TERM PAPER OF THE PROJECT

Introduction:

Healthcare is a crucial domain in every society, and there is a need to provide timely and accurate diagnoses to improve patient outcomes. Machine learning and deep learning techniques can help healthcare professionals diagnose and treat diseases more efficiently. The aim of this term paper is to discuss the implementation of three machine learning models - Gaussian Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest and Logistic Regression - for the classification of 42 diseases. The code implementation for these models is discussed in detail, along with the evaluation metrics used to assess their performance.

Background:

Machine learning algorithms have been widely used in healthcare to assist medical professionals in diagnosis, disease prediction, and treatment planning. In recent years, deep learning models have shown exceptional performance in various medical applications, including image and signal analysis, prediction, and classification of various diseases. The early and accurate detection of diseases can save lives and reduce healthcare costs. Therefore, it is essential to use appropriate models to diagnose and classify diseases. This term paper describes the implementation of three machine learning models to classify 42 diseases.

Methods:

The code implementation for three machine learning models - Gaussian Naive Bayes, Support Vector Machine (SVM), Decision Tree and Logistic Regression - for the classification of 42 diseases is discussed in detail below.

Data Preparation:

The data used in this project consists of a set of features that describe various medical conditions, as well as a target variable indicating which disease is present. The data is split into a training set and a testing set, with the training set being used to train the machine learning models and the testing set being used to evaluate their performance.

Before training the models, we pre-process the data using the Simple Imputer class from the Scikit-Learn library to fill in any missing values with the mean value of that

feature. We then use the Label Encoder class to encode any categorical features as binary features. The training and testing data are concatenated and Label Encoded, and then split back into the original training and testing sets.

Model Training:

We use three different machine learning models to classify the diseases: Gaussian Naive Bayes, Support Vector Machines (SVM), Logistic Regression and Decision Tree. These models are all implemented using the Scikit-Learn library.

After training the models on the training data, we use them to make predictions on the testing data. We compute the accuracy, precision, recall, and F1 score for each model to evaluate their performance. The results are printed to the console, and a confusion matrix is also generated and plotted using the seaborn library.

Method for Random Forest:

Importing the required libraries: The code imports the necessary libraries such as NumPy, pandas, sklearn, and metrics.

Reading the datasets: The code reads two CSV files, 'Training.csv' and 'Testing.csv', and stores them in pandas data frames, 'trainset' and 'test set', respectively.

Splitting the dataset: The code splits the datasets into features (x_train, x_test) and the target variable (y_train, y_test) using the iloc method. The first 132 columns are considered as features, and the last column is considered as the target variable.

Creating a Random Forest Classifier object: The code creates a RandomForestClassifier object with 100 decision trees (n_estimators = 100).

Fitting the model: The code fits the Random Forest Classifier on the training set (x_train, y_train) using the fit() method.

Making predictions: The code predicts the target variable for the testing set (x_test) using the predict () method.

Evaluating the model: The code evaluates the performance of the Random Forest Classifier using the accuracy_score() and confusion_matrix() methods from the metrics library.

Result:

The results of our experiments show that all of the models were able to achieve high levels of accuracy in classifying the diseases. The Gaussian Naive Bayes model achieved an accuracy of 100%, the SVM achieved an accuracy of 100%, and the Logistic Regression model achieved an accuracy of 100%, Decision Tree achieved an accuracy of 97.61% and Random forest achieved an accuracy of 97.67% .For the all the models precision score, re-call score and F1 score is also 100%.

Conclusion:

In this project, we have explored the use of machine learning and deep learning models for the task of disease classification. We have shown that these models can achieve high levels of accuracy in classifying a set of 42 diseases. The results demonstrate the potential of these techniques for improving the diagnosis and treatment of medical conditions.