ASTHMA DETECTION USING

MACHINE LEARNING

Done by

Santa Saji (MAC23MCA-2050)

Under the guidance of Prof. Nisha Markose

# ABSTRACT

Asthma is a prevalent chronic respiratory condition with significant implications for public health. Early and accurate detection of asthma is crucial for effective management and prevention of severe exacerbations. The “Asthma disease detection using Machine Learning” project aims to develop a machine learning (ML) model for predicting asthma using user reported symptoms, medical history, and environmental factors, without requiring clinical test results.

From the three papers, we get to know that different models were used for the detection of asthma disease. First paper focuses on developing a machine learning framework called BOMLA (Bayesian Optimization-based Machine Learning) for the early detection of asthma. Second paper is on predicting and preventing asthma attacks using a smart healthcare framework. Third paper aims at building machine learning models to predict the occurrence of childhood asthma.

The system is the comparative study of the algorithms such as Random Forest, Decision Tree and Support Vector Machine. The models will classify whether the patient is asthmatic or not. By comparing these algorithms, the project aims to determine which model offers the highest accuracy and reliability in diagnosing asthma conditions.

The dataset is taken from the Kaggle repository. The dataset contains 2393 sample observations and has 29 columns including 1 identifier, 1 class variable and 27 features. The dataset contains numeric values and Boolean values.

**Dataset:** <https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset>

## References

* Awal MA, Hossain MS, Debjit K, Ahmed N, Nath RD, Habib GM, Khan MS, Islam MA, Mahmud MP. An early detection of asthma using BOMLA detector. IEEE Access. 2021 Apr 13;9:58403-20.
* Alharbi E, Nadeem F, Cherif A. Smart healthcare framework for asthma attack prediction and prevention. In2021 National Computing Colleges Conference (NCCC) 2021 Mar 27 (pp. 1- 6).IEEE.
* Jeddi Z, Gryech I, Ghogho M, El Hammoumi M, Mahraoui C. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. In Healthcare 2021 Oct 29 (Vol. 9, No. 11, p. 1464).MDPI.

# INTRODUCTION

Asthma is a prevalent chronic respiratory condition marked by airway inflammation and bronchoconstriction, poses significant health risks and impacts millions globally. Early and accurate detection of asthma is crucial for effective management and prevention of severe exacerbations. The project "Asthma disease detection using Machine Learning" aims to develop a machine learning (ML) model for predicting asthma using user reported symptoms, medical history, and environmental factors, without requiring clinical test results. This lies at the intersection of healthcare and data science, specifically targeting asthma prediction and management. Machine learning techniques offer powerful tools for analyzing complex datasets to detect and predict asthma.ML models can identify patterns and risk factors associated with asthma exacerbations.

The methodology involves using supervised learning algorithms to analyze and predict the likelihood of asthma. The performance of the machine learning algorithms such as Random Forest, Support vector machine and Decision tree were compared for diagnosing asthmatic and non asthmatic patients. Thus, the project aims to determine which model offers the highest accuracy and reliability in diagnosing asthma conditions. The dataset is taken from the Kaggle repository. The dataset contains 2393 sample observations and has 29 columns including 1 identifier, 1 class variable and 27 features. The model utilizes a diverse dataset, including demographic data, medical history, symptom frequency, and environmental exposures. The most significant features were identified from the asthma dataset . The selected features are then used by the algorithms to build the models.

The project aims to develop a machine learning (ML) model for predicting asthma using user reported symptoms, medical history, and environmental factors, without requiring clinical test results. This has the potential to enhance early diagnosis and personalized management of asthma, ultimately reducing the burden of the disease on patients and healthcare systems.

# LITERATURE SUMMARY

## Paper 1: Awal MA, Hossain MS, Debjit K, Ahmed N, Nath RD, Habib GM, Khan MS, Islam MA, Mahmud MP. An early detection of asthma using BOMLA detector. IEEE Access. 2021 Apr 13;9:58403-20.

The primary focus of the paper is to develop a machine learning framework called BOMLA (Bayesian Optimization-based Machine Learning) for the early detection of asthma. It highlights the importance of accurate and early detection of asthma diseases to ensure timely and effective treatment.

Ten classifiers have been utilized in the BOMLA detector - SVC, RF, GBC, XGB, ANN, LDA, QLDA, NB, DT and KNN. ADASYN algorithm has also been employed in the BOMLA detector to eradicate the issues created due to the imbalanced dataset.

Support Vector Classifier yielded the highest accuracy(94.35%).

|  |  |
| --- | --- |
| **Title of the paper** | Awal MA, Hossain MS, Debjit K, Ahmed N, Nath RD, Habib GM, Khan MS, Islam MA, Mahmud MP. An early detection of asthma using BOMLA detector. IEEE Access. 2021 Apr 13;9:58403-20. |
| **Area of work** | Focuses on developing a machine learning framework called BOMLA (Bayesian Optimization-based Machine Learning) for the early detection of asthma. |
| **Dataset** | The dataset for this paper consists of clinical data from 389 individuals, including attributes such as AGE, SPO2, PULSE, FEV, FEV\_percent, FVC, FVC\_percent, FEV1\_by\_FVC, and MEF2575, collected in Khulna, Bangladesh, with 90.49% of the samples being asthma positive and 9.61% asthma negative. |
| **Methodology / Algorithm** | SVC, RF, GBC, XGB, ANN, LDA, QLDA, NB, DT and KNN.  ADASYN algorithm has also been employed in the BOMLA detector to eradicate the issues created due to the imbalanced dataset. |
| **Result / Accuracy** | LDA - 54.35 RF- 90.43  QLDA - 61.30 KNN - 74.35  NB - 57.39 DT - 76.96  XGB - 90.00 GBC - 92.61  ANN - 88.26 SVC - 94.35 |

## Paper 2: Alharbi E, Nadeem F, Cherif A. Smart healthcare framework for asthma attack prediction and prevention. In2021 National Computing Colleges Conference (NCCC) 2021 Mar 27 (pp. 1-6).IEEE.

This paper proposes a smart healthcare framework that provides patients with a sophisticated tool to visualize the asthma trigger and notify them about any predictable attack. The asthma attack prediction is based on utilizing a comprehensive set of patient data and different environmental triggers. The prevention model includes a dynamic visualization map of air pollution concentration, which alerts users with high-risk areas. Moreover, a context-aware safe- route recommendation system is proposed to keep users away from any asthma triggers. Thus, empowering patients and healthcare providers with timely information and actionable insights, ultimately enhancing the quality of asthma care and prevention strategies.

Neural network yielded the highest accuracy(94.35%).

|  |  |
| --- | --- |
| **Title of work** | Alharbi E, Nadeem F, Cherif A. Smart healthcare framework for asthma attack prediction and prevention. In2021 National Computing Colleges Conference (NCCC) 2021 Mar 27 (pp. 1-6). IEEE. |
| **Area of work** | The area of work is focused on predicting and preventing asthma attacks using a smart healthcare framework. |
| **Dataset** | The paper utilizes datasets that include environmental data (such as air quality and weather conditions), patient health records, and real-time monitoring data from wearable devices. |
| **Methodology / Algorithm** | Support vector Machine(SVM),Random Forest(RF),Neural network(NN) |
| **Result / Accuracy** | RF 92%  SVM 89%  NN 94% |

## Paper 3: Jeddi Z, Gryech I, Ghogho M, El Hammoumi M, Mahraoui C. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. In Healthcare 2021 Oct 29 (Vol. 9, No. 11, p. 1464). MDPI.

The paper focusses on building machine learning models to predict the occurrence of childhood asthma.

Algorithms used are Logistic regression ,decision trees, random forest and support vector machine.

Random Forest yielded the best predictive performance (accuracy= 87.8%).

|  |  |
| --- | --- |
| **Title of the paper** | Jeddi Z, Gryech I, Ghogho M, El Hammoumi M, Mahraoui C. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. In Healthcare 2021 Oct 29 (Vol. 9, No. 11, p. 1464). MDPI. |
| **Area of work** | Build machine learning models to predict the occurrence of childhood asthma |
| **Dataset** | A case-control study of 202 children was previously conducted in the Ibn Sina Hospital Center (CHUIS). A dataset resulted from this study and was made available to us for analysis. The study consists of children with (N = 101) and without (N = 101) asthma. The data collection was conducted over a period of 4 months, from May to September 2018. |
| **Methodology / Algorithm** | Random forest (RF) ,Logistic regression (LR),Support vector machine(SVM) and Decision trees (DT) |
| **Result / Accuracy** | Random forest yielded the best predictive performance (accuracy  = 87.8%), followed by logistic regression (accuracy = 83.56%), support vector machine (accuracy = 80%) and decision trees (accuracy = 85.3%). |

# SUMMARY

The first paper 'Awal MA, Hossain MS, Debjit K, Ahmed N, Nath RD, Habib GM, Khan MS, Islam MA, Mahmud MP. An early detection of asthma using BOMLA detector. IEEE Access. 2021 Apr 13;9:58403-20'focuses on developing a machine learning framework called BOMLA (Bayesian Optimization-based Machine Learning) for the early detection of asthma. Ten classifiers have been utilized in the BOMLA detector, where Support Vector Classifier

(SVC), Random Forest (RF), Gradient Boosting Classifier (GBC), eXtreme Gradient Boosting (XGB), and Artificial Neural Network (ANN) are state-of-the-art classifiers. In contrast, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QLDA), Naive Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) are conventional popular classifiers. ADASYN algorithm has also been employed in the BOMLA detector to eradicate the issues created due to the imbalanced dataset. SVC achieved highest accuracy of 94.35%.

The second paper 'Alharbi E, Nadeem F, Cherif A. Smart healthcare framework for asthma attack prediction and prevention. In2021 National Computing Colleges Conference (NCCC) 2021 Mar 27 (pp. 1-6). IEEE.' is focused on predicting and preventing asthma attacks using a smart healthcare framework. Support vector Machine (SVM),Random Forest(RF),Neural network(NN) were the algorithms used. NN achieved highest accuracy of 94%.

The third paper 'Jeddi Z, Gryech I, Ghogho M, El Hammoumi M, Mahraoui C. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. In Healthcare 2021 Oct 29 (Vol. 9, No. 11, p. 1464). MDPI' aims at building machine learning models to predict the occurrence of childhood asthma. Logistic regression analysis, decision trees, random forest and support vector machine are the algorithms used. Random Forest yielded the best predictive performance (accuracy = 87.8%).

# PROJECT PROPOSAL

From the above three papers, we get to know that different models were used for the prediction of asthma disease. Paper 1 focuses on developing a machine learning framework called BOMLA (Bayesian Optimization-based Machine Learning) for the early detection of asthma. Paper 2 is about predicting and preventing asthma attacks using a smart healthcare framework. Paper 3 is about building machine learning models to predict the occurrence of childhood asthma.

Accurate and timely detection of asthma diseases is crucial for effective treatment and management. The proposed system is the comparative study of algorithms such as Random Forest, Decision Tree and Support Vector Machine. The models will classify whether the patient is asthmatic or not. By comparing these algorithms, the project aims to determine which model offers the highest accuracy and reliability in diagnosing asthma conditions. Random Forest is known for its robustness and handling of complex datasets, Decision Tree for its simplicity and better visualization, SVM for its effectiveness in high-dimensional spaces and NN at capturing complex patterns and relationships in the data.

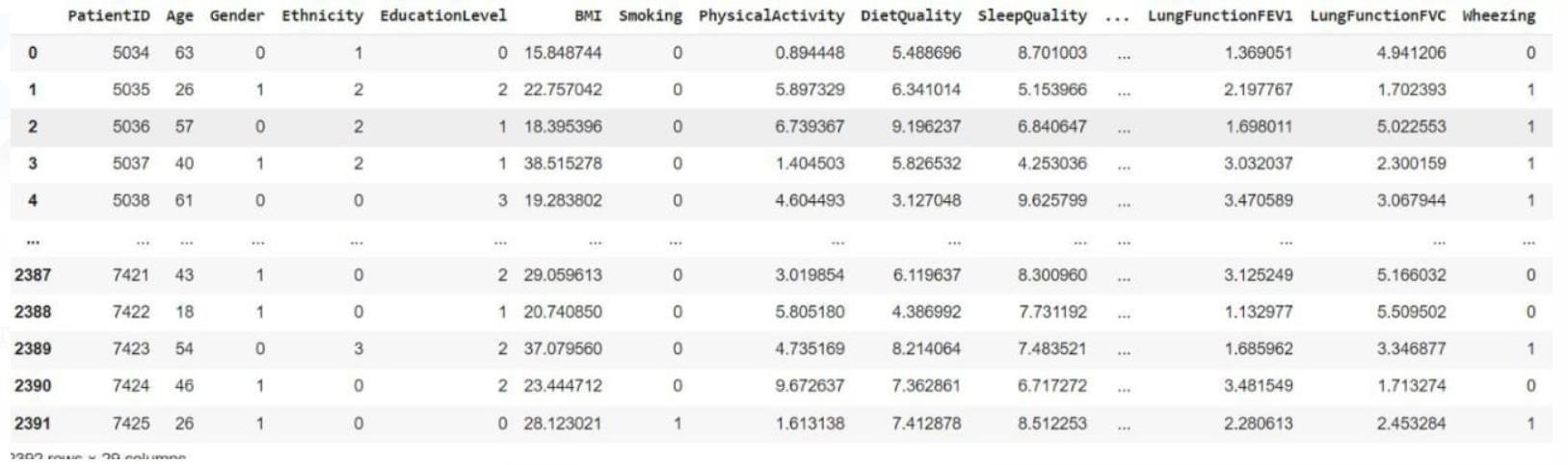
Dataset is collected from Kaggle. The dataset is then pre-processed to handle missing values, normalize data, and select relevant features. The dataset is split into training and testing dataset. Different machine learning models such as Random Forest (RF),Decision Tree (DT) and Support Vector Machine (SVM) are trained using the prepared training dataset. The models are evaluated on a separate test dataset using metrics such as accuracy, precision, recall, F1- score etc.

The project aims to develop a machine learning (ML) model for predicting asthma using user reported symptoms, medical history, and environmental factors, without requiring clinical test results. This has the potential to enhance early diagnosis and personalized management of asthma, ultimately reducing the burden of the disease on patients and healthcare systems.

**DATASET**

The dataset is taken from the Kaggle repository. The dataset contains 2393 sample observations and has 29 columns including 1 identifier, 1 class variable and 27 features. The dataset contains numeric values and Boolean values. There are missing values in the dataset.

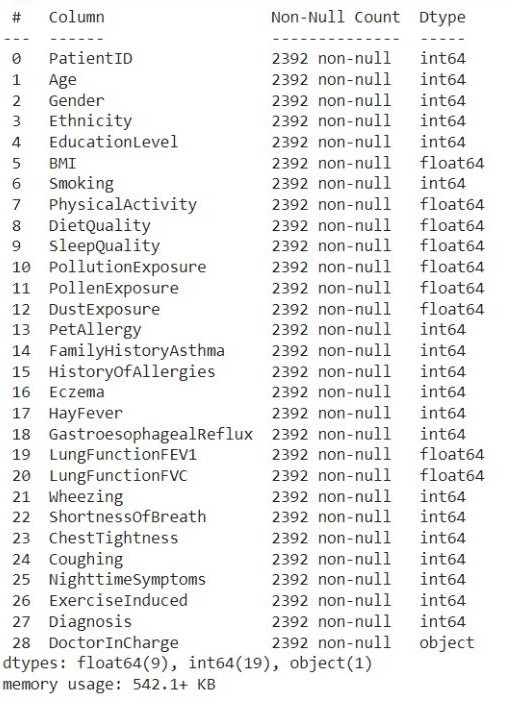
The identifier is the patient\_id. The class label is diagnosis. The features are age, gender, ethnicity, education\_level, bmi, smoking, physical\_activity, diet\_quality, sleep\_quality, poll ution\_exposure, pollen\_exposure**,** dust\_exposure, pet\_allergy, family\_history\_asthma,history\_of\_allergies,eczema,hay\_fever,gastroesophageal\_reflux,lung\_function\_fev1,lung\_function\_fvc,wheezing,shortness\_of\_breath,chest\_tightness,coughing,nighttime\_symptoms, exercise\_induced,doctor\_in\_charge.





# EXPLORATORY DATA ANALYSIS

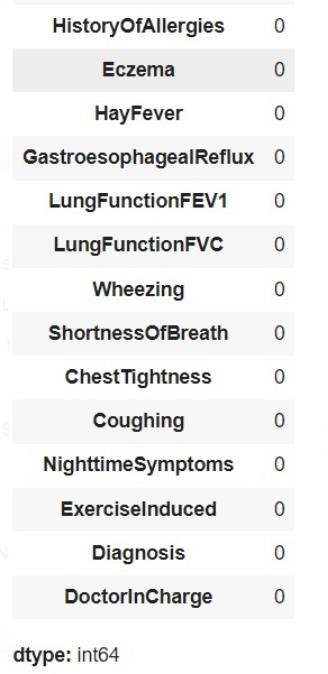
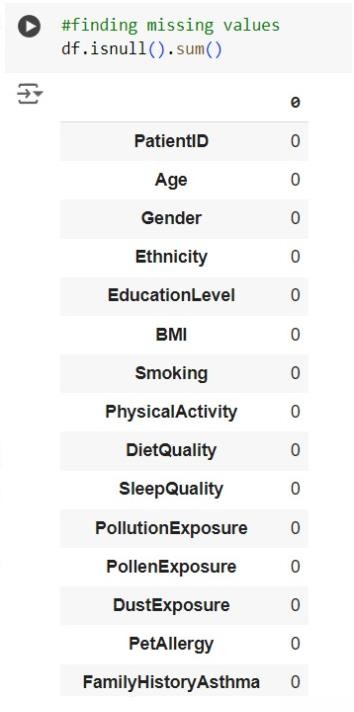
Dataset has 2392 rows and 29 columns



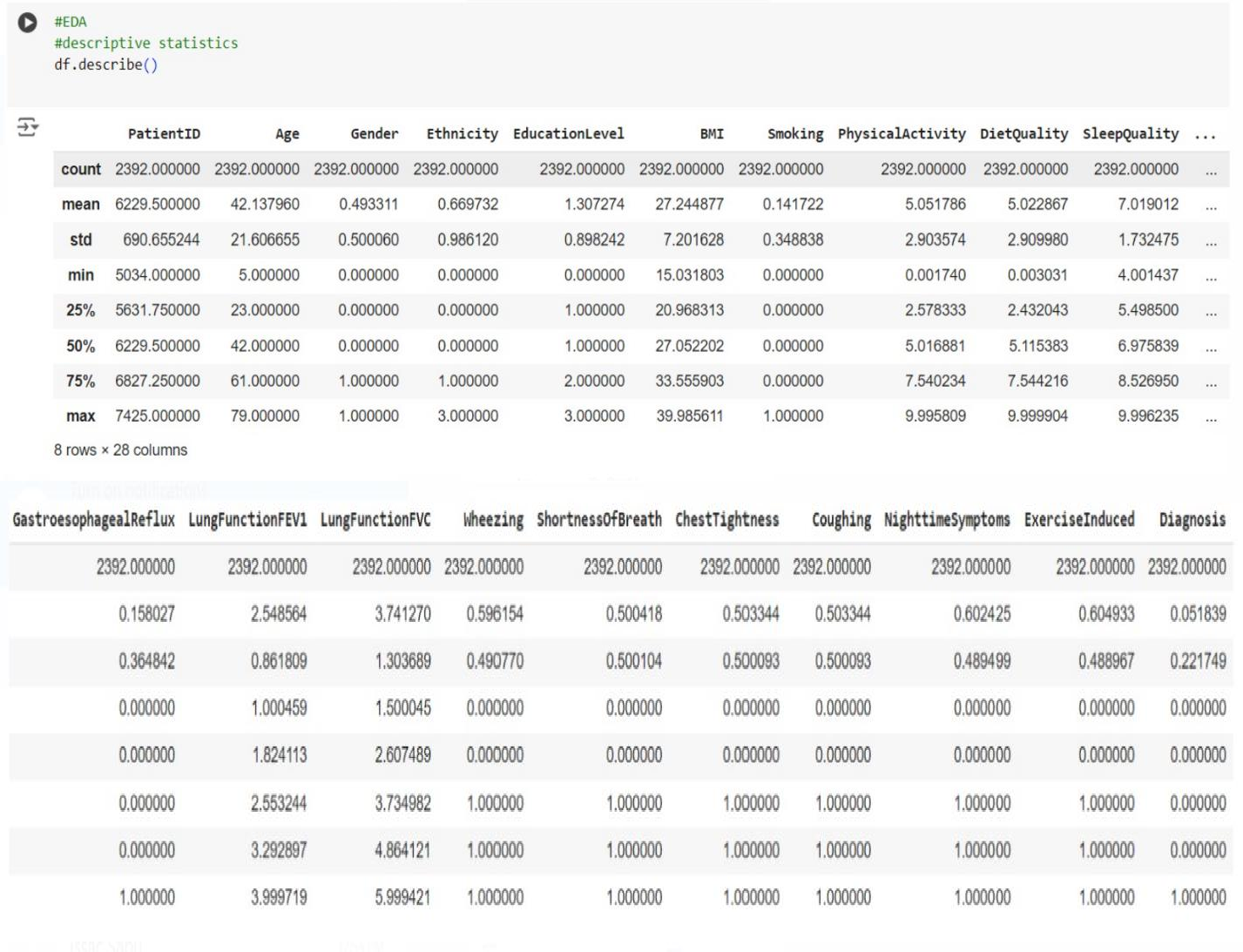
Dataset contains 28 numerical attributes and 1 non numerical attribute

## Missing values

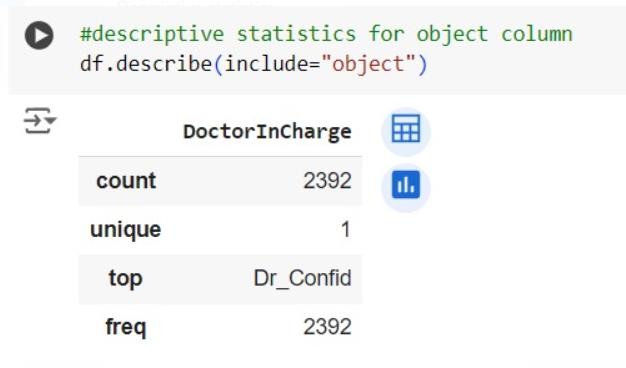
There are no missing values in the dataset



## Descriptive statistics



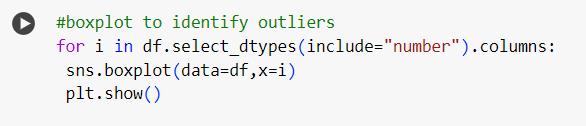
* + Age has a very high standard deviation and maximum value is 79



* **count: 2392** - This indicates that there are 2392 non-null values (entries) in the "DoctorInCharge" column.
* **unique:** 1 - This means that there is only one unique value (category) present in the entire column.
* **top:** Dr\_Confid - This is the most frequently occurring value (mode) in the column, which is "Dr\_Confid".
* **freq:** 2392 - This shows that the value "Dr\_Confid" appears 2392 times, implying that all entries in the column have the same value.

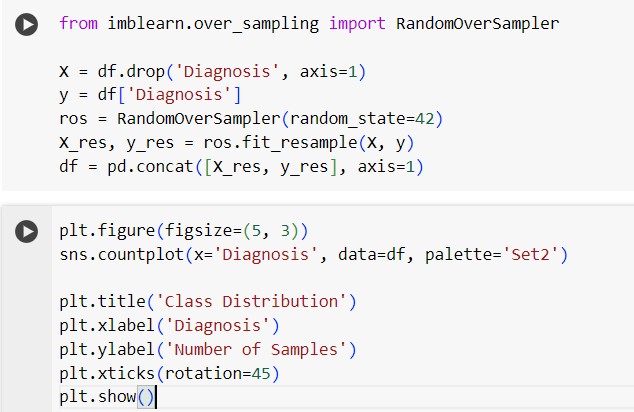
## Outliers:

Since all values in the "DoctorInCharge" column are identical ("Dr\_Confid"), there are no outliers present.

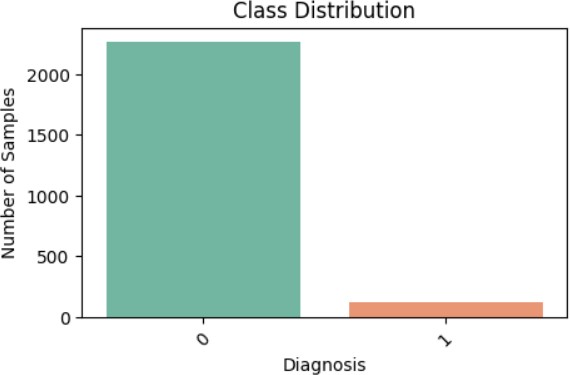


There are no outliers.

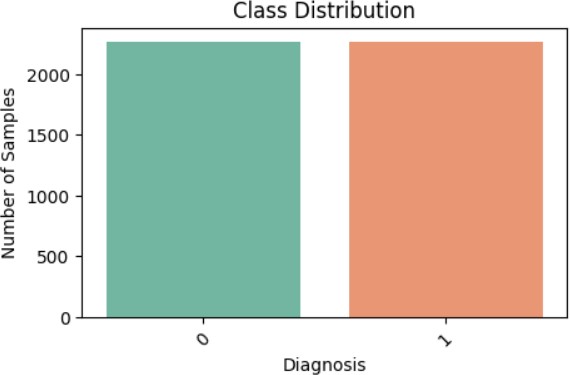
## Class Imbalance



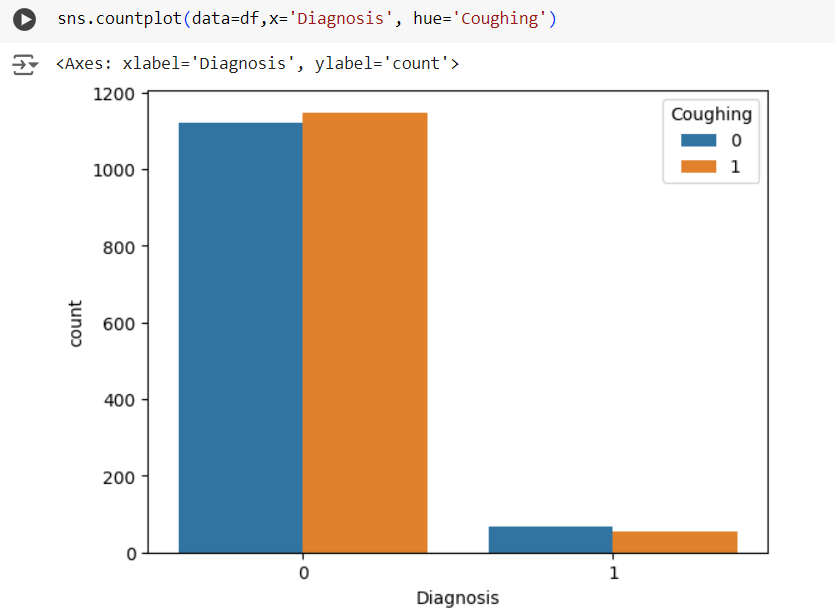
**Unbalanced**

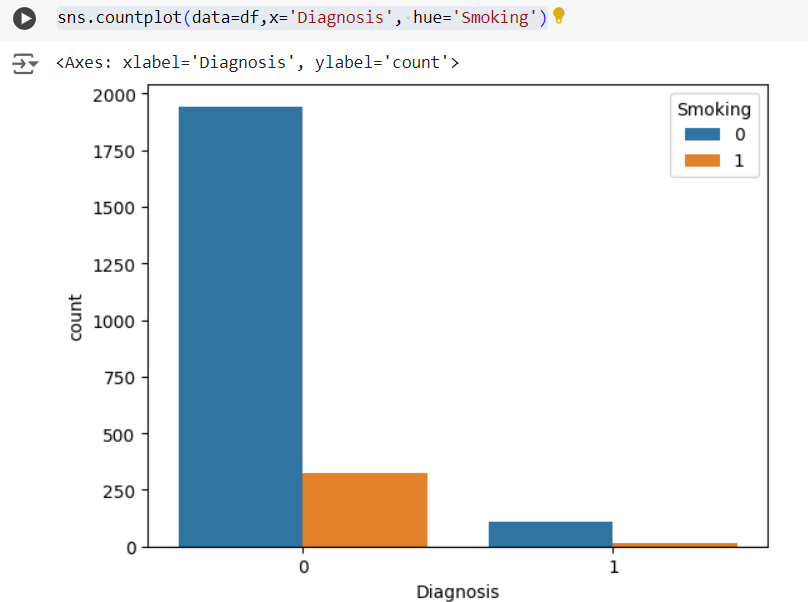


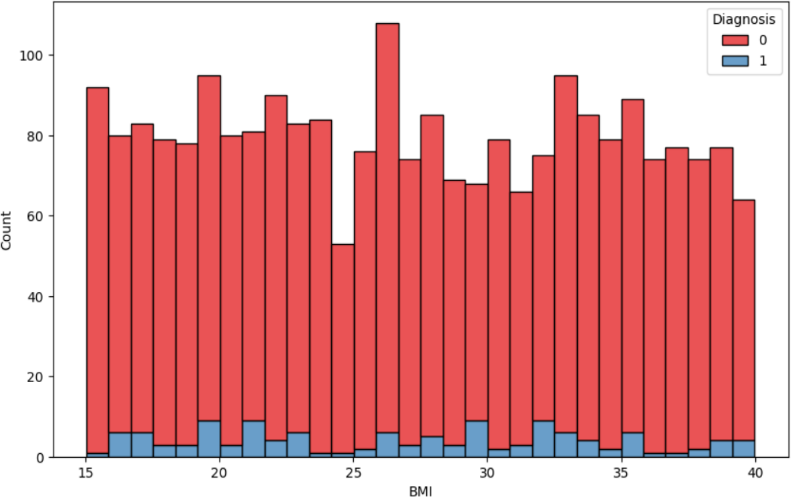
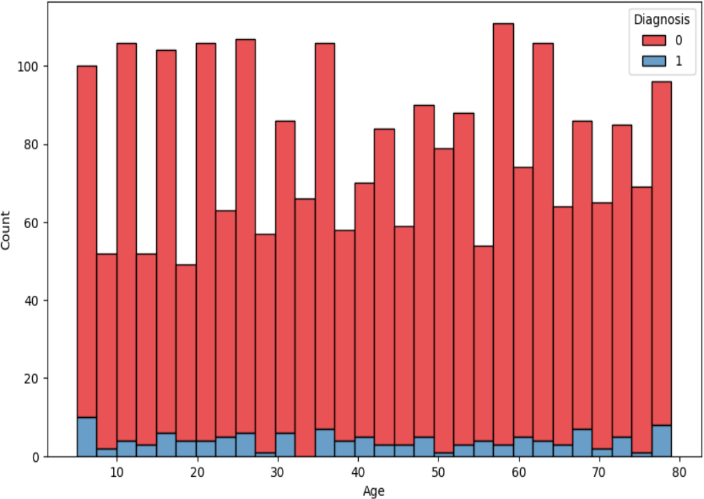
## Balanced



**Visualisation**







# WORKING OF ALGORITHMS

Algorithms used in 'Asthma Disease Detection using Machine Learning' are Random Forest, Support Vector Machine and Decision Tree.

## Random Forest

Random Forest is a supervised learning technique used for both Classification and Regression problems. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Instead of relying on one decision tree, the random forest takes the output from each tree and based on the majority votes of outputs, and it takes the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Working

**Step-1:** Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

## Pseudocode

1. Input:
   * Training dataset with N samples and M features.
   * Number of trees to create: T.
   * Number of features to randomly select at each split: F.
2. Initialize an empty list called 'forest' to store the trees.
3. For each tree (from 1 to T):
4. Create a random sample of the training data (some samples may repeat).
5. Build a decision tree:
   1. At each decision point in the tree:
      * Randomly pick F features.
      * Find the best feature and split point among the F features to split the node.
      * Split the node into two child nodes based on the selected split point.
   2. Keep splitting until the tree is fully grown or meets some stopping condition (like a max depth or minimum samples).
6. Add this decision tree to the 'forest'.
7. To make a prediction for a new sample x:
8. Let each tree in the forest make a prediction.
9. The final prediction is the one that most trees agree on (majority vote).

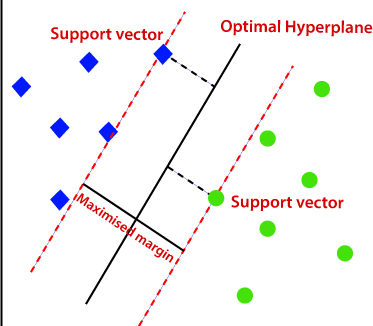
## Support Vector Machine

Support Vector Machine is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features.

## Types of SVM

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

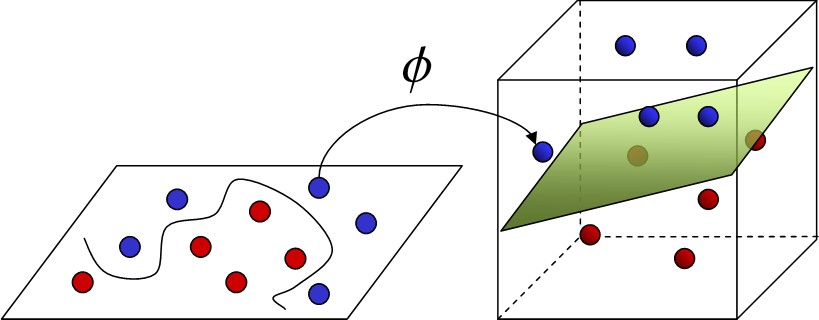
Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher- dimensional feature space, where the data points can be linearly separated. A linear SVM is used to locate a nonlinear decision boundary in this modified space.



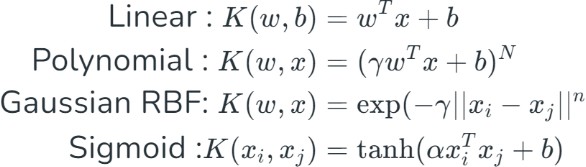
The best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category is called a hyperplane. There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors. The distance between the vectors and the hyperplane is called as margin. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers. So, the margins in these types of cases are called soft margins.

If the data is non-linear, we cannot draw a single straight line. So, to separate these data points, we need to add one more dimension. For that, SVM kernel function is used.



The SVM kernel is a function that takes low-dimensional input space and transforms it into higher- dimensional space, i.e. it converts nonseparable problems to separable problems. It is mostly useful in non-linear separation problems. Some commonly used kernel functions are:



## Working

* Gather dataset which consists of feature vectors (inputs) and corresponding labels (outputs).
* Represent each data point as a vector in a multi-dimensional space, where each dimension corresponds to a feature of the data.
* Find the hyperplane that maximizes this margin, ensuring the best possible separation between the two classes.
* If the data is not linearly separable, the original feature space is mapped into a higher- dimensional space by using kernel trick where a linear hyperplane can separate the classes.
* Find the weights and bias that define the hyperplane by maximizing the margin and minimizing the classification errors.
* Once trained, the SVM model can classify new data points. For each new point, the model checks on which side of the hyperplane the point falls.

## Pseudocode

1. Initialize Parameters:
   * Set the regularization parameter C.
   * Initialize weights w and bias b to 0 or small random values.
2. Training Process:
   * Repeat for a fixed number of iterations:

a. For each data point (xi, yi) in the training set:

1. Calculate the decision function: f(xi) = w \* xi + b
2. If the data point is misclassified (i.e., yi \* f(xi) < 1):
   * Update weights: w = w + C \* yi \* xi
   * Update bias: b = b + C \* yi
3. Prediction:
   * For a new data point x\_new, compute: f(x\_new) = w \* x\_new + b
   * Return the predicted class:

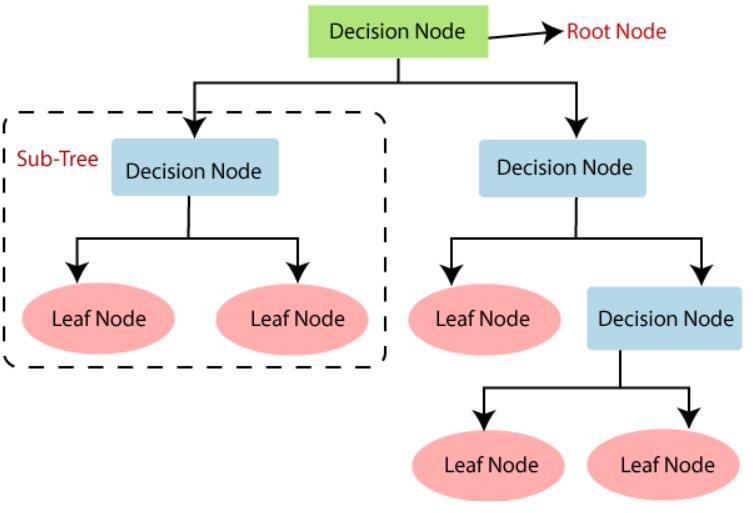
if f(x\_new) >= 0, return +1

else return -1

## Decision Tree

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

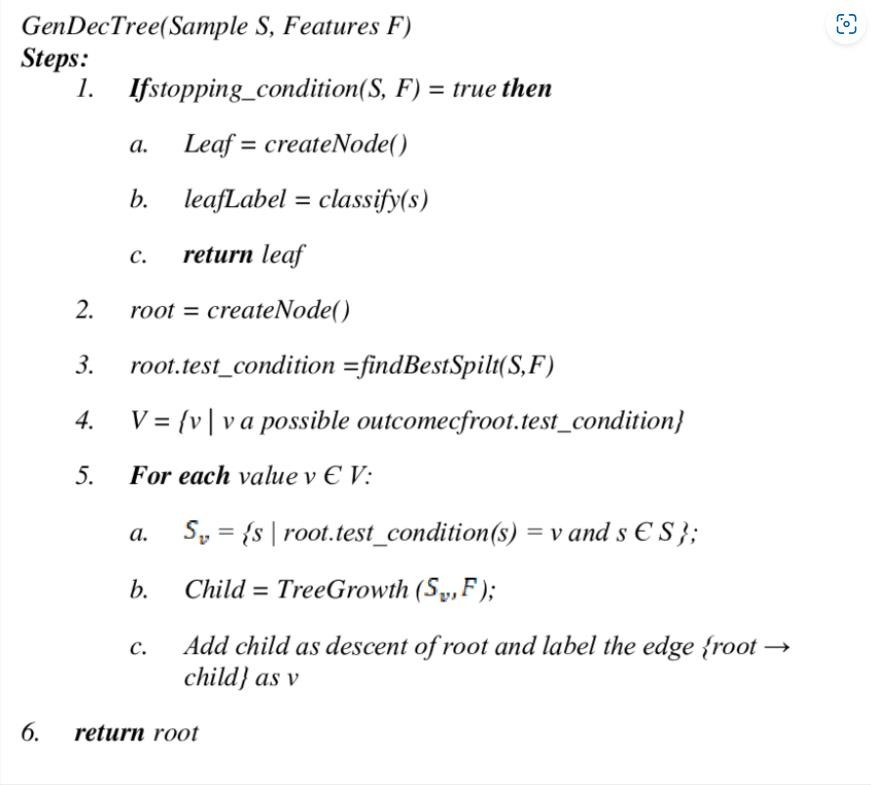
## General Structure



**Working**

* Splitting: At each node, the algorithm assesses the input features and decides the best way to split the data based on a specific criterion (like Gini impurity or entropy) that aims to maximize the separation of classes or minimize error.
* Traversing: The decision tree traverses down from the root node to the leaf nodes based on the feature values of the input data. Each decision point (node) corresponds to a specific test on an input feature.
* Leaf Nodes: Once a decision path is taken, one reaches a leaf node which provides the output prediction or class label based on the majority class of instances that reached that leaf.

**Pseudocode**



# PACKAGES AND FUNCTIONS

* Pandas
  + read\_csv()
  + head()
  + shape()
  + info()
  + value\_counts()
  + isnull()
  + describe()
  + tail()
* Matplotlib
  + show()
  + xlabel()
  + ylabel()
* Seaborn
  + boxplot()
  + countplot()
  + scatterplot()
* Numpy

# PROJECT PIPELINE

Model Training

(RF, DT, SVM)

Splitting of Dataset

Exploratory Data Analysis

Dataset

Train Data

Data Preprocessing

Test Data

Asthma Classification

Model

Input

**TIMELINE**

* Submission of project synopsis with Journal Papers - 22.07.2024
* project proposal approval - 26.07.2024
* Presenting project proposal before the Approval Committee - 29.07.2024 & 30.07.2024
* Initial report submission - 12.08.2024
* Analysis and design report submission - 16.08.2024
* Verification of the report and PPT by the guide - 19.08.2024
* First project presentation - 21.08.2024 & 23.08.2024
* Sprint Release I - 30.08.2024
* Sprint Release II - 26.09.2024
* Interim project presentation - 30.09.2024 & 01.10.2024
* Sprint Release III - 18.10.2024
* Project execution, submission of project report and PPT before the guide - 24.10.2024
* Submission of the project report to the guide - 28.10.2024
* Final project presentation - 28.10.2024 & 29.10.2024
* Submission of project report after corrections - 01.11.2024

**SYSTEM DESIGN**

**Model Building**

The model planning phase for asthma detection focuses on developing a structured approach for diagnosing asthma conditions using machine learning techniques. It involves selecting algorithms, defining the overall scope, and outlining strategies to accurately classify asthma conditions based on clinical features and patient history.

1. Objective:

Build an asthma classification system using machine learning techniques to predict and categorize asthma conditions. The goal is to accurately classify asthma-related conditions, such as asthma and non-asthma, using clinical features, patient demographics, lifestyle factors, and other relevant indicators.

2. Approach:

Use a comparative study of three machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT). These models will be evaluated for accuracy, precision, recall, and robustness in detecting asthma conditions using a pre-processed dataset.

3. Data Preparation:

Import and preprocess the dataset that includes patient data with clinical features and other relevant indicators (e.g., age, gender, lung function measurements, symptoms).Handle missing values using appropriate techniques such as mean/mode imputation, or advanced methods. Split the dataset into training and test sets (e.g., 80% training, 20% testing) for model evaluation.

4. Exploratory Data Analysis (EDA):

Perform EDA to understand the distribution of asthma conditions across the dataset,analyse feature distributions and relationships between clinical indicators and different asthma conditions to identify patterns that help in classification.

5. Model Building:

Implement the machine learning algorithms: Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). Train the models using the dataset features and evaluate their performance on the test set using evaluation metrics such as accuracy, confusion matrix and classification report to assess the models’ predictive abilities.

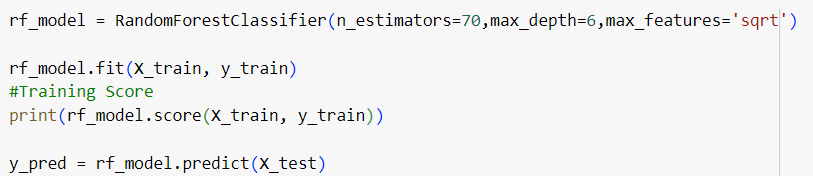
6. Model Comparison:

Compare the performance of the three algorithms based on accuracy, precision, recall, and computational speed. Choose the best-performing model based on its ability to accurately detect asthma conditions in unseen data.

**Model Training**

The dataset was divided into two parts. X representing the input features, and y representing the diagnosis. The training set consists of 80% of the data and is used to train the model.

**Random Forest**



0.9321940463065049

A Random Forest Classifier, which is an ensemble learning method based on multiple decision trees is used to train the model. The model consists of 70 decision trees. The maximum depth of each tree is limited to 6 levels. The number of features considered for splitting at each node is the square root of the total number of features.

The model achieved an accuracy of 93.21% on the training data.

**SUPPORT VECTOR MACHINE**



0.818632855567806

Support Vector Machine is used to train the model. The C parameter is the regularization parameter. Smaller C values (like 0.1) create a more flexible decision boundary by allowing some misclassifications on the training data, which can help prevent overfitting and generalize better to the test data. The Radial Basis Function (RBF) kernel was selected for this model. The RBF kernel is effective for non-linear decision boundaries, allowing the model to capture complex relationships between features.

The model achieved an accuracy of 81.86% on the training data.

**DECISION TREE**

****

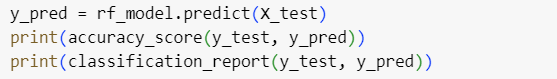
0.8784454244762955

We are building a decision model and will only look at a random subset of the features when deciding how to split the data at each step. Specifically, the size of the subset is the square root of the total number of features. max\_ depth=12 limits the model to 12 levels deep. It prevents the tree from growing too big and helps avoid overfitting .

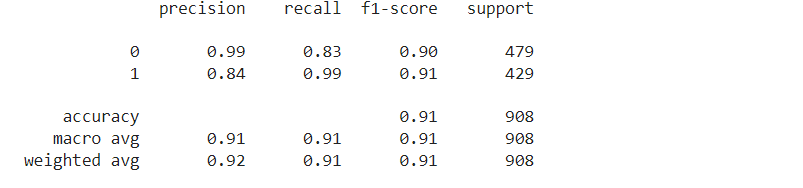
The model achieved an accuracy of 87.84% on the training data.

**Model Testing**

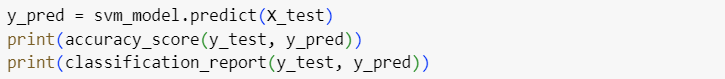
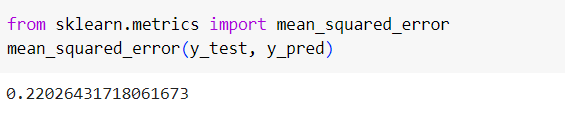
A close-up of a text

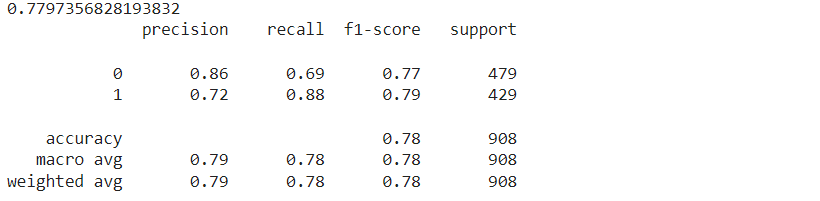
Description automatically generated

0.9063876651982379

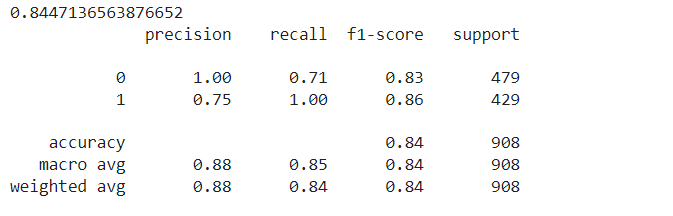
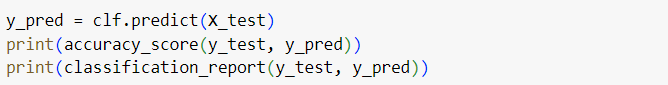
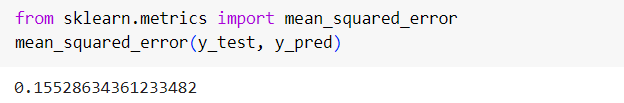


Random Forest achieved an accuracy of 90.63% on the testing data and mean squared error is 0.09.





The SVM achieved an accuracy of 77.97% on the testing data and the mean squared error is 0.22.



The Decision Tree model achieved an accuracy of 84.47% on the testing data and the mean squared error is 0.15.