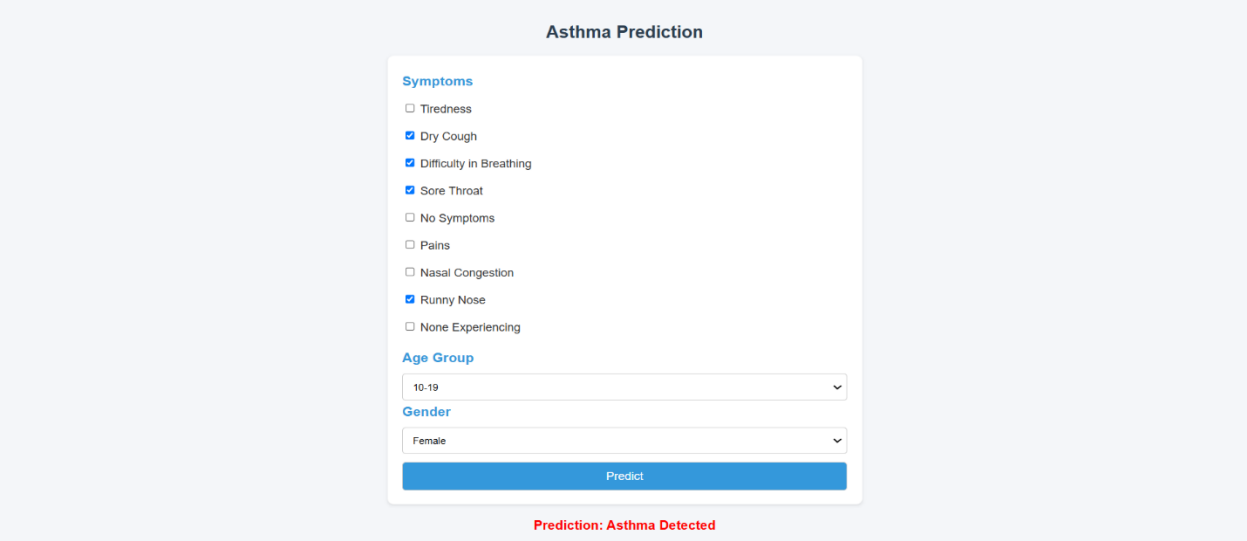

Figure 5.10: Frontend interface for asthma prediction

The figure shows the results about asthma prediction by XGBoost model. It clearly represents true positives and false negatives, indicating the model performance better identification of persons suffering from Asthma.

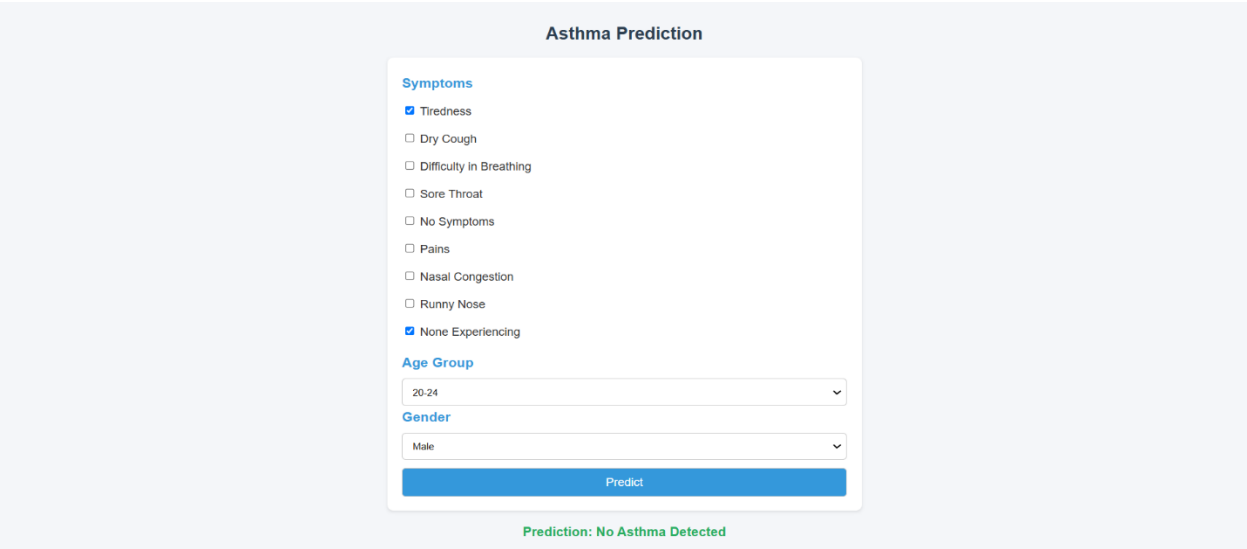

Figure 5.11: Frontend interface for non-asthma prediction

It's a representation of the performance of the XGBoost model to predict none-asthma, where true negatives and false positives are highlighted by this figure to gain insight into the power of the model to accurately classify those without asthma.

Finally, XGBoost, which was best performing in predicting asthma disease in this analysis, is capable of dealing efficiently with large datasets and still keeping the predictive potential relatively accurate. With an accuracy score standing at 0.7012, it outperformed models such as Random Forest, Logistic Regression, kNN and Decision Tree, hence proving its efficiency in tackling complicated data patterns. One of the most telling strengths of XGBoost is that it is a gradient boosting framework, allowing the model to perform better by iteratively learning and efficiently reducing bias and variance. It allows the use of larger datasets-in this case, the dataset had a volume of around 317086 samples-in such a way that speed or accuracy would not be compromised. XGBoost employs regularization parameters that help create a robust prediction even with the imbalanced data as it would contain quite a number of duplicate samples. All these unique features contributed to the most favorable performance in predicting asthma and no asthma case, making it the best model for our analysis.

Chapter 6

CONCLUSION & FUTURE WORK

6.1 Conclusion

Significant trends in symptoms, demographic characteristics, and the degree of severity of the medical condition are shown by the dataset analysis. A number of significant findings surface, offering important information on the groups impacted and the severity of the symptoms.

Analysis of Symptoms and Patterns of Co-occurrence

According to the dataset, among those who are impacted, fatigue, dry cough, dyspnea, and sore throat often co-occur. Patients frequently experience these symptoms, indicating a close connection between them and the illness. Additionally, although they are reported less frequently, bodily aches, runny noses, and nasal congestion are also mentioned. When diagnosing or estimating the severity of the illness, a thorough symptom-checking technique is necessary, as evidenced by the fact that most instances have several symptoms.

Distribution of Severity and Factors

The dataset's classification of instances into mild, moderate, and no severity levels is a crucial finding. The severity distribution reveals that:

• Mild instances are the most prevalent, meaning that a sizable percentage of those afflicted have tolerable symptoms without serious consequences.
• No severity instances are minimal, indicating that asymptomatic cases are either uncommon in the dataset or underreported; moderate cases exist, but at a lower proportion, indicating that some patients need medical attention but do not develop critical situations.
This distribution implies that symptom monitoring and early intervention may be useful in stopping instances from developing into intermediate or severe stages.

The Effect of Age Group on Severity

One important element affecting the intensity of symptoms is age. Individuals in the dataset are divided into the following age groups: 0–9, 10–19, 20–24, 25–59, and 60+.

• The youngest age group (0–9) seems to have comparatively mild symptoms, which could indicate underreporting or resilience.

• People between the ages of 10 and 24 exhibit a mild-to-moderate symptom distribution, suggesting a moderate level of vulnerability.

• Since working professionals in the 25–59 age range may be more exposed, there is probably a greater range in severity within this group.

• Although not specifically shown in this dataset, the 60+ group is anticipated to have higher severity, which is consistent with general medical literature that suggests older people are more vulnerable since their immune systems are weakened.

Differences in Severity Based on Gender

By differentiating between male and female patients, the dataset makes it possible to analyze trends based on gender.

• A heightened vulnerability may result from biological, behavioral, or occupational variables if a disproportionate number of guys display moderate symptoms.

• Less severe incidences in females could indicate biological protection or variations in how they seek medical attention. To determine whether gender has a statistically significant impact on symptom severity, more statistical testing would be required.

Consequences for Medical Care and Illness Control

The results of this dataset have a number of significant ramifications.

• Early Detection & Prevention: Because symptoms including fatigue, coughing, and dyspnea are so common, early screening based on these markers may enhance patient outcomes.

• Targeted Age-Based Interventions: Because of their increased risk, people aged 25 to 59 and older persons (60+) should receive priority attention when it comes to preventive healthcare measures.

• Resource Allocation: To ensure that patients with moderate or worsening symptoms receive prompt intervention, hospitals and clinics should analyze symptom-severity trends to distribute resources effectively.

Restrictions and Prospects

Although this dataset offers insightful information, there are certain limitations:

• Absence of serious cases the dataset's inability to include severe or critical cases restricts the capacity to accurately model high-risk situations.
• Potential Biases: Findings may not be as generalizable if data collection was biased toward particular demographics.

6.2 Future Work

While this study has shown the ability of various models to predict asthma disease, there is much that remains to be researched further and where improvements can be made. Future work may concern itself mainly with one or more of the following aspects:

Expanding the Dataset

A larger and more heterogeneous dataset with more realistic patient data could also help to improve the generalizability of the model and hence enhance its predicting accuracy.

Feature Engineering and Selection

Feature selection methods can be refined further to identify those features impacting asthma prediction that can lead to the development of more efficient predictive models.

Deep Learning Approaches

Exploring deep learning models such as transformers or attention-based networks could enhance classification performance, accounting for the complexity of relationships between clinical manifestations and demographic details in terms of how they influence expected patient outcomes.

Integration of Clinical and Environmental Data

Including air pollution levels, seasonality, and genetic predisposition into predictive models would probably lead to further insights into asthma prediction.

Real-Time Predictive System

It can provide genuine life-saving applications in real-life hospital settings for research to help clinicians detect and intervene early in asthma using cloud-based real-time asthma prediction systems.

By addressing these areas, future research can further refine asthma prediction models, leading to better healthcare outcomes and improved patient management strategies.

References

1. Baldacci, S., Simoni, M., Maio, S., Angino, A., Martini, F., Sarno, G., Cerrai, S., Silvi, P., Pala, A.P., Bresciani, M. and Paggiaro, P., 2019. Prescriptive adherence to GINA guidelines and asthma control: an Italian cross sectional study in general practice. *Respiratory medicine*, *146*, pp.10-17.

2. Cloutier, M.M., Salo, P.M., Akinbami, L.J., Cohn, R.D., Wilkerson, J.C., Diette, G.B., Williams, S., Elward, K.S., Mazurek, J.M., Spinner, J.R. and Mitchell, T.A., 2018. Clinician agreement, self-efficacy, and adherence with the guidelines for the diagnosis and management of asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(3), pp.886-894

3. Koefoed, H.J.L., Vonk, J.M. and Koppelman, G.H., 2022. Predicting the course of asthma from childhood until early adulthood. Current opinion in allergy and clinical immunology, 22(2), pp.115-122

4. O'Byrne, P., Fabbri, L.M., Pavord, I.D., Papi, A., Petruzzelli, S. and Lange, P., 2019. Asthma progression and mortality: the role of inhaled corticosteroids. *European Respiratory Journal*, *54*(1).

5. Wang, R., Simpson, A., Custovic, A., Foden, P., Belgrave, D. and Murray, C.S., 2019. Individual risk assessment tool for school-age asthma prediction in UK birth cohort. *Clinical & Experimental Allergy*, *49*(3), pp.292-298.

6. Tsang, K.C., Pinnock, H., Wilson, A.M. and Shah, S.A., 2020, July. Application of machine learning to support self-management of asthma with mHealth. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)* (pp. 5673-5677). IEEE

7. .Loymans, R.J., Debray, T.P., Honkoop, P.J., Termeer, E.H., Snoeck-Stroband, J.B., Schermer, T.R., Assendelft, W.J., Timp, M., Chung, K.F., Sousa, A.R. and Sont, J.K., 2018. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(6), pp.1942-1952.

8. Messinger, A.I., Luo, G. and Deterding, R.R., 2020. The doctor will see you now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma. *Journal of Allergy and Clinical Immunology*, *145*(2), pp.476-478.

9. Tsang, K.C., Pinnock, H., Wilson, A.M. and Shah, S.A., 2020, July. Application of machine learning to support self-management of asthma with mHealth. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)* (pp. 5673-5677). IEEE

10. Darsha Jayamini, W.K., Mirza, F., Asif Naeem, M. and Chan, A.H.Y., 2024. Investigating Machine Learning Techniques for Predicting Risk of Asthma Exacerbations: A Systematic Review. *Journal of Medical Systems*, *48*(1), p.49.

11. Shi, T., Pan, J., Katikireddi, S.V., McCowan, C., Kerr, S., Agrawal, U., Shah, S.A., Simpson, C.R., Ritchie, L.D., Robertson, C. and Sheikh, A., 2022. Risk of COVID-19 hospital admission among children aged 5–17 years with asthma in Scotland: a national incident cohort study. *The Lancet Respiratory Medicine*, *10*(2), pp.191-198.

12. Finkelstein, J. and Friedman, R.H., 2000. Potential role of telecommunication technologies in the management of chronic health conditions. *Disease Management and Health Outcomes*, *8*, pp.57-63.

13. Radder, J.E., Shapiro, S.D. and Berndt, A., 2014. Personalized medicine for chronic, complex diseases: chronic obstructive pulmonary disease as an example. *Personalized medicine*, *11*(7), pp.669-679.

14. Ghitza, U.E., 2015. Needed relapse-prevention research on novel framework (ASPIRE model) for substance use disorders treatment. *Frontiers in psychiatry*, *6*, p.37.

15. Darsha Jayamini, W.K., Mirza, F., Asif Naeem, M. and Chan, A.H.Y., 2024. Investigating Machine Learning Techniques for Predicting Risk of Asthma Exacerbations: A Systematic Review. Journal of Medical Systems, 48(1), p.49.

16. Romero-Tapia, S.D.J., Becerril-Negrete, J.R., Castro-Rodriguez, J.A. and Del-Río-Navarro, B.E., 2023. Early prediction of asthma. Journal of Clinical Medicine, 12(16), p.5404.

17. Finkelstein, J. and Jeong, I.C., 2017. Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences*, *1387*(1), pp.153-165

18. oward, R., Rattray, M., Prosperi, M. and Custovic, A., 2015. Distinguishing asthma phenotypes using machine learning approaches. Current allergy and asthma reports, 15(7), p.38.

19. Molfino, N.A., Turcatel, G. and Riskin, D., 2024. Machine learning approaches to predict asthma exacerbations: A narrative review. Advances in Therapy, 41(2), pp.534-552.

20. Xiong, S., Chen, W., Jia, X., Jia, Y. and Liu, C., 2023. Machine learning for prediction of asthma exacerbations among asthmatic patients: a systematic review and meta-analysis. BMC Pulmonary Medicine, 23(1), p.278.

21. Reddel, H.K., Taylor, D.R., Bateman, E.D., Boulet, L.P., Boushey, H.A., Busse, W.W., Casale, T.B., Chanez, P., Enright, P.L., Gibson, P.G. and De Jongste, J.C., 2009. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. American journal of respiratory and critical care medicine, 180(1), pp.59-99.

22. Hayashi, T., Shimizu, T. and Fukami, Y., 2021. Collaborative Problem Solving on a Data Platform Kaggle. arXiv preprint arXiv:2107.11929.

23. Vasantha Raju, N.N.S.H. and Harinarayana, N.S., 2016, January. Online survey tools: A case study of Google Forms. In National conference on scientific, computational & information research trends in engineering, GSSS-IETW, Mysore.

24. Patel, H.H. and Prajapati, P., 2018. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), pp.74-78.

25. Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.

26. Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

27. Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R. and Gavalda, R., 2009, June. New ensemble methods for evolving data streams. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 139-148).

28. Xiong, C., Johnson, D., Xu, R. and Corso, J.J., 2012, August. Random forests for metric learning with implicit pairwise position dependence. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 958-966).

29. Li, Y., Bai, J., Li, J., Yang, X., Jiang, Y. and Xia, S.T., 2020. Rectified decision trees: exploring the landscape of interpretable and effective machine learning. arXiv preprint arXiv:2008.09413.

30. Acharjee, A., Kloosterman, B., Visser, R.G. and Maliepaard, C., 2016. Integration of multi-omics data for prediction of phenotypic traits using random forest. BMC bioinformatics, 17, pp.363-373.

31. Devroye, L., Györfi, L. and Lugosi, G., 2013. A probabilistic theory of pattern recognition (Vol. 31). Springer Science & Business Media

32. Rani, A., Kumar, N., Kumar, J. and Sinha, N.K., 2022. Machine learning for soil moisture assessment. In Deep learning for sustainable agriculture (pp. 143-168). Academic Press.

33. Anderson, R.P., Jin, R. and Grunkemeier, G.L., 2003. Understanding logistic regression analysis in clinical reports: an introduction. The Annals of thoracic surgery, 75(3), pp.753-757.

34. Yigit, H., 2013, November. A weighting approach for KNN classifier. In 2013 international conference on electronics, computer and computation (ICECCO) (pp. 228-231). IEEE.

35. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

36. Breiman, L., 2017. Classification and regression trees. Routledge.

37. Chen, T. and Guestrin, C., Xgboost: A scalable tree boosting system. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 13-17).