**Python for Machine Learning**

**Final Term Project - Report**

1. Vinay Sai Ravulapati – CB.EN.U4AIE21153

Content:

|  |  |  |
| --- | --- | --- |
| 1 | Introduction |  |
| 2 | Motivation |  |
| 3 | Data Collection and Preprocessing |  |
| 4 | Machine learning algorithms |  |
| 5 | Conclusion |  |

**Topic**:

**Heart Disease Classification and Analysis using Machine Learning Algorithms Logistic Regression, SVM, KNN.**

**INTRODUCTION**:

Heart disease is a significant health concern globally and early detection is crucial for effective treatment and prevention. Machine Learning algorithms have shown promising results in the field of healthcare by enabling accurate classification and analysis of heart disease. This report presents a study that focuses on heart disease classification and analysis.

The objective of this study is to build predictive models that can accurately classify patients as having or not having heart disease based on various medical attributes. The dataset used in this study contains information such as age, sex and other relevant factors are individually analyzed to understand their impact on heart disease. Bar plots are used to visualize the relationships between these features and the target variable. The algorithms Logistic Regression, Support Vector Machines and K-Nearest Neighbors are implemented in this study. These algorithms are chosen due to their effectiveness in classification tasks and their suitability for the heart disease dataset. The accuracy score is used as the performance metric to assess the model performance.

**Machine Learning:**

* **Supervised:** We are given input samples (X) and output samples (y) of a function y = f(X). We would like to “learn” f, and evaluate it on new data. Types:
* **Unsupervised:** Given only samples X of the data, we compute a function f such that y = f(X) is “simpler”.

**Techniques:**

Machine learning algorithms we Implemented:

**Supervised Learning:**

* Logistic Regression
* Support Vector Machines (SVM)
* K-Nearest Neighbors

**Motivation:**

The motivation behind this study is to develop an effective machine learning model for the classification of heart disease. By leveraging machine learning algorithms, we aim to assist healthcare professionals in making accurate predictions and identifying individuals at risk of heart disease.

The motivation for using machine learning techniques in this context stems from several factors:

1.**Improved Accuracy**: Machine learning algorithms have the potential to analyze large and complex datasets, identifying patterns and relationships that may not be readily apparent to human observers.

2.**Early Detection:** Early detection of heart disease is crucial for effective treatment and management. Machine learning models can learn from historical data and identify subtle patterns and risk factors that might be indicative of heart disease at an early stage. The model can be trained to consider a wide range of variables and personalize the prediction process based on an individual’s unique characteristics.

3.**Decision Support:** Machine learning models can act as decision support tools for healthcare professionals. By providing insights and predictions based on data analysis.

By leveraging these techniques, we aim to contribute to the advancement of cardiovascular healthcare and ultimately enhance patient care and outcomes.

**Data Collection and Preprocessing**:

**Steps**:

1.**Dataset Selection**: We selected a specific dataset for heart disease classification. The dataset contains information about individuals, including demographic data, medical measurements and diagnostic features.

2.**Importing Libraries**: We began by importing essential libraries such as numpy, pandas, matplotlib, seaborn and sklearn. These libraries provide functions and tools for data manipulation, visualization and machine learning algorithms.

3.**Loading the Dataset**: We loaded the heart disease dataset using the pandas library. The dataset was stored in a CSV file format and the ‘read\_csv()’ function used to read the file and create a dataframe object.

4.**Exploring the Dataset:** To understand the structure and content of the dataset we performed several exploratory steps. We checked the shape of the dataset using the ‘shape’ attribute, which gives the number of rows and columns. We also printed a few rows using the ‘head()’ function to examine the column names and sample data.

5.**Data Description:** We utilized the ‘describe()’ function to obtain descriptive statistics of the dataset. This helps in understanding the distribution and range of values for each column.

6.**Data Information:** The ‘info()’ function was used to retrieve information about the dataset including the column names, data types and the presence of any missing values. This step helps in identifying the need for data cleaning or handling missing values.

7.**Checking for Missing Values:** We found that there were no missing values in the dataset. This is crucial for machine learning models as missing values can impact the performance and accuracy of the model.

8.**Analyzing the Target Variable:** We examined the target variable which indicates the presence or absence of heart disease. We checked the unique values and their frequencies using the ‘unique()’ and ‘value\_counts()’ functions respectively. This allows us to understand the distribution of the target variable.

9.**Checking Correlation**: We calculated the correlation between features in the dataset to identify any significant relationships. We created a correlation matrix using the ‘corr()’ function and visualized it using a heatmap. This step helps in understanding the interdependence of features and identifying potentials predictors of heart disease.

10.**Feature Analysis**: We analyzed specific features in the dataset that are known to be associated with heart disease. We used visualization techniques such as bar plots to examine the relationships between these features and the target variable.

11.**Train-Test Split**: To evaluate the performance of machine learning models we split the dataset into training and testing sets. The ‘train\_test\_split()’ function from sklearn was used to randomly divide the data into a specified ratio.

12.**Scaling the Data:** As part of data preprocessing we applied feature scaling to the dataset. This step ensures that all features are on a similar scale which is important for certain machine learning algorithms. We used ‘StandardScaler()’ function from sklearn to standardize the features.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

**Machine learning algorithms and result:**

1. **LOGISTIC REGRESSION:**

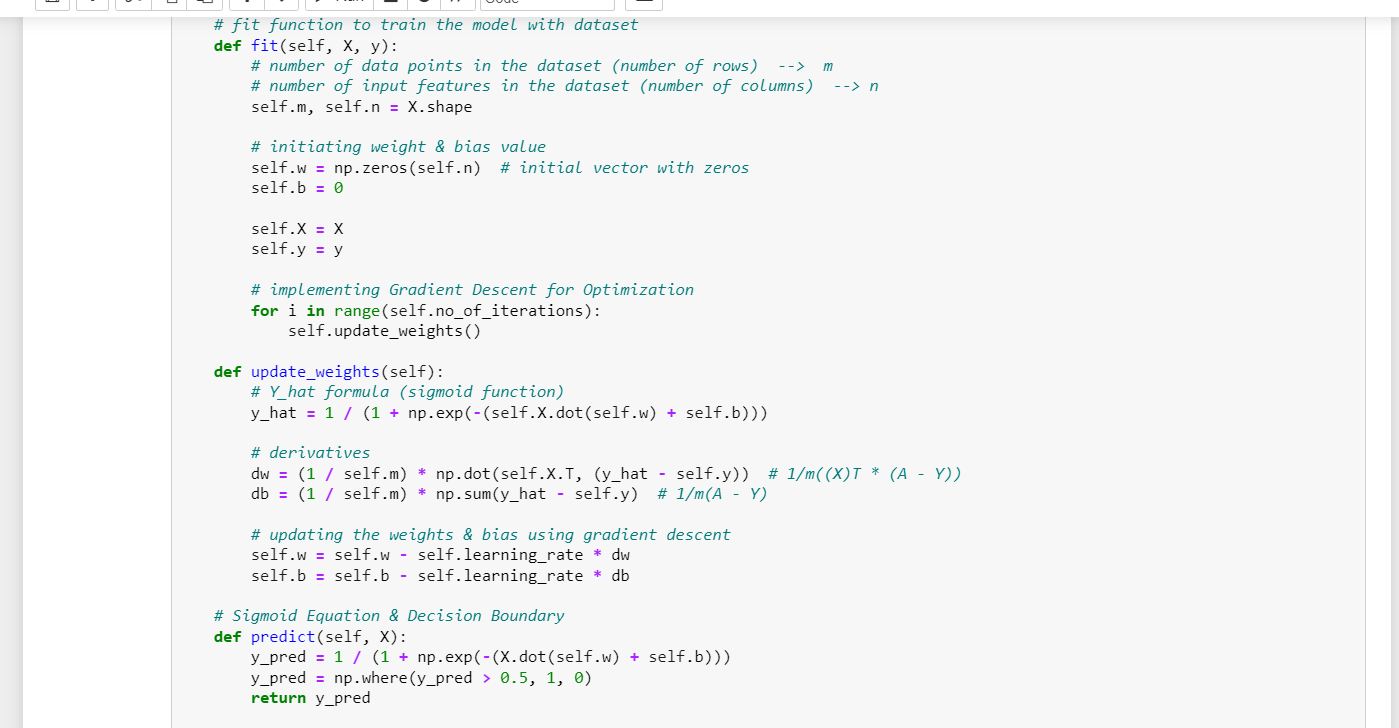
* Logistic Regression is a Machine Learning Classification Algorithm that is used to predict the probability of a categorical dependent variable.
* In logistic Regression, the dependent variable is a Binary variable that contains data coded as(yes, success, etc.) or (No, failure, etc.)
* In other words, the Logistic Regression model predicts P(Y=1) as a function of X.
* Logistic regression models the probability of the default class (e.g. 0)
* A logistic regression model makes a prediction by combining the weighted sum of the input features with a bias term
* The model estimates the probability of the default class using the logistic (sigmoid) function
* The decision boundary is the line that separates the area where the model predicts a class of 0 and where it predicts a class of 1.

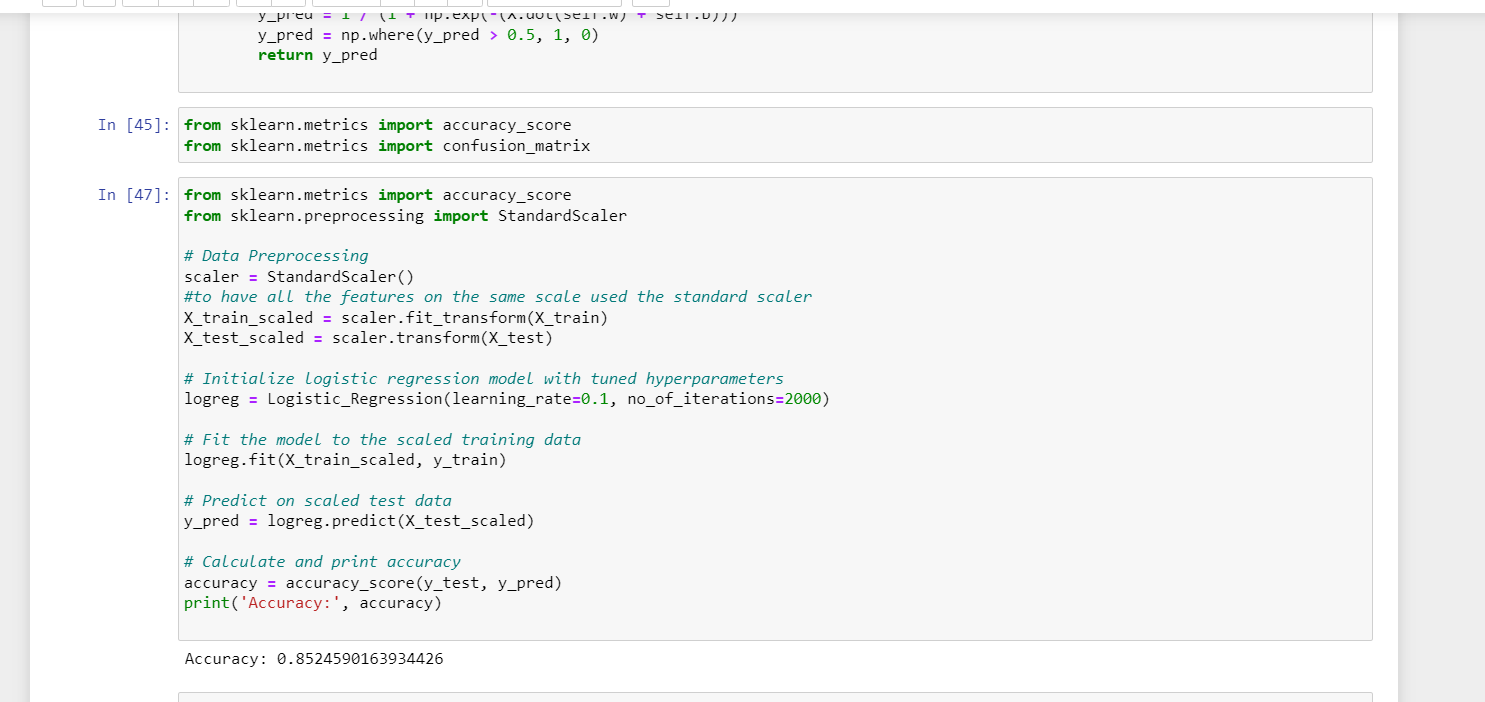
A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence



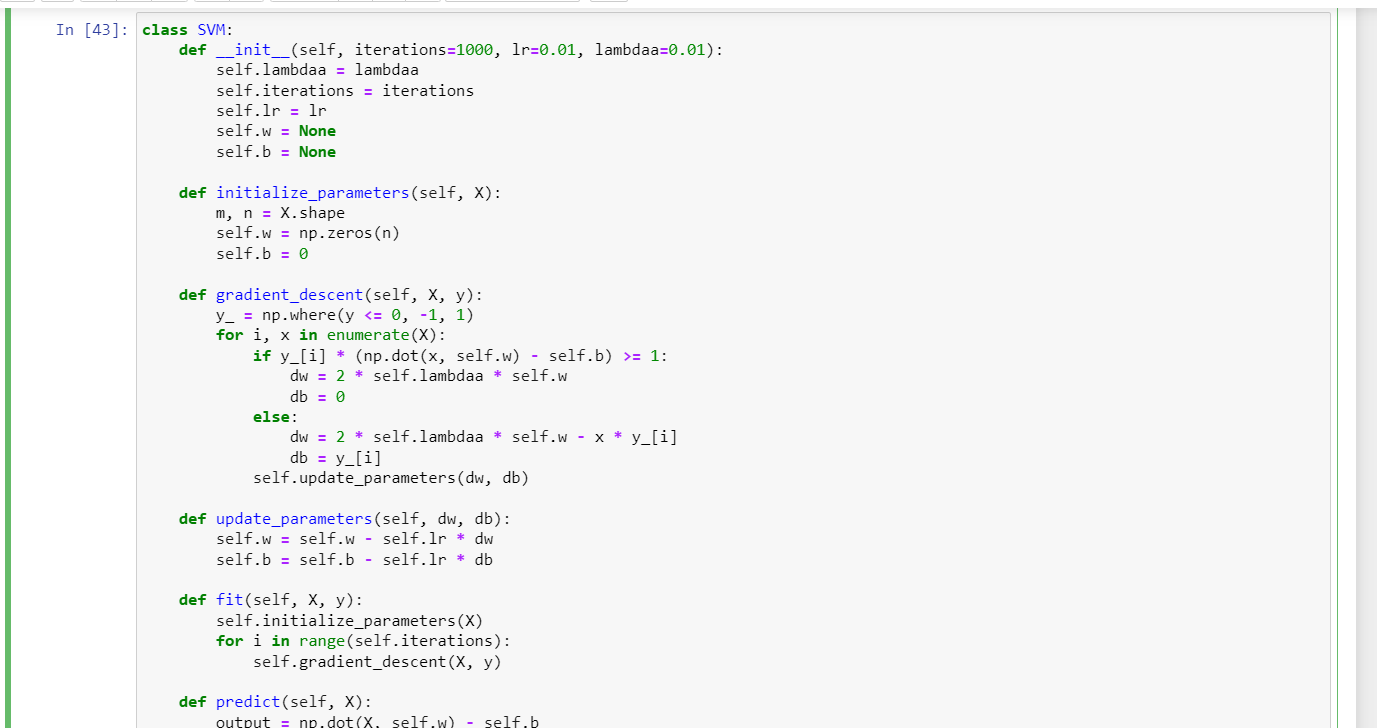


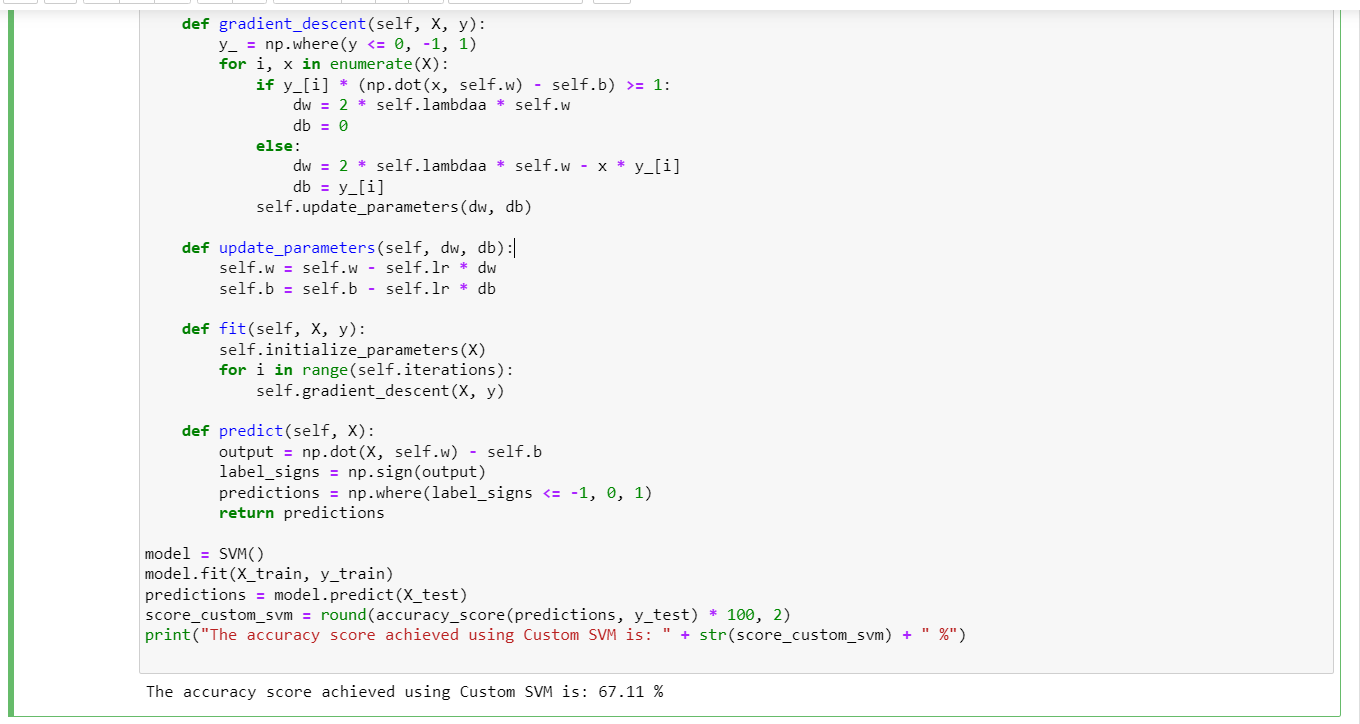
2. **SVM:**

* Support Vector Machine is a powerful supervised learning algorithm used for both classification and regression tasks.
* SVM aims to find an optimal hyperplane that separates the data points into different classes while maximizing the margin between the classes.
* SVM aims to achieve a robust and generalized decision boundary that can handle new and unseen data differently.
* The dataset is divided into train and test which defines in SVM the decision boundary is determined by the SVM algorithm based on the provided kernel.
* Support vectors are the data point that lie closest to the decision boundary they play crucial role in defining the decision boundary and margin.
* The formulation of SVM with a soft margin the objective function aims to minimize the weight vector ‘w’ while allowing for misclassifications based on the regularization parameter ‘c’. The term ‘lambda’ represents the regularization parameter in the custom SVM implementation.

A screenshot of a computer program

Description automatically generated with low confidence





3.**KNN**:

* KNN is Supervised Machine Learning Model where the model predicts by learning from the past data available
* KNN is used for classification based on feature similarities
* Here the data will be classified based on how its neighbours are classified
* So the new input data is classified based on the similarities of the input which we used for training the model
* In KNN, K is an parameter that specifies the no of neighbours it should consider to classify the input data
* We cannot give whatever value we want for K because it is an important factor for better accuracy of the model
* So generally there are two ways:

1. Some will consider K= (where n is no of data points)
2. A random value which is odd but much greater(frequently used K=3,5)

* Here if we take big K value it will lead us to a confusion and classification can also not be that much perfect.
* To find the K-nearest neighbours we will calculate the distance
* Euclidean distance =

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

**Conclusion**:

In conclusion the Logistic Regression, SVM, KNN were applied to the dataset for classification tasks. The accuracy scores were calculated to evaluate the performance of the models which allows us to make informed decisions. We have achieved a better understanding of their performance in the predicting the target variable.

INFERENCE: