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:** Logistic regression is a technique for classifying low-dimensional data with nonlinear bounds. It also shows the percentage difference between the dependent variables and the rank of each variable. The fundamental goal of Logistic Regression is to figure out what the right answer is for each variable. The Logistic Model, also known as Logistic Regression, is a categorical variable having two categories, such as light or dark.
2. **Decision Tree Regression:** A decision tree is one of the most used ways for building classifiers. Each internal node represents a condition on an attribute, each branch reflects the condition's conclusion, and each leaf node represents the class label, similar to the flowchart structure. After weighing all of the factors, a conclusion is taken. A path from a root to a leaf is used to illustrate classification criteria.
   Decision trees are used in the medical business to determine the order of attributes. It starts by generating a list of problems that have been solved. After then, the entire set is divided into two sections: a training set and a testing set. When constructing a decision tree, a training set is used. The testing set is used to figure out how many people are in a group.

3. **Random Forest Regression:** It's a categorization method that involves creating a multiple-choice tree and then voting on specific branches to create a class. The technique generates a forest of decision trees based on attribute placements picked at random. It has the advantage of enhancing forecast accuracy without significantly increasing processing costs.
4. **Support Vector Regression:** SVM**:** SVMs, or statistical learning models, are becoming increasingly popular. SVMs are supervised learning models that can be used to address regression problems as well as classification difficulties. A binary classifier, such as an SVM, separates training data into two groups.
   The SVM technique converts characteristics into a higher-dimensional vector space, where a maximum margin hyperplane is found. The distance between the hyperplane and the closest data point is maximized on both sides. The strategy of increasing the margin, and therefore establishing the greatest feasible distance between the separating hyper-plane and the instances on either side of it, has been shown to lower the expected generalization error greatly.

5. **Gaussian Naïve:** Gaussian Naive Bayes accepts continuous valued features and models them all as Gaussian (normal) distributions. To build a simple model, assume the data is characterized by a Gau ssian distribution with no covariance (independent dimensions) between the parameters.KNN Alg orith: "K-Nearest Neighbour" is the abbreviation for "K-Nearest Neighbour." It's a machine learning algorithm that's supervised. Both classification and regression problem statements can be solved usin g the approach. The sign 'K' represents the number of nearest neighbors to a new unknown variable t hat must be predicted or categorised.
6. **KNN Algorith:** "K-Nearest Neighbour" is the abbreviation for "K-Nearest Neighbour." It's a machine learning algorithm that's supervised. Both classification and regression problem statements can be so lved using the approach. The sign 'K' represents the number of nearest neighbors to a new unknown variable that must be predicted or categorised.

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
},

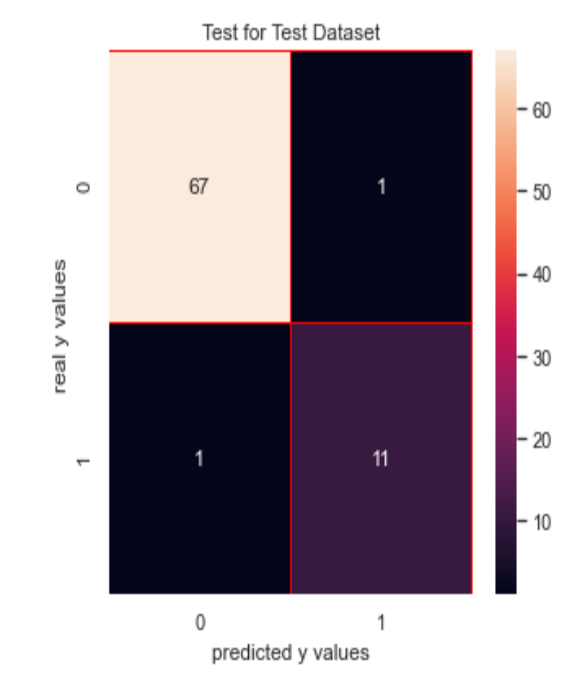
]

Outcomes of data mining techniques.

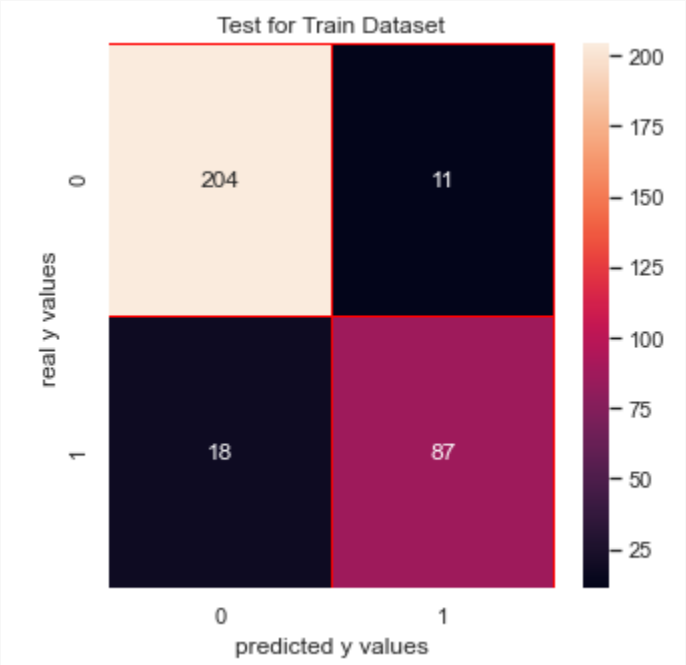
The following outcomes are obtained after fitting the testing data into the trained model:

Logistic Regression

For Testing Data:



For Trained data:

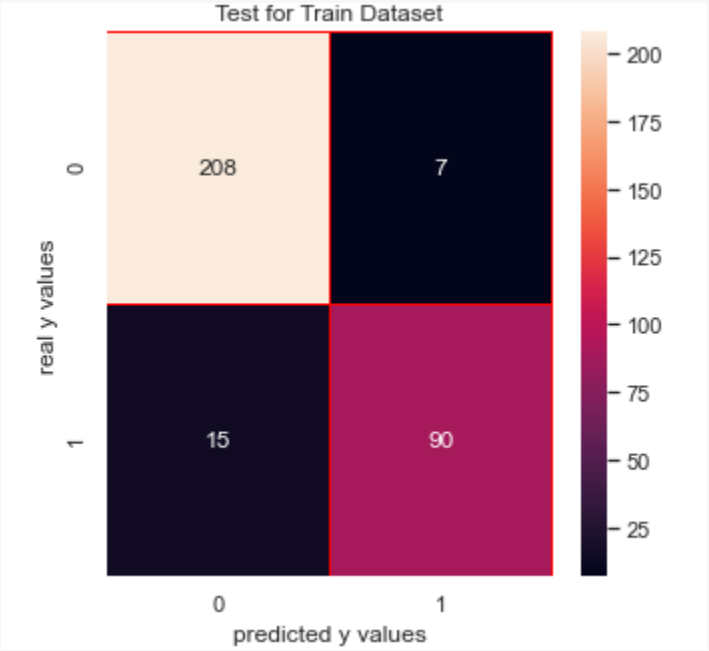



```
precision score:  0.9166666666666666
recall score:  0.9166666666666666
f1_score:  0.9166666666666666
```

Support Vector Machine

Training:

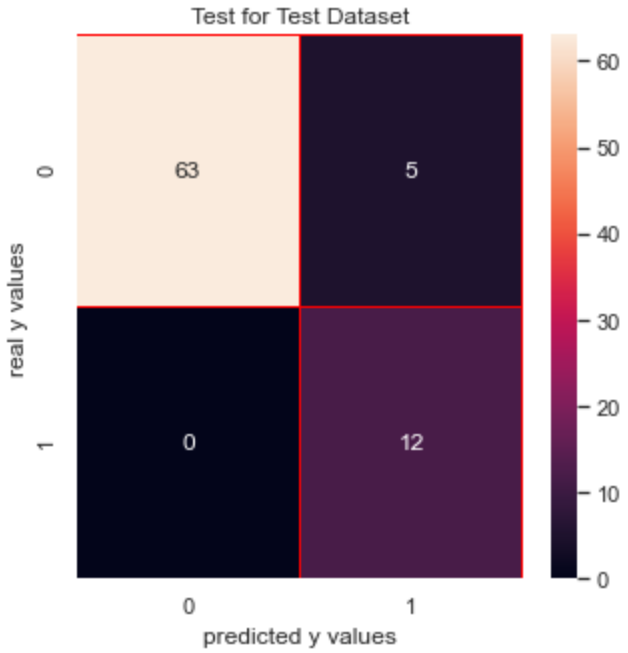
Testing:



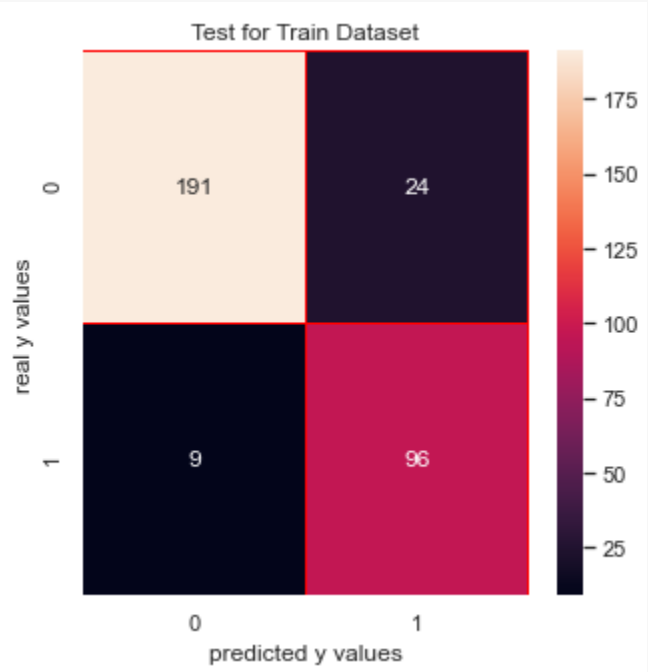

```
precision score:  0.7692307692307693
recall score:  0.8333333333333334
f1_score:  0.8
```

3. Gaussian Naive Bayes

Training:



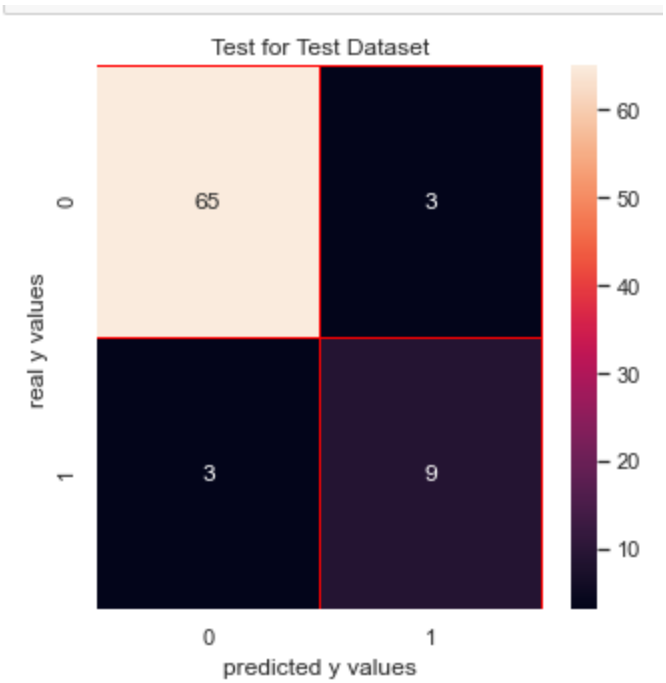
Testing:

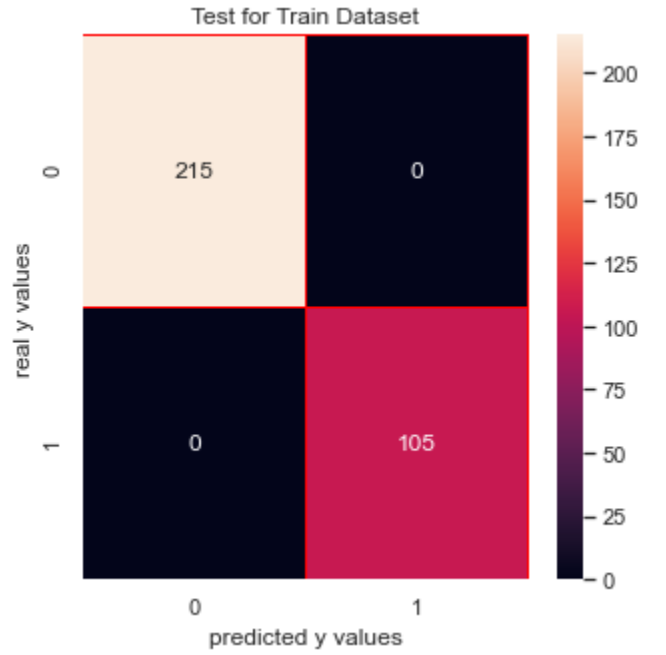



```
precision score:  0.7058823529411765
recall score:  1.0
f1_score:  0.8275862068965517
```

4. Decision Tree

Training:

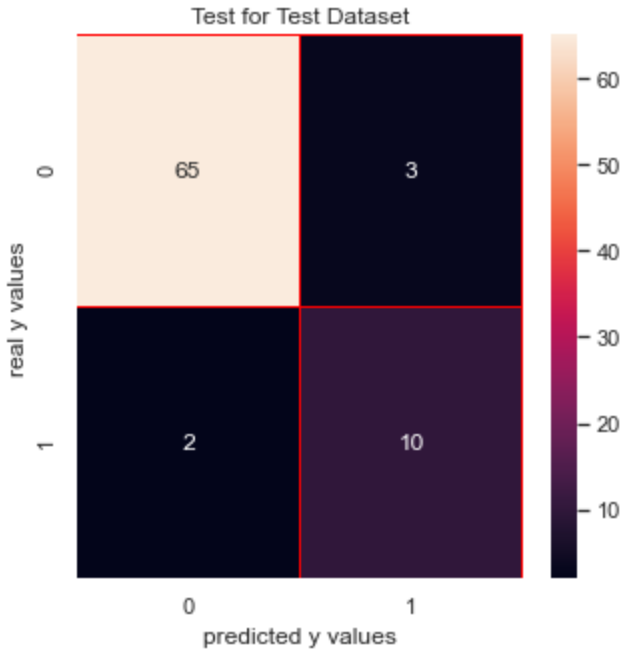




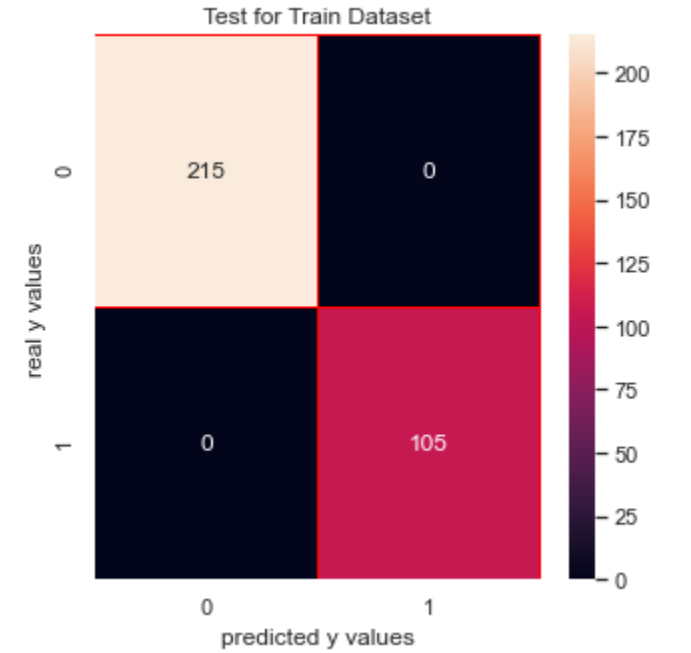

```
precision score:  0.75
recall score:  0.75
f1_score:  0.75
```

5. Random Forest

Training Data:

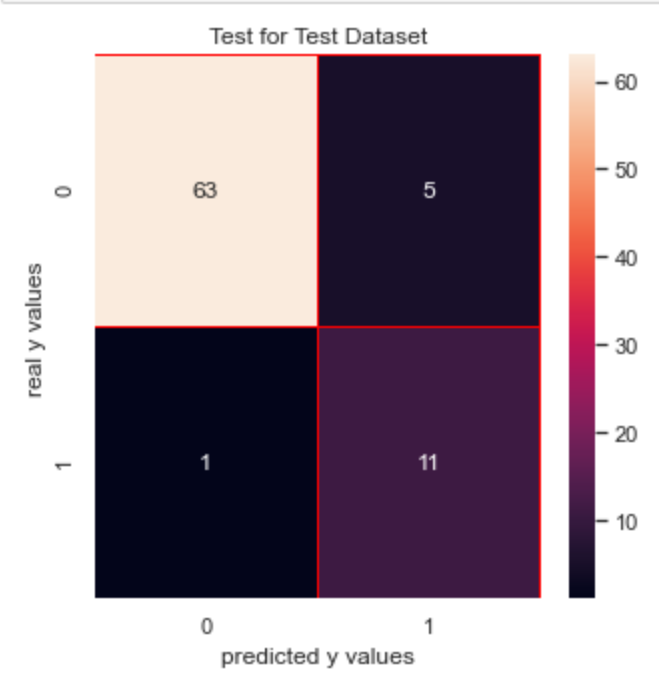


Testing:

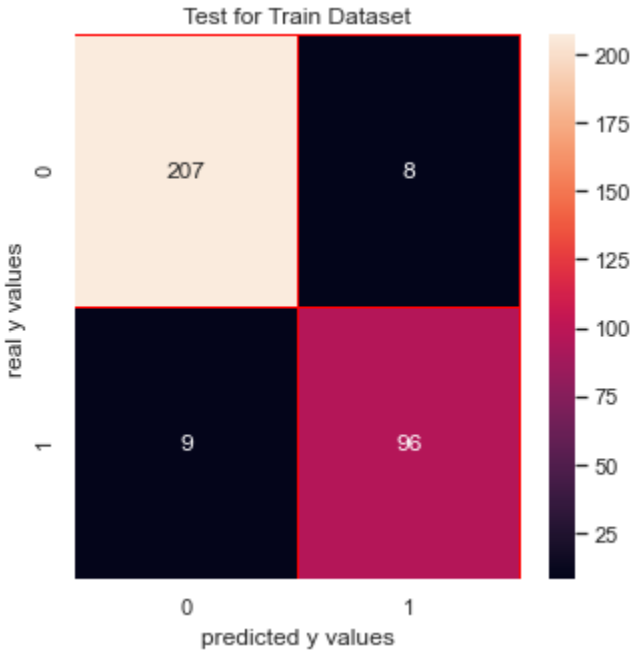



```
precision score:  0.7692307692307693
recall score:  0.8333333333333334
f1_score:  0.8
```

6.K Nearest Neighbors



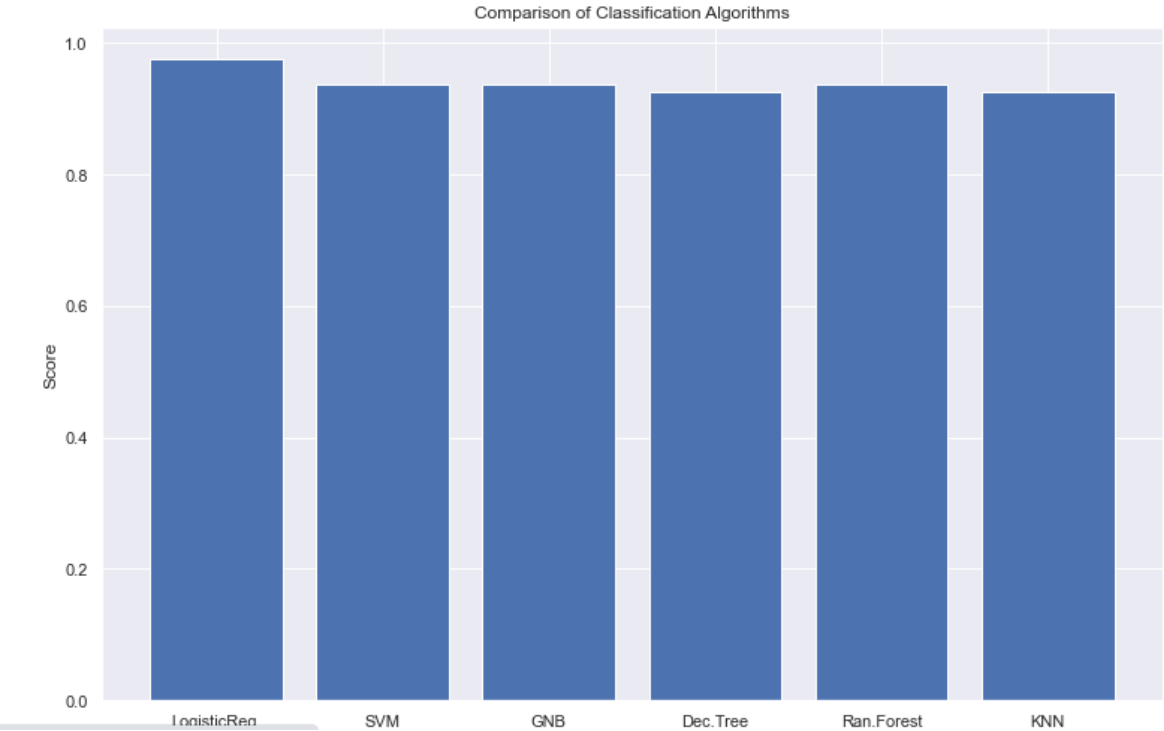
Testing:




```
precision score:  0.6875
recall score:  0.9166666666666666
f1_score:  0.7857142857142857
```

Comparison of Classification Algorithms:

The following Analysis are draw plotting all the algorithms output:



Conclusion:

Previously the data modeling starts with dividing the data into training and testing data. Here, we take t he 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 th e use Linear regression the outcome is 91% predictive.

Github : https://github.com/prasanth50/Phase6