UNIVERSITY OF ESSEX

Department of Mathematical Sciences

Subject: CE802 Machine Learning

Wordcount: 1079

INTRODUCTION

The healthcare sector is presently experiencing an increasing problem in dealing with the incidence of chronic illnesses such as diabetes. Diabetes treatment may be considerably improved by early identification and intervention, resulting in the betterment of the patient's outcomes and a lower load on healthcare systems. In this context, we are using machine learning (ML) to identify individuals who are at high risk of contracting diabetes gives a great opportunity for healthcare practitioners to engage proactively with preventative treatments. We will outline a report in this proposal that will evaluate the feasibility and effectiveness of applying ML techniques we have used to forecast people at high risk of acquiring diabetes using data from electronic medical records.

Task 1:

In this exercise, we will predict whether or not the patient will be diagnosed with diabetes by applying machine learning methods such as gradient boosting, random forest, and decision trees.

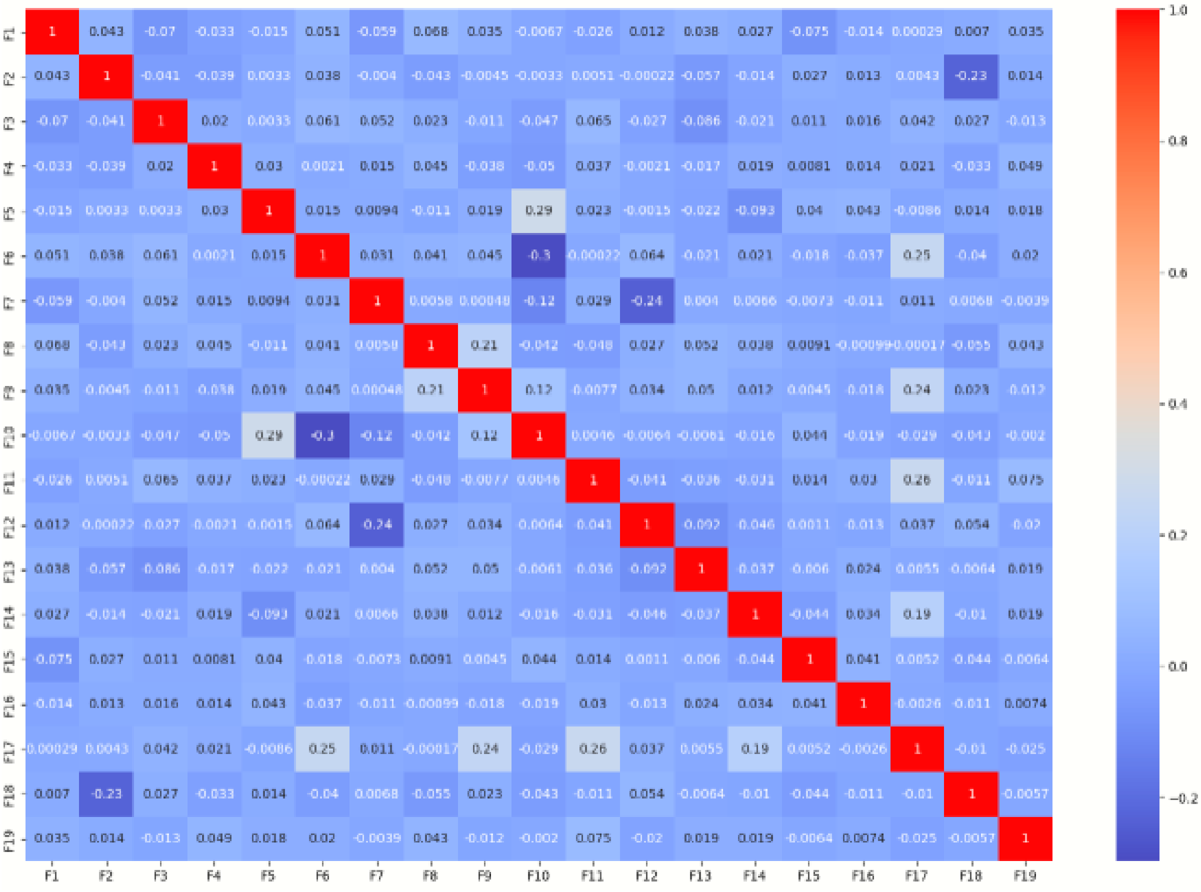
Pre-processing

Pre-processing the data is a key technique for improved prediction in order to develop a model with greater accuracy. We will use the null function to check for null values in the dataset; in our train data, there are over 500 missing values in the F19, which may be fixed by removing the whole column or replacing it with its mean values. The categorical values in a CLASS column (TRUE OR FALSE) are then replaced with binary values (0 and 1) utilizing their replace function.

Correlation matrix:

A correlation matrix is a table that displays the coefficients of correlation between several variables in a dataset. Correlation matrices can be used in machine learning to investigate the correlations between features, detect possible duplicates or unnecessary features, and choose the most important features for the model. Correlation coefficients vary from -1 to +1, with lower values suggesting a stronger linear link between the variables. A heatmap or a scatterplot matrix can be used to visualize a correlation matrix. It is a valuable tool for data analysis and model development since it may increase the accuracy and interpretability of machine learning models.

These is the correlation plot diagram,



	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
F1	1	0.043	-0.07	-0.033	-0.015	0.051	-0.059	0.068	0.035	-0.0067	-0.026	0.012	0.038	0.027	-0.075	-0.014	0.00029	0.007	0.035
F2	0.043	1	-0.041	-0.039	0.0033	0.038	-0.004	-0.043	-0.0045	-0.0033	0.0051	-0.00022	-0.057	-0.014	0.027	0.013	0.0043	-0.23	0.014
F3	-0.07	-0.041	1	0.02	0.0033	0.061	0.052	0.023	-0.011	-0.047	0.065	-0.027	-0.086	-0.021	0.011	0.016	0.042	0.027	-0.013
F4	-0.033	-0.039	0.02	1	0.03	0.0021	0.015	0.045	-0.038	-0.05	0.037	-0.0021	-0.017	0.019	0.0081	0.014	0.021	-0.033	0.049
F5	-0.015	0.0033	0.0033	0.03	1	0.015	0.0094	-0.011	0.019	0.29	0.023	-0.0015	-0.022	-0.093	0.04	0.043	-0.0086	0.014	0.018
F6	0.051	0.038	0.061	0.0021	0.015	1	0.031	0.041	0.045	-0.3	-0.00022	0.064	-0.021	0.021	-0.018	-0.037	0.25	-0.04	0.02
F7	-0.059	-0.004	0.052	0.015	0.0094	0.031	1	0.0058	0.00048	-0.12	0.029	-0.24	0.004	0.0066	-0.0073	-0.011	0.011	0.0068	-0.0039
F8	0.068	-0.043	0.023	0.045	-0.011	0.041	0.0058	1	0.21	-0.042	-0.048	0.027	0.052	0.038	0.0091	-0.00099	-0.00017	-0.055	0.043
F9	0.035	-0.0045	-0.011	-0.038	0.019	0.045	0.00048	0.21	1	0.12	-0.0077	0.034	0.05	0.012	0.0045	-0.018	0.24	0.023	-0.012
F10	-0.0067	-0.0033	-0.047	-0.05	0.29	-0.3	-0.12	-0.042	0.12	1	0.0046	-0.0064	-0.0061	-0.016	0.044	-0.019	-0.029	-0.043	-0.002
F11	-0.026	0.0051	0.065	0.037	0.023	-0.00022	0.029	-0.048	-0.0077	0.0046	1	-0.041	-0.036	-0.031	0.014	0.03	0.26	-0.011	0.075
F12	0.012	-0.00022	-0.027	-0.0021	-0.0015	0.064	-0.24	0.027	0.034	-0.0064	-0.041	1	-0.092	-0.046	0.0011	-0.013	0.037	0.054	-0.02
F13	0.038	-0.057	-0.086	-0.017	-0.022	-0.021	0.004	0.052	0.05	-0.0061	-0.036	-0.092	1	-0.037	-0.006	0.024	0.0055	-0.0064	0.019
F14	0.027	-0.014	-0.021	0.019	-0.093	0.021	0.0066	0.038	0.012	-0.016	-0.031	-0.046	-0.037	1	-0.044	0.034	0.19	-0.01	0.019
F15	-0.075	0.027	0.011	0.0081	0.04	-0.018	-0.0073	0.0091	0.0045	0.044	0.014	0.0011	-0.006	-0.044	1	0.041	0.0052	-0.044	-0.0064
F16	-0.014	0.013	0.016	0.014	0.043	-0.037	-0.011	-0.00099	-0.018	-0.019	0.03	-0.013	0.024	0.034	0.041	1	-0.0026	-0.011	0.0074
F17	0.00029	0.0043	0.042	0.021	-0.0086	0.25	0.011	-0.00017	0.24	-0.029	0.26	0.037	0.0055	0.19	0.0052	-0.0026	1	-0.01	-0.025
F18	0.007	-0.23	0.027	-0.033	0.014	-0.04	0.0068	-0.055	0.023	-0.043	-0.011	0.054	-0.0064	-0.01	-0.044	-0.011	-0.01	1	-0.0057
F19	0.035	0.014	-0.013	0.049	0.018	0.02	-0.0039	0.043	-0.012	-0.002	0.075	-0.02	0.019	0.019	-0.0064	0.0074	-0.025	-0.0057	1

GRADIENT BOOSTING

Gradient Boosting is a popular machine-learning technique for building prediction models. It is an ensemble strategy that combines numerous weak prediction models to generate a powerful predictive model. The technique works by constantly adding new models to the ensemble, with each new model attempting to correct the flaws of the previous models. Gradient Boosting is extremely useful in scenarios with complex relationships and a large number of characteristics in the data. It is presently one of the most often utilized algorithms in machine learning competitions and real-world applications.

Gradient boosting accuracy exists,

```
hyperparameters: {'learning_rate': 0.5, 'max_depth': 6, 'n_estimators': 100}
GBM Accuracy score: 0.9012500000000001
Gradient Boosting Machine using grid: 0.895
```

DECISION TREE

Decision trees are a form of machine learning technique that is often used for classification and regression issues. They operate by dividing the data into subsets based on the values of the characteristics, with each split resulting in a tree node. The divides are determined based on a set of criteria, such as the Gini index or information gain, that maximizes class separation or variance reduction. Once formed, the tree may be used to make predictions on fresh data by traversing the tree from the root to the leaf node corresponding to the expected class or value. Decision Trees are useful in a variety of applications because they are basic and easy to understand.

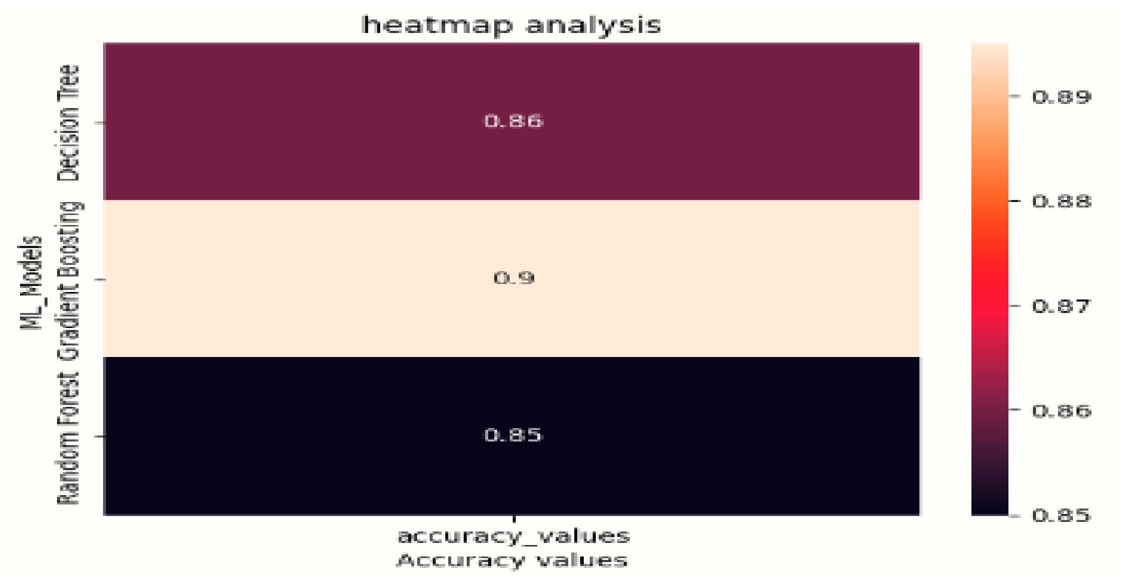
Decision tree accuracy exists of,

```
Decision Tree accuracy: 0.85
Random Forest accuracy: 0.86
```

RANDOM FOREST

Random Forest is a machine learning ensemble learning approach that is an extension of the decision tree algorithm. It works by generating a large number of decision trees, each trained on a different set of data and attributes. Each tree in the forest makes a forecast throughout the training phase, and the final prediction is the average or majority vote of all the trees. Because it can handle a large number of input features, handle missing values, and prevent overfitting, Random Forest is a powerful and popular algorithm. It is commonly utilized in a wide range of applications like classification, regression, and feature selection.

The heatmap study of accuracy in the ml model,



Comparison

By comparing all three models that we have used for this particular task,it is evident that gradient boost is more accurate the predict whether the person is going to be diabetic or not**.**

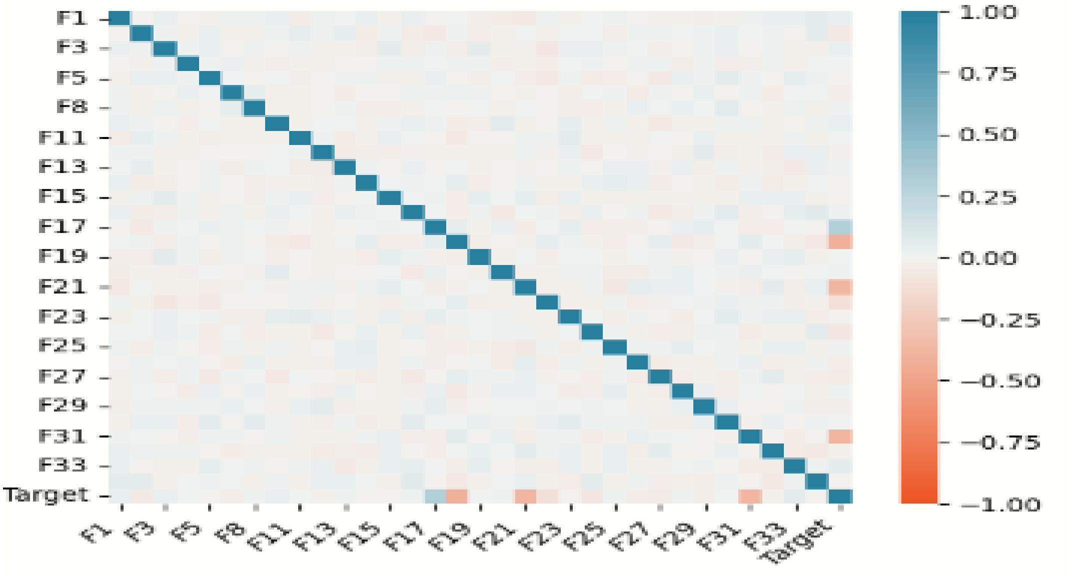
Task 2 :

This talk is about determining the patient’s glucose level

Data pre-processing:

Since the data has no null values, we need to change the categorical values from the data so that we can pre-process the data to create a model for prediction.

The graph below depicts the correlation between variables.

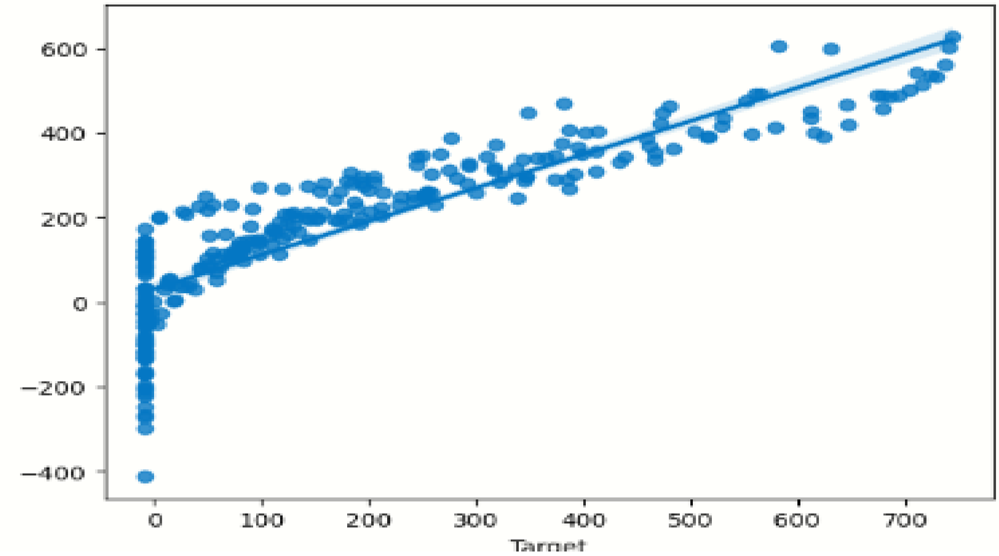


Following that, we will separate the train data into train and test before creating the model.

LINEAR REGRERSSION:

Linear Regression is a prominent machine learning technique that performs regression tasks and is based on supervised learning. It is a statistical approach to predictive analysis that is mostly used to determine the link between variables and predictions. The algorithm predicts continuous/real or numeric variables such as sales, salary, age, product price, and so on…

Mean squared error value- `10794.64495149472`



RANDOM FOREST REGRESSION USING GRID SEARCH CV

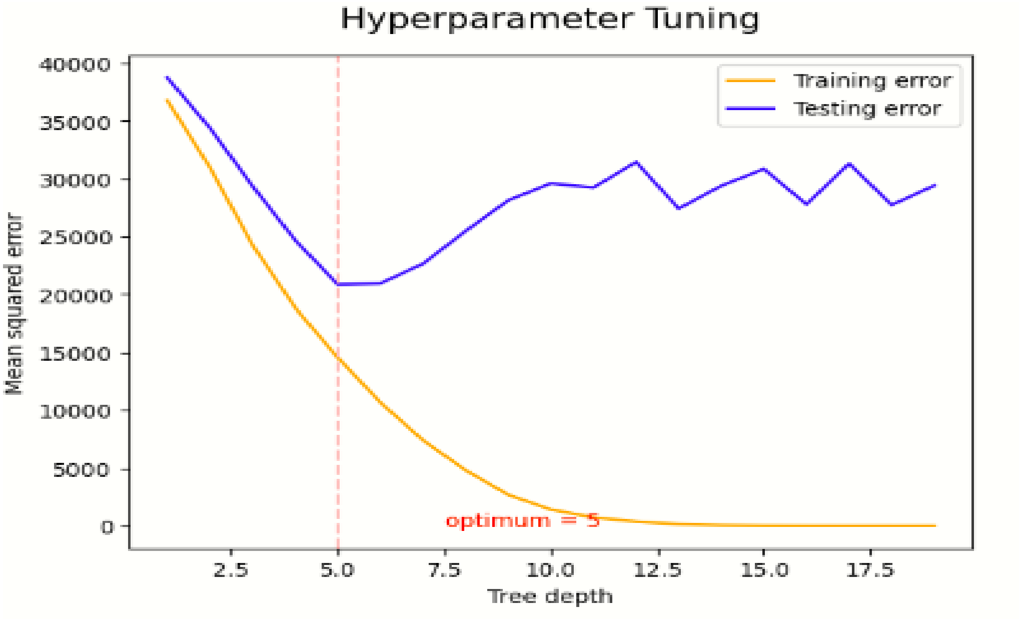
Grid search is a technique for optimizing the random forest model's hyperparameters. Random forest is a well-known machine-learning method for diabetes prediction. Grid search evaluates all possible combinations of the parameters defined in a grid to find the hyperparameter combination that maximizes the model's performance. Grid search may be used to enhance hyperparameters such as the number of trees, the maximum depth of the trees, and how many samples are required to divide a node in half.

Mean squared error value- `20931.888415245026`

DECISION TREE REGRESSOR

Grid search is a technique for optimizing the hyperparameters of the decision tree regressor model, a machine-learning methodology for regression issues. Grid search is a strategy that investigates every conceivable combination of hyperparameters put in a grid in order to find the one that maximizes model performance. Grid search may be used to increase hyperparameters like maximum tree depth

Mean squared error value- `19800.329276409037`.



COMPARISON:

By analyzing all of the models that have been utilized, we can conclude that logistic regression is best suited for this data due to its mean square error value.

CONCLUSION:

They have to give some of the graph plots acquiring the value in this diagrams.