UNIVERSITY OF ESSEX

CE802 – MACHINE LEARNING

2213688

Word Count – 1030

Introduction:

About Problem 2:

The Medical client wanted to develop a machine learning model that can determine from a patient's electronic medical information whether or not they will receive a diabetes diagnosis. The given dataset was a collection of labelled historical data, each of which represents a patient and their current stage of diagnosis.

The ML model was created and It was predicted whether the patient has diabetes or not.The target variable we did forecast to the ‘Class’ column. If the patient has been given a diabetes diagnosis, the value will take as “TRUE”, otherwise it will take the value as “FALSE”.

About Problem 3:

An another medical clients wanted to design a system that not only predicted the likelihood of a patient developing diabetes but also the extent to which their average blood glucose level would exceed the diagnostic threshold if untreated.

With the help of given dataset, The model was trained to learn the patterns and relationships in the data, and then tested to ensure that it could accurately predict the extent of blood glucose level exceeding the diagnostic threshold for new patients.

Datasets Overview:

Two datasets have been provided by the medical client for Problem 2 for training and testing purposes. The training dataset consists of 1000 records with 20 columns, plus an additional column called "Class." The 'Class' column, which contains boolean values like 'True' and 'False,' is regarded as the target column.

The 'Class' Column was used as the target variable in the ML Model, while every other Column was used as a features variable. The machine learning model is trained using the

training dataset and target variable data. The machine learning model's performance is assessed solely using the testing dataset that is provided. This dataset contains same kind of features as the training dataset but that's not include the target values.

Problem2- Training Data

F2	F3	F4	F5	F6	F7	F8	F9	F10	...	F12	F13	F14	F15	F16	F17	F18	F19	F20	Class
.0180	0	2429.040	1.5729	-6901.08	-13.5480	-2204.77	-43.1740	-4.3000	...	-1.96	-7144.9	1	-4.4310	2451.300	12843.744	12.90000	-121.720	NaN	False
.4216	1	3365.160	4.3960	-8939.28	-12.7050	-149.17	-43.0300	-3.8674	...	-1.96	-7022.7	0	-6.3270	2274.330	12726.420	10.35480	-975.600	NaN	True
.3124	0	3616.860	1.8304	-8944.38	-7.2657	-1898.57	-42.7656	-4.0435	...	-1.96	-6071.9	0	-4.3770	2111.937	12842.944	21.25500	-362.220	NaN	False
.4268	0	2263.965	1.5357	-8197.68	-12.9780	-2195.97	-45.5070	-5.1800	...	-0.96	-6703.3	0	-2.3304	4110.300	12812.392	15.83700	-318.680	24.8	True
.6242	0	1669.560	1.5597	-8021.43	-8.9535	-2091.79	-45.5200	-6.4990	...	-0.96	-6999.9	1	-3.0663	3516.600	5396.160	9.69801	-285.372	24.3	True

columns

A dataset comparable to the one in Problem2 has been provided by the medical client in Problem3. However, there is additional data in this dataset that can be used to forecast blood glucose levels. 35 columns and 1400 records make up the dataset, which also includes a target variable column that will serve as the model's output. This project's objective is to create a machine learning model that can correctly predict the target variable from the values of other features in the dataset.

To accomplish this, all columns other than the target variable column are viewed as model features.The client has provided a separate test dataset without the target variable column, which will be used to evaluate the model's performance. The machine learning model will be trained using the training dataset, and the target variable will be predicted using the test dataset.

Problem3- Training Data

Out[3]:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	...	F26	F27	F28	F29	F30	F31	F32	F33	F34	Target
0	28.74	-268.77	62.76	4	38.46	-535.96	538.08	18.37	3.34	-11.67	...	6.20	4110.78	-2.36	27.60	922.07	-1006.92	12	28.38	High	184.93
1	60.30	-11.37	61.32	6	76.09	-441.44	-1614.94	1.73	0.88	-10.30	...	23.58	2050.63	-0.43	55.40	1151.12	-493.71	21	59.49	Very high	83.33
2	-59.58	-305.94	67.35	16	104.74	-448.60	-154.42	72.08	11.36	-5.58	...	2.16	1003.06	-2.81	66.80	329.39	-710.04	6	36.69	Very low	-8.93
3	171.36	-51.48	69.72	10	82.82	-470.62	-1056.76	0.00	4.82	-6.01	...	10.72	11991.41	-1.49	58.32	669.98	-953.94	15	31.35	High	11.72
4	-97.74	-54.00	66.69	8	54.23	-372.46	-480.02	18.13	3.90	-13.63	...	3.08	-223.59	0.06	51.74	679.19	-2098.38	9	34.89	Low	484.05

5 rows × 35 columns

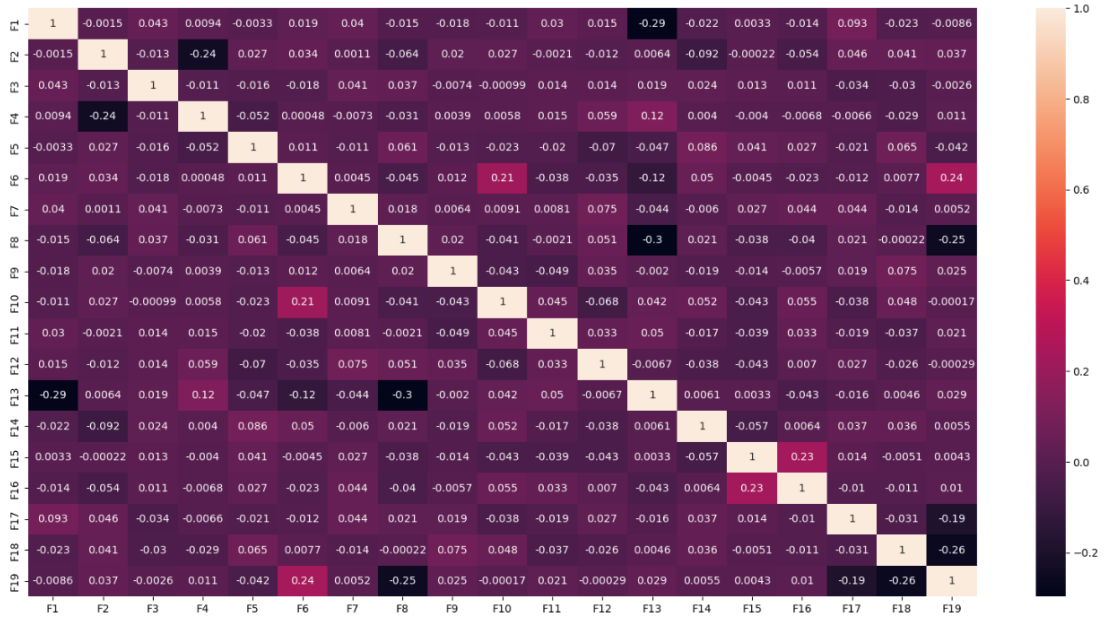
Data Pre-processing:

The given two problems, The data pre-processing was done before creating ML model's.

In the problem2 training data, The 'F20' Column has 500 Null values. So, we removed that column from the features and remaining column was taken as input the ML model.

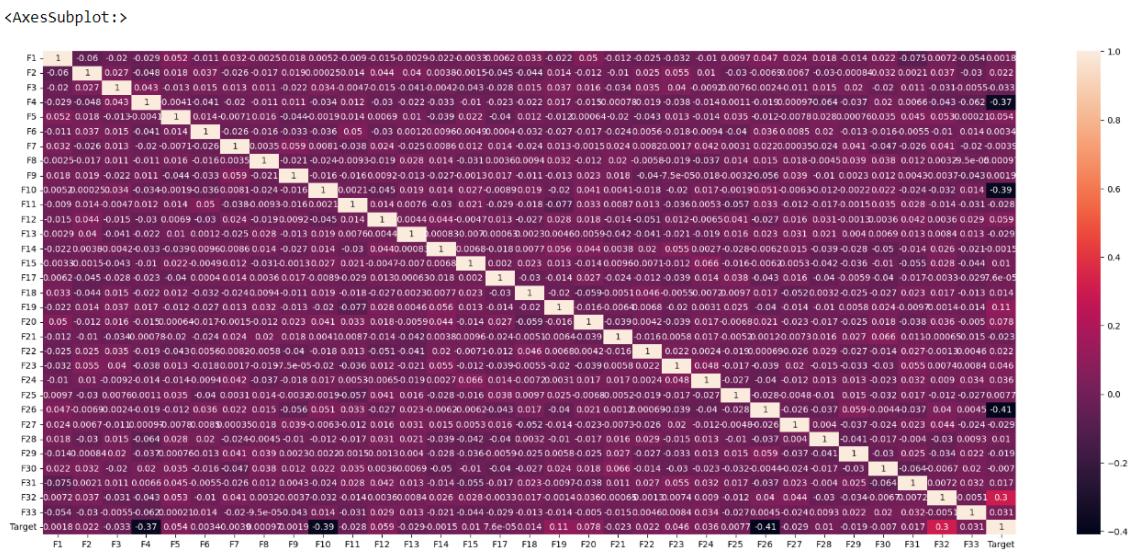
We also used Standard Scalar() Method to vectorize the input data to feed input to the ML model and the Column 'Class' considered as a output Column to the ML models.

The Correlation map of given dataset for the problem2,



In the Problem3, there is NO Null values in the training data. However, we have two columns called ‘F16’ and ‘F34’ has a categorical data. So, we have changed the information from categorical data to binary variables using get_dummies() method in pandas.

The Correlation map of given dataset for the problem3,



The same pre- processing steps was done to the test datasets for both problems1 and 2 respectively.

Model Selection:

The following ML models was trained and developed for the problem2 like,

- Decision Tree
- Random Forest
- Gradient Boosting Machine

For Problem3,

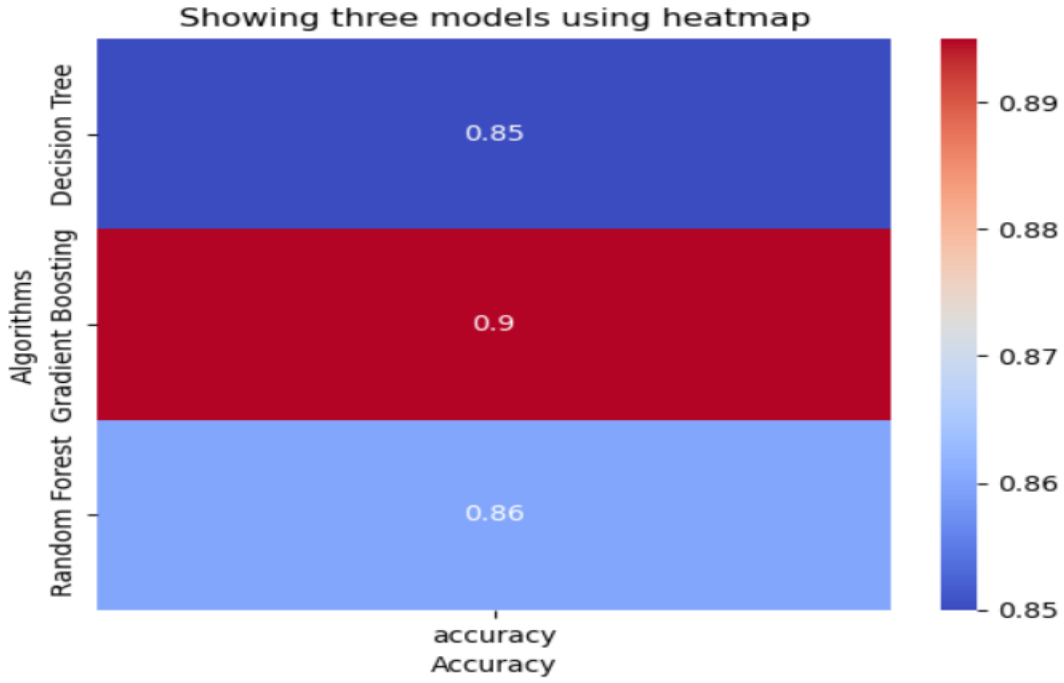
- Linear Regression
- Random Forest
- Decision Tree

Model Evaluation:

For Problems 2 and 3, we trained all machine learning models using the provided datasets. After training each model, we calculated its accuracy rate using the same training data. In order to determine which model to use for predicting the test dataset, we selected the model with the highest accuracy rate.

For Problem 2, the Gradient Boosting Machine model was found to have the highest accuracy rate compared to the other two models. We used the selected model to predict the target variable in the test dataset for Problem 2. By doing this, we were able to evaluate the performance of the model on new, unseen data. The accuracy of the model's predictions was evaluated by comparing them to the actual values in the test dataset.

The following graph shows the accuracy rate of selected three Machine Learning Models for problem2,



The accuracy rate of the Random Forest is 86%, that of the Decision Tree is 85%, and that of the Gradient Boosting algorithm is almost 10% higher than that of the others in the previous graph.

For Problem3, We trained the three models indicated above and used them to estimate the output value. Instead of finding accuracy score in this case, we found mean squared error method.

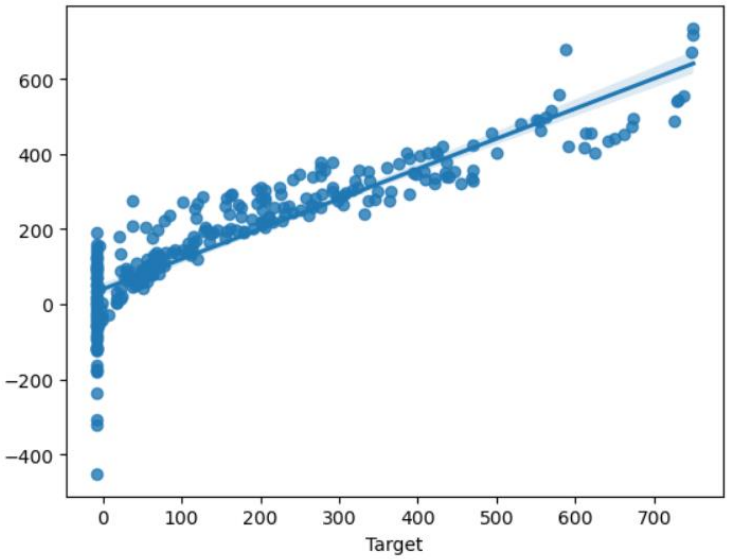
The average squared difference between projected values and actual values is determined by the mean squared error (MSE), which is a mathematical formula. Because it penalises bigger errors more severely than smaller ones, MSE is a frequently used metric for regression tasks. It also reveals the average difference between the projected and actual values. The model is operating better the lower the value of MSE.

This means that I used the linear regression model to make predictions for the output variable in the test dataset, and the evaluated the performance of the model by comparing the predicted values to the actual values in the test dataset.

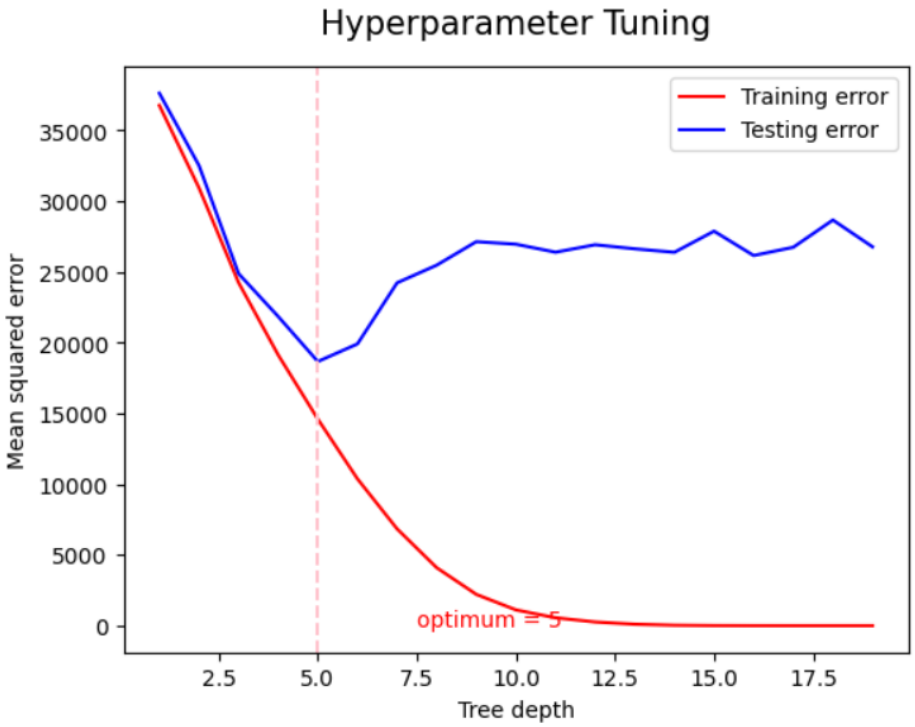
Based on the MSE value, the linear regression model has performed better than the other two models. Therefore, you have decided to use the linear regression model to make predictions for the output variable in the test dataset.

- **The Liner model mean squared error value is 8716.**
- **Random Forest mean squared error value is 21648.**
- **Decision Tree mean squared error value is 21847.**

The Linear Model regression plot as below,



The Hyper tune graph for decision tree classifier,



Conclusion:

The given dataset has been analysed and ML model was trained for the given two problems.