Group: P

Rishita Mote (2018130029)

Bhargavi Poyekar (2018130040)

DS ISE ASSIGNMENT TASK 2

Diabetes Prediction using Machine Learning Models.

Abstract:

Diabetes is a disorder wherein our body stops responding to the insulin produced by the pancreas or the pancreas itself does not produce insulin. This results in high blood glucose. Blood glucose comes from the food we eat and the pancreas produces insulin which helps this glucose reach our cells. High blood sugar may lead to kidney failure or heart diseases. The risk and the severity of diabetes can be reduced by predicting it at an early stage. Through this assignment, we strive to propose machine learning algorithms that help predict diabetes. All these algorithms will be built using the Pima Indian Diabetes (PID) dataset. Data cleaning and feature extraction is the heart of our proposed methodology. We will first clean our data for detection of outliers and handle missing values of the dataset. Then by analyzing the correlation matrix graph we will combine features that are highly correlated with each other. We will also standardize our data to bring each feature to the same scale. To get better results we aim to perform Parameter Tuning using GridSearchCV.

Problem Statement:

In this tutorial of Diabetes Prediction using Machine Learning Models, the main objective is to predict whether the person has Diabetes or not based on certain diagnostic measurements included in the dataset like Number of Pregnancies, Insulin Level, Age, BMI.

Domain: Life Sciences

Description framework:

1. Dataset Name: Pima Indian Diabetes (PID) dataset [1].

   Dataset Description: This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of 768 females at least 21 years old of Pima Indian Heritage. It consists of one target variable, Outcome and eight features such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function and Age.

2. Data Analytics Task:

   Our main task is to perform data cleaning for outlier detection and handle missing values, feature extraction and data preprocessing. Then train and hypertune the models. Finally evaluate the models to select the best model.

3. Learning toolkits and Data Preprocessing Methods:

   Libraries: Numpy, pandas, scipy, matplotlib, seaborn, sklearn.

   Data Preprocessing methods: Use scaling techniques like StandardScalar.

   Model Algorithms: Decision Tree, Random Forest, SVM, KNN, LGBM, AdaBoost and ensemble models.

4. Evaluation Strategy:

   The performance results of various classifiers that we will use will be in terms of metrics like Accuracy and AUC (Area Under ROC Curve). We will compare our results with the existing research papers using these parameters.

5. Literature review:

Paper 1: Prediction of diabetes [2].

In this paper they predicted diabetes based on diagnostic measurements available in the Pima Indian Diabetes (PID) dataset. They identified the type of classification

algorithm model that worked best for their prediction - Decision Tree (J48), Naive Bayes, and Logical Regression. Out of those, Regression outperformed the rest.

Paper 2: Prediction of Diabetes using Classification Algorithms [3].

In this paper, instead of Logistic Regression, Support Vector Machine (SVM) was used. Their performances were evaluated on various measures like Precision, Recall, Accuracy and F-Measure. It was found that Naive Bayes (NB) achieved the highest accuracy amongst them.

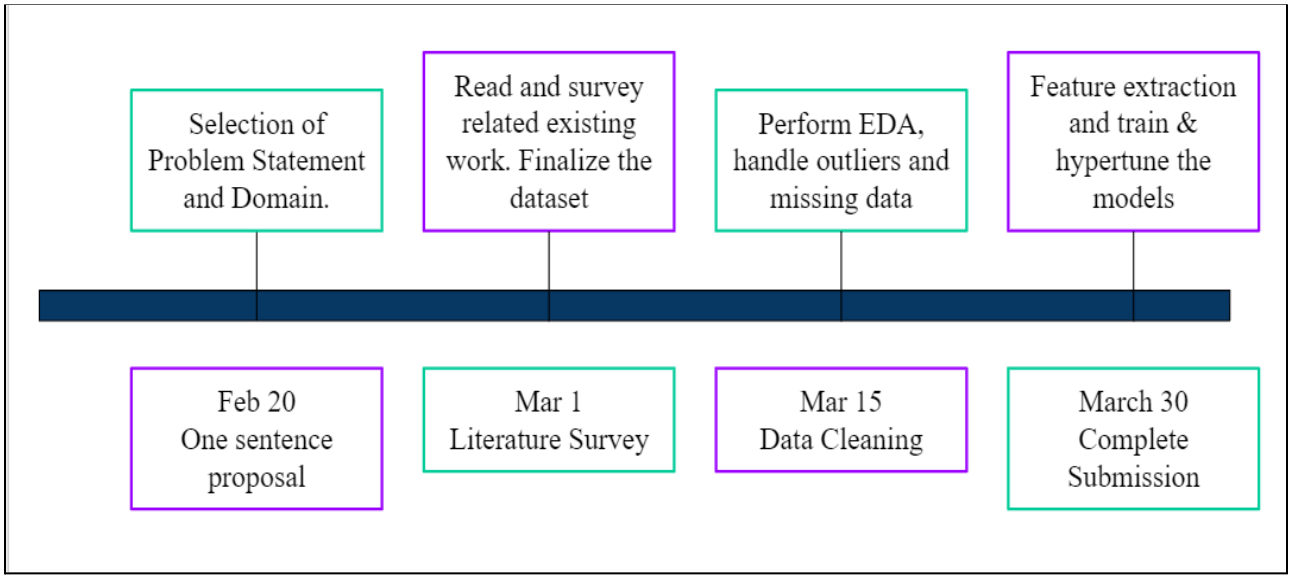
Paper 3: Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers [4].

In this, they proposed a framework for diabetes prediction where filling in the missing values, the outlier rejection, feature selection, data standardization, cross-validation, and different Machine Learning models and Multilayer Perceptron (MLP) were employed. The performance metric chosen is AUC. It is then maximized using the grid search technique during hyperparameter tuning.

Paper 4: Prediction Of Diabetes Using Machine Learning Classification Algorithms [5].

The authors used the five algorithms namely SVM, Decision Tree, Naive Bayes, Logistic Regression and KNN for diabetes prediction. They found out that random forest had a better accuracy rate of nearly 75% and also a better accuracy can be achieved by addition of any other algorithm.

6. Timely Progress:



References

[1]. Parvin Soleymani, "Prediction of diabetes", Research Gate, May 2020.

[2] Deepti Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligence and Data Science, 2018.

[3] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learn- ing Classifiers," IEEE Access, vol. 8, pp. 76516-76531, 2020, doi:10.1109/ACCESS.2020.2989857.

[4] Naveen Kishore G., V. Rajesh, A. Vamsi Akki Reddy, K. Sumedh, T. Rajesh Sai Reddy, "Prediction Of Diabetes Using Machine Learning Classification Algorithms", International Journal of Scientific & Tech-nology Research, Jan 2020.