CSCI – 6674 Data Mining

Team Name – Miners

Your report should include the name of all team members, their UNH emails, your team name.

PRASANTH REDDY THUMMA - pthum2@unh.newhaven.edu

SUMA MODUGU - smodu1@unh.newhaven.edu

MOHAMMAD NAZEER - mshai14@unh.newhaven.edu

Please introduce your selected data set and research question.

Dataset:

https://www.kaggle.com/mohansacharya/graduate-admissions

- The dataset we are going to use in this prediction is extracted from Kaggle, where all major study and analysis of dataset's take place.
- We choose Kaggle as our source of dataset, snice it provides a real-time data.
- The collected data is used for only study analysis, and we make sure to follow all the guidelines and copyright issues by the dataset owner.

Research questions: Does the student get admission from the university? In addition, we included more questions to describe more about the data and making necessary data analysis.

List of Data Mining Techniques.

The following are the list of data mining techniques used:

1. **Logistic Regression:** The Logistic regression is typically used to classify low dimensional data with nonlinear boundaries. also, it provides the difference in the percentages of the dependent variables and the rank of each variable. The main purpose of Logistic Regression is to determine the correct result of each variable Logistic regression is also known as Logistic Model, which is a categorical variable with two categories, for instance light or dark, slim/healthy.
2. **Decision Tree Regression:** One of the most frequent methods for creating classifiers is to use a decision tree. It's like the flowchart structure, in which each internal node represents a condition on an attribute, each branch reflects the condition's conclusion, and each leaf node represents the class label. After computing all qualities, a decision is made. Classification rules are represented via a path from a root to a leaf.
   In the medical industry, decision trees are used to decide the order of qualities. It first generates a set of solved problems. The entire set is then separated into two parts: a training set and a testing set. Where a training set is used for the induction of a decision tree. The testing set is used to determine the accuracy of the system**.**

3. **Random Forest Regression:** It is a classification approach that is developed by training a multiple-choice tree and then voting individual branches to produce a class. Based on attribute positions chosen at random, the algorithm creates a forest of decision trees. It offers the advantage of raising forecast accuracy without increasing processing expenses much.
4. **Support Vector Regression:** SVM**:** Statistical learning models such as SVMs are becoming more popular. SVMs are supervised learning models that are applied mainly for classification, but they can also be used to solve regression problems. An SVM is a binary classifier that divides training data into two categories.
   The SVM algorithm maps features into a higher-dimensional vector space, where in this space, a maximum margin hyperplane is established. On each side, the distance between the hyperplane and the closest data point is maximized. The method of maximizing the margin, and thus producing the largest possible distance between the separating hyper-plane and the instances on either side of it, has proven to significantly reduce the expected generalization error.

Model Parameters and Hyperparameters.
Model Parameters: These are the parameters in the model that must be determined using the training dat a set. These are the fitted parameters.
Hyperparameters: These are adjustable parameters that must be tuned in order to obtain a model with op timal performance.

Logistic Regression: -
Parameters = [
{
C=0.23
solver= 'liblinear'
},
]

Decision Tree: -
Parameters = [
{
criterion= 'entropy'
min samples leaf= 13
min samples split= 2
max_depth= 3
splitter= 'random'
},

]

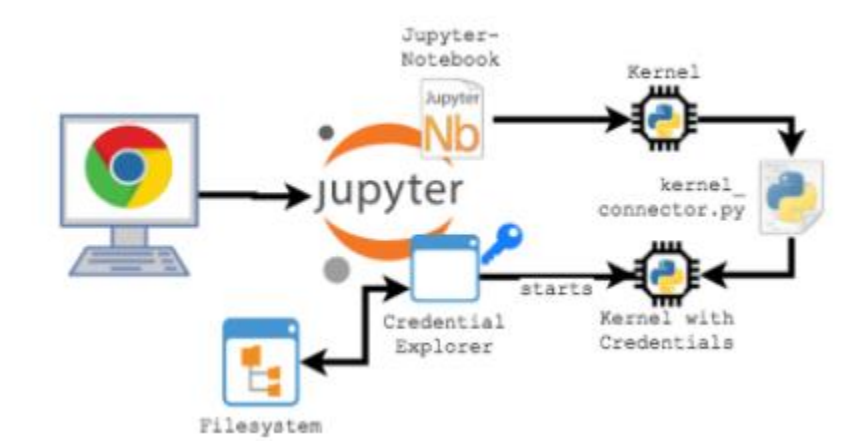
Support Vector Machine: -
Parameters = [
{
kernel= 'rbf'
gamma= 0.1, C= 1.0
},
]

Random Forest: -

Parameters = [
{
criterion= 'gini',
min samples leaf= 1
min samples split= 2
no of estimators= 200
max_features= 'sqrt'
max_depth= 50
},

]

Brief Description of Hardware Used:
Processor: i5
Hard disk: 512GB SSD
Memory: 8 GB
Operating System: Windows 11 64-bit
Language: Python3
Dataset Source: Kaggle
Tool: Jupyter Notebook



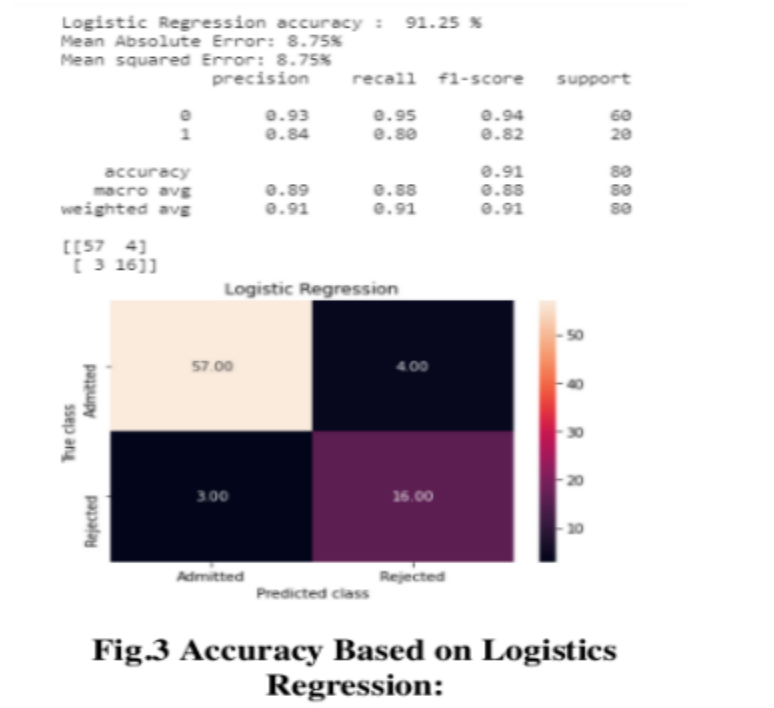


Fig.3 Accuracy Based on Logistics Regression:

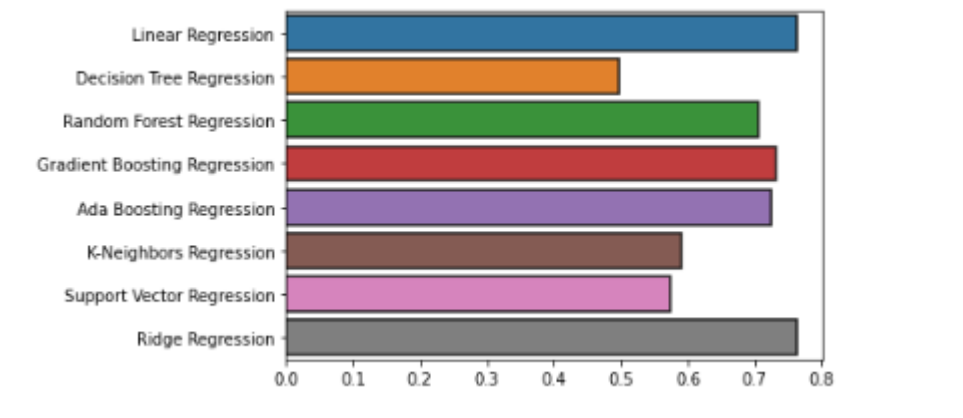
Outcomes of data mining techniques.

The following outcomes are obtained after fitting the testing data into the trained model:

Linear Regression: 0.7633591517711498
Decision Tree Regression: 0.4968653502015902
Random Forest Regression: 0.7064079522923155
Gradient Boosting Regression: 0.7325623738426896
Ada Boosting Regression: 0.7243490916447404
K-Neighbors Regression: 0.591216177044829
Support Vector Regression: 0.5747083714216271
Ridge Regression: 0.7624408186173257

From the obtained results it is clear that Linear Regression with 76.3%, so usage of this model produces more decent predictions followed by Ridge Regression with 76.2% and Gradient Boosting Regression with 73.2%.

The following is a bar chart which plot the obtained results:



The data modeling starts with dividing the data into training and testing data. Here, we take the test_size = 0.2 which gives 20% data to train the model and remaining 80% of data is used for testing the data. Fitting this trained model to the testing data with different modeling techniques gives different predictive analysis based on the internal structure of each algorithm. So, to conclude we can see with the use Linear regression the outcome is 76% predictive.

Github : https://github.com/prasanth50/Phase5