CE802 Machine Learning

Name: Pinky Ramnath Mehta

Registration Number: 2211433

Date: 03/05/2023

Department: School of Computer Science and Electronic Engineering

Word count excluding tables and figures: 1494

Introduction:

The aim of this study is to explore the performance of machine learning techniques for the patient's blood glucose level exceeds the diagnostic threshold if untreated. The four algorithms that used are Linear Regression (LR), Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), and Random Forest Regressor (RFR).

Methodology:

The first step is to load file of training dataset and perform EDA. Upon EDA, the information retrieved was, of 34 columns two of the feature columns are categorical and the rest are numerical values, there are no duplicate or missing values. The 35th column is the target variable which is numerical and is continuous hence it is a regression problem.

For the categorical features to be encoded into numerical we are using OneHotEncoder. Encoding the values to numerical form is necessary as algorithms cannot handle categorical directly. OneHotEncoder preserves the information about the categories while encoding to numerical values. After encoding, we will pass it in original file. Concatenation of original and encoded file is performed.

We need to pre-process the data to improve performance and accuracy of models. Pre-process steps includes ColumnTransform() and StandardScaler() and is done before training models. ColumnTranform() pre-processes both the categorical encoded and continuous data and fits into one dataset. Passing the numerical columns of the dataset with a string variable is not necessary but helps for debugging purposes. StandardScaler() standardizes the numeric features and the encoder is used to one-hot encode categorical features. The first transformer (numeric features) is applied to columns 'F1' through 'F34' and the second transformer (categorical features) is applied to columns 'F29' and 'F32'.

The next step is to split dataset into training and testing for ML algorithms. A new dataframe is created with variable name X by dropping Target column, common pre-processing step. Target variable needs to separate from input features as it is used for prediction. The dropped variable is stored in a new dataframe as variable Y.

Using train_test_split function the X and Y are split into X_train and Y_train which is training subset of data that will be used to train the ML model and X_Test and Y_test being the testing subset of the data which will be used to evaluate the performance of the trained model on new and unseen data.

A pipeline is used to simplify the process of building, evaluating and deploying algorithms. For every pipeline we are passing on pre-processor and the chosen algorithm. The advantage of this step is it helps to reduce the likelihood of error that can occur when transforming the data and training the model separately.

The fit() method is used to train the selected ML algorithms. For instance, we are first making use of Logistic Regression pipeline and we are fitting the X_train and Y_train which are training data with feature and target variables. This is basically done for training the model.

With the trained pipeline, we predict the target values for X_test and predicted score is stored in variable.

Score() method used, helps in calculating coefficient of determination R^2 of prediction for the test set. Higher R^2 indicates it a better fit of model to the test data.

Types of Evaluation performed in our Regression Model:

1. **Median Absolute Error (MAE) and Root MAE**

Evaluating with MAE and its Root is, the given dataset contains outliers. MAE is the median of the absolute difference between predicted and actual values. Reason why median is performed as the metric is more robust to the presence of outliers. Square root MAE puts the error metric back into the same units as the original target variable, making it easier to interpret.

2. **Mean Squared Error (MSE) and Root MSE (RMSE).**

Reasons for this approach is they both measure average difference between actual and predicted values of target variable and the RMSE is square root of obtained MSE score. The lower the values of MSE and RMSE, the better is the model's performance. Both MSE and RMSE gives a higher weight to large error, which makes them particularly sensitive to outliers.

3. **R-Squared (R2) score.**

It is commonly used in regression analysis to evaluate the performance of model with range between 0 and 1 where 1 indicating that model perfectly fits the data and 0 indicating does not fit the data at all and to compare different models.

R2 score is used as one of the evaluation metrics for regression models to check the model's predictive power and how well it can generalize to unseen data.

4. **Cross Validation with KFolds.**

By setting k=5, we are performing 5-fold cv, which means that the data will be divided into 5 equally sized subsets, and the model is trained and evaluated 5 times, with a different subset used for validation each time. The cvs function returns an array of 5 scores, one for each fold.

The reason we use cv is to estimate the performance of the model on an independent dataset. By using multiple folds, we can get a more reliable estimate of the model's performance than by just using a single train/test split. This is especially useful when the dataset is small, and we want to make the most out of the available data. In addition, cross-validation helps to reduce overfitting, as it provides a more accurate estimate of the model's generalization performance.

As the obtained scored are displayed in form of array we are passing the scoring values as Median Absolute Error and taking the mean of the square root of cvs.

Final step of cross validation is by taking the Mean, Median and Standard Deviation of the obtained cross_val_score. The mean score gives an indication of the average performance of the model, while the median provides a measure of central tendency that is less sensitive to extreme values. The standard deviation gives an indication of the variance in the performance of the model across the different folds.

Strength and Weakness of models:

LR is considered to be simple which has good interpretability making easy to understand the relationship between variables with an efficiency to handle large dataset with ease but is also important to have a balance between overfitting and underfitting while using LR model. Overfitting

can occur in LR when the features are highly correlated and this model is not suitable for non-linear data.

DTR is opposite to LR in reference with linearity and is easy to understand and interpret. The advantage of using this model, it can handle mixed data types by neither making use of scaling nor the one-hot encoding. As the name suggest it is a greedy algorithm and disadvantages is its instability that can cause changes in trees when a small change is made in data. Moreover, the model is biased when the data is unbalanced.

GBR is one of best methods to use for regression problems as it tackles well by providing high accuracy and predictive power due to its ensemble nature that combines multiple learners to form a strong learner. A flexible algorithm that handles both numerical and categorical data and is best suited for complex regression problems. Disadvantage of GBR is, needs more time and effort to fine-tune the hyperparameters and is expensive when compared with other models.

RFR is the model that tries to reduce the risk of overfitting and can estimated the importance of features for output and the easy use of handling missing data but is said to be slow to train datasets and requires higher memory when compared to DTR

Results:

The results of our investigation are summarized in Table 1 below. The table shows the RMSE, RMAE, R2 Score and Cross validation score for each of the four machine learning techniques.

ML Algorithms	RMAE	RMSE	R2 Score	Mean Square cross_val_score
Linear Regression	7.528	98.556	0.7446	7.9090
Decision Tree Regressor	9.339	146.283	0.4374	8.5995
Gradient Boosting Regressor	6.825	77.333	0.8427	7.3411
Random Forest Regressor	8.137	104.127	0.7149	8.4821

Table Figure 1.1

The results show that the gradient boosting regressor achieved the lowest scores for RMAE, RMSE, and mean Square CVS with values 6.825, 77.333, and 7.3411 respectively. The R2 Score 0.8427 which is closer to 1 also indicates that the model is the best fit. The linear regression also performed well, achieving a RMAE and RMSE of 7.528 and 98.556 with R2 score as 0.7446 but not reaching the expected values of CVS.

Comparison scores on graph for various ML Models:

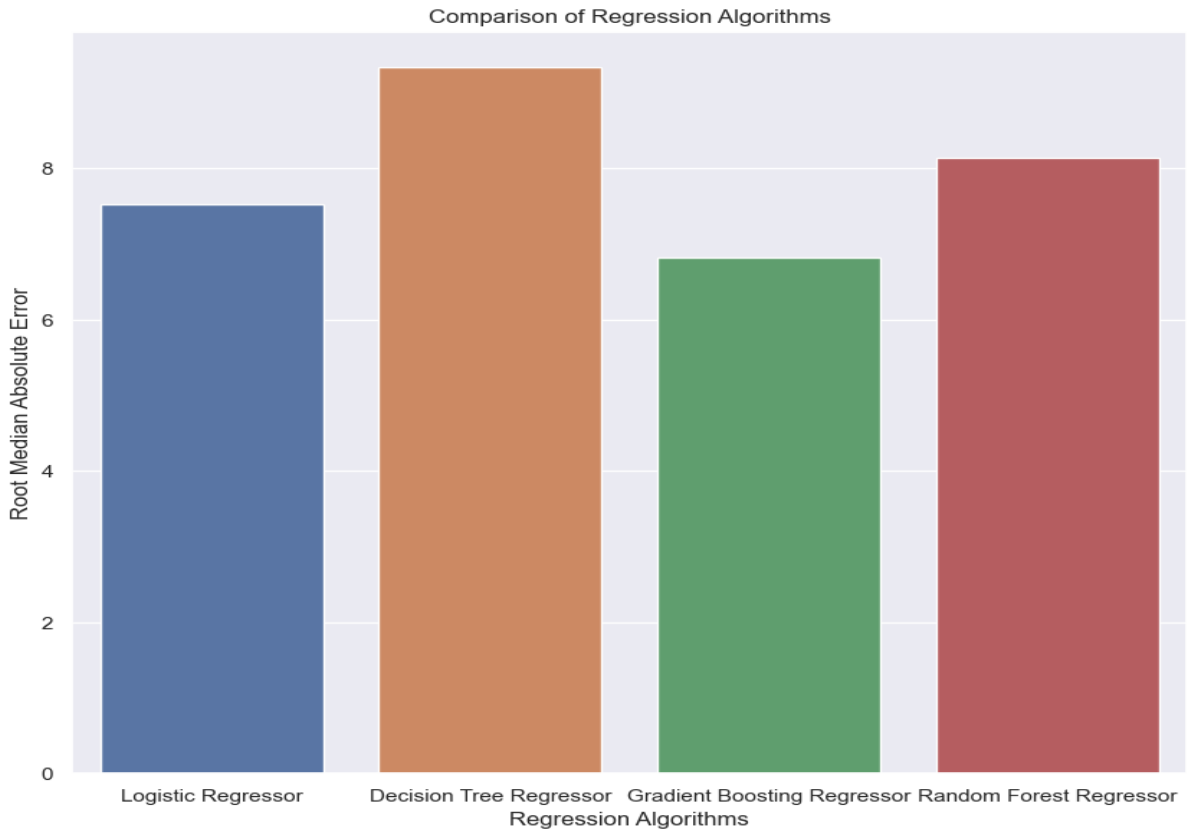

Fig 1.2 Root Median Absolute Error evaluation of models

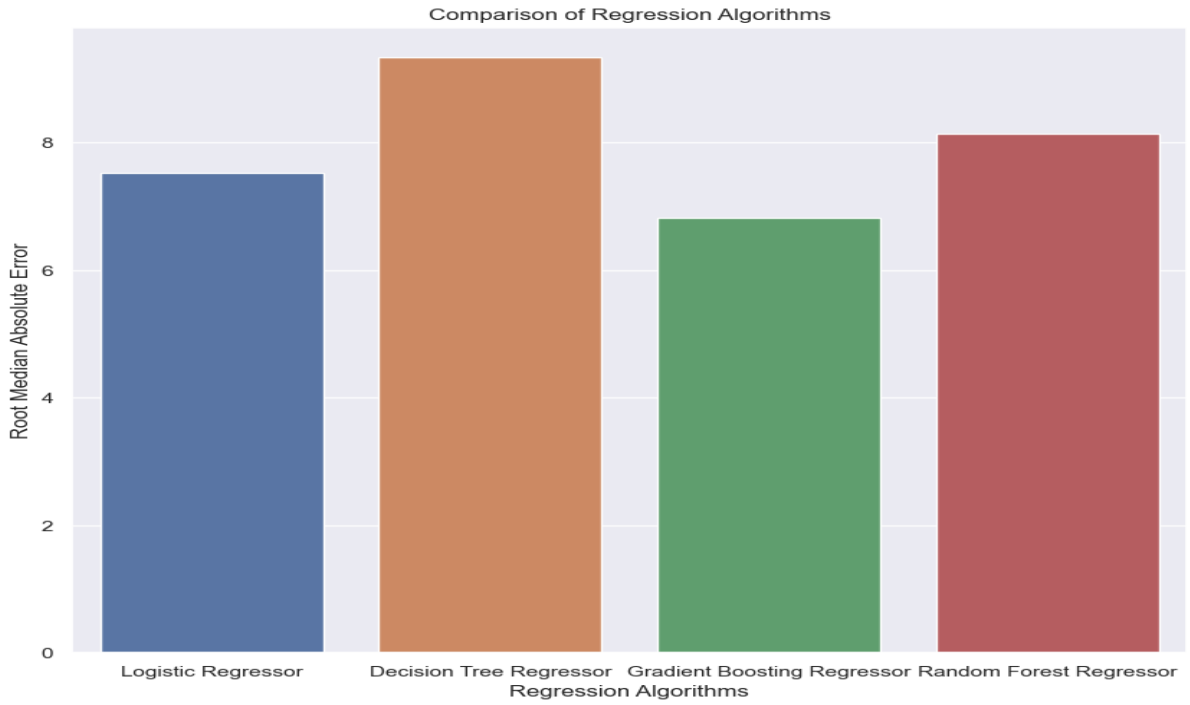

Fig 1.3 Root mean squared error evaluation of models

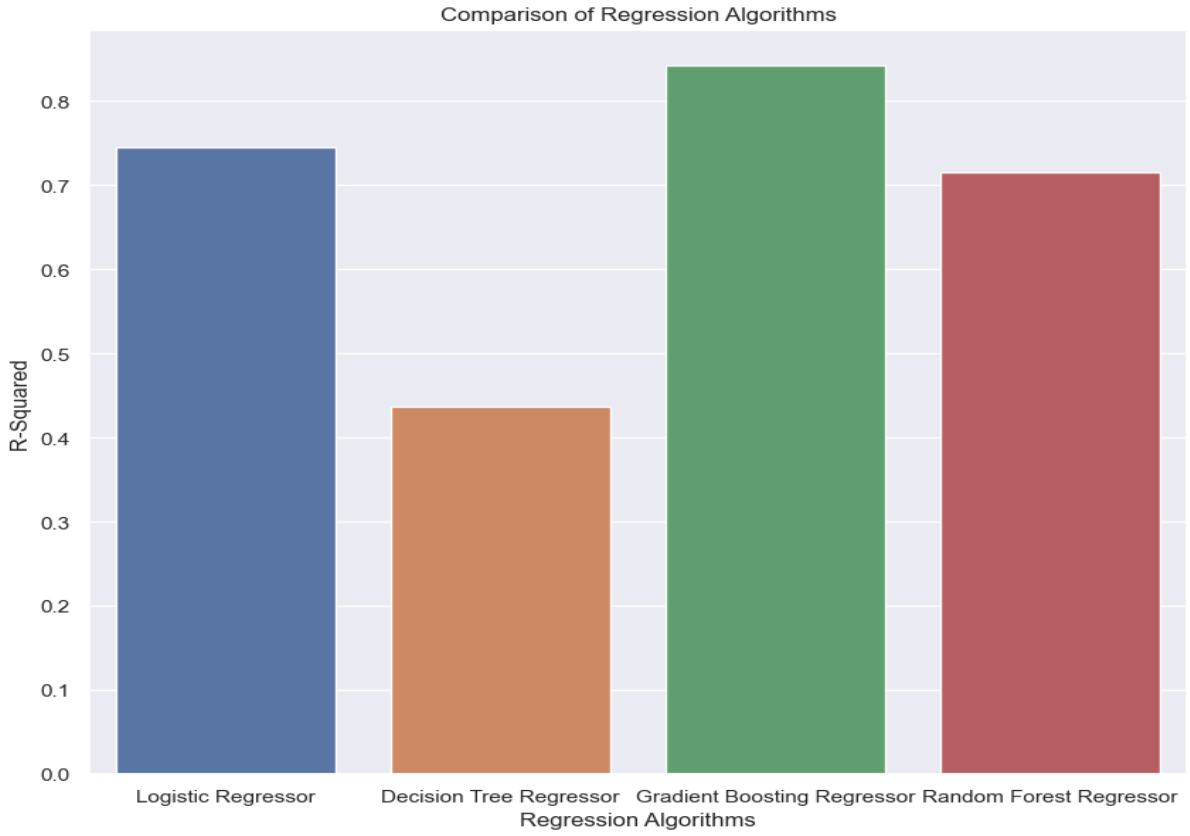

Fig 1.4 R2 Scores of all Models

Discussion:

The results suggest that the GBR and LR are the most suitable techniques. Both of these achieved RMSE and RMAE values that were considerably lower than the other techniques.

Storing all the used models into a single variable and predicting the best fit model for part B by passing on the MSE scores where the outcome with best fit model was GBR. Hence, a pipeline is used and MSE score was predicted and this pipeline is passed onto the test dataframe to predict the test data.

Conclusion:

The investigation suggests that GBR is the most suitable techniques for predicting extent by which a patient's blood glucose level exceeds the diagnostic threshold. This achieved the lowest RMSE and are is most accurate. The LR technique also performed less well but may not be suitable for this particular problem.