CSE303: Statistics for Data Science [Summer 2021]

Project Report

Submitted by:

Student ID	Student Name	Contribution Percentage
2019-1-60-204	Noshin Faria	28
2018-2-60-008	Shamima Yesmin	24
2018-2-60-031	Fahmida Nusrat Promy	24
2019-1-60-264	Md. Shamsur Rahman Talukdar	24

1. Introduction

This project is based on student satisfaction dataset. Maximum features contain categorical data. Here, we use different kind of regression and classification algorithms to train the model to get the optimal solution. Parameters of different algorithm plays a vital role in some cases. Maximum models perform so well and some are not. Among them logistic regression's performance is optimal based on training score and prediction capability. On the other hand, polynomial regression performs worst which is logical for this linear dataset.

Dataset Characteristics and Exploratory Data Analysis

"Student Satisfaction Survey Dataset" is collected from EWU's students. This dataset contains 543 rows and 33 columns. It contains the personal data of 543 students and their satisfaction level for all kind of activities or actions taken by the university throughout these online semesters. The maximum features of this dataset contain categorical data.

The correlation among the features is -

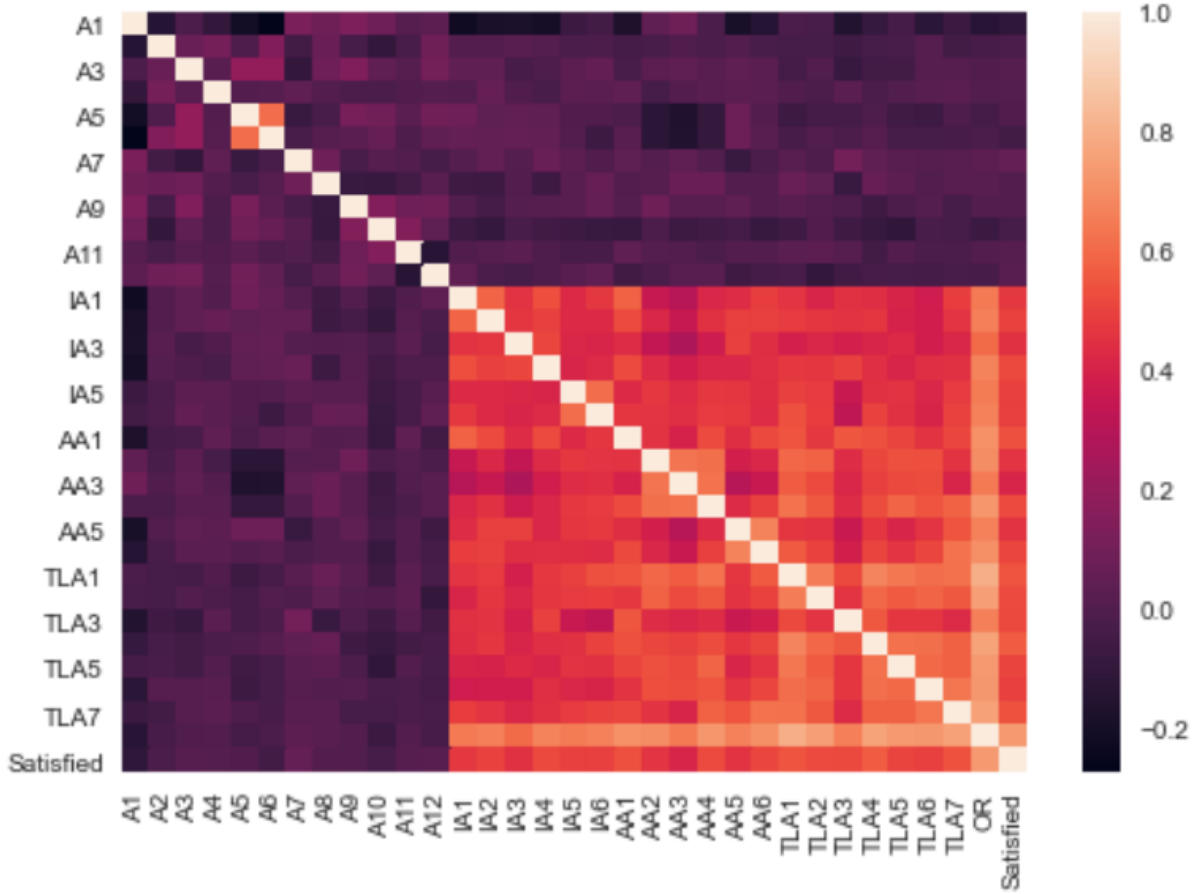


Figure: Correlation matrix

The above matrix represents that the features from "A1 to A12" are least correlated with all the features. From "IA1 to Satisfied" are highly correlated with each other. Among them, the correlation between "OR" and "IA1 to Satisfied" is highest. So, we can assume that "IA1 to Satisfied" will have some positive impact on our models.

```
In [47]: runcell('no 5', 'C:/Users/nus34/
Documents/cse303/project.py')
A1     -0.118190
A2     -0.024485
A3      0.009541
A4      0.004186
A5     -0.010237
A6     -0.049142
A7      0.054039
A8      0.000900
A9     -0.004867
A10    -0.037041
A11     0.011517
A12     0.017467
dtype: float64
```

Figure- Correlation among A1- A12 with Satisfied column

It shows the correlation of A1-A12 with the satisfied column. Minimum correlation exists among them. Some of their correlation even in negative direction. Among them, A7 column's correlation with "Satisfied" is 0.05. this column is the most correlated with satisfied column. A7 represents "Currently getting any Scholarship from the University"

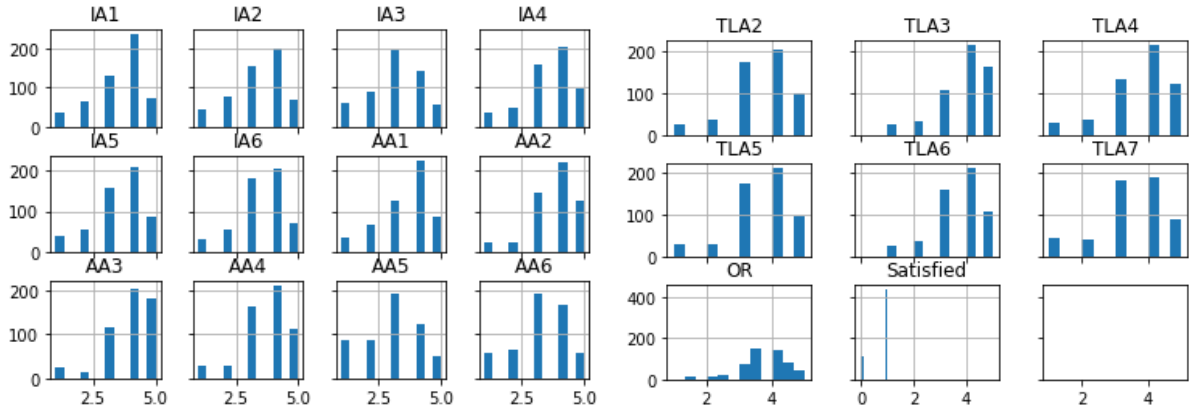


Figure: Histogram of data distribution

1. Machine Learning Models

OLS (ordinary least squares) model:

This is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. It is a very popular and widely used model it can gives overall overview of a model. It provides description of the following terms,

R-squared: It means the proportion of the variance in the dependent variable that is predictable/explained

Adj. R-squared: Adjusted R-squared is the modified form of R-squared adjusted for the number of independent variables in the model. Value of adj. R-squared increases, when we include extra variables which actually improve the model.

F-statistic: it determines ratio of mean squared error of the model to the mean squared error of residuals. It determines the overall significance of the model.

t: the value of t-statistic. It is the ratio of the difference between the estimated and hypothesized value of a parameter, to the standard error.

Linear regression model:

linear regression is very useful for predictive analysis. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

Polynomial Regression:

Polynomial regression is a supervised machine learning algorithm. It uses for nonlinear data where straight line can’t fit properly.

The polynomial models can be utilized in those circumstances where the connection among study and explanatory factors is curvilinear. It a special case of Multiple linear regression which estimates the relationship as an (nth) degree polynomial. We see between the actual value and the best fit line, which we anticipated and it appears to be that the actual value has some sort of curve in the diagram and our line is no where near to cutting the mean of the point.

Lasso Regression (L1):

Overfitting is a very common issue in the field of machine learning. L1 and L2 are the techniques that can be used to address the overfitting issue. Overfit means that with respect to the training dataset we get low error but with respect to the testing dataset we get high error. And underfit means getting high error with respect to the both training and testing dataset. A good model should always have low variance and low bias. By Using l1 and l2 we can convert high variance to the low variance. So basically, L1 regularization works by adding a term to the error function used by the training algorithm.

Ridge Regression (L2):

L2 regularization forces weights toward zero but it does not make them exactly zero. Using l2 we find the line that result in the minimum sum of squared residuals. The main idea behind l2 regression is to find a new line that does not fit the training data.

Logistic regression model:

A common job of machine learning algorithms is to recognize objects and being able to separate them into categories this type of algorithms are classifiers and logistic regression is one of them.

Logistic regression analysis is used to examine the association of (categorical or continuous) independent variables.

It estimates the effects of independent variables on the result variables as probability. The logistic regression ensures the determination of the risk factors as probability is a method that investigates the relationship of the result variables with independent variables in binary or multiple phases.

Support Vector Machine

SVM method applies a kernel function to perform classification. It performs well with large number of samples. SVM has additional parameters such as penalty normalization which applies L1 and L2.

K-Nearest Neighbours using Backward Feature Elimination

Sequential Feature Selection: Sequential Feature Selector can be forward feature selection or backward feature elimination. For backward feature elimination, it starts with all the features and remove one by one based on their cross-validation scores.

K-Nearest Classifier: The k-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm which used to solve both classification and regression problems. It uses n number of neighbors data points to predict its dependent feature. k-nearest neighbors (KNN) calculate the Euclidean distance from test value to others value and chooses the n number of closest neighbor's value to get the predicted result.

Data Preprocessing

Null Value: This dataset doesn't content any null value.

```
In [51]: runcell(1, 'C:/Users/Noshin/
A1           0
A2           0
A3           0
A4           0
A5           0
A6           0
A7           0
A8           0
A9           0
A10          0
A11          0
A12          0
IA1          0
IA2          0
IA3          0
IA4          0
IA5          0
IA6          0
AA1          0
AA2          0
AA3          0
AA4          0
AA5          0
AA6          0
TLA1         0
TLA2         0
TLA3         0
TLA4         0
TLA5         0
TLA6         0
TLA7         0
OR           0
Satisfied    0
dtype: int64
```

Encoding: We use Label Encoding technique here to encode labels (which contains string values) with a value between 0 to n-1 where n is the number of distinct labels.

In this dataset, (A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, Satisfied) these columns contain string labels. So, we use label encoding technique on these columns and convert all the labels into numeric value.

```
columns = ['A1','A2','A3','A4','A5','A6','A7','A8','A9','A10','A11','A12','Satisfied']
le = LabelEncoder()
df[columns] = df[columns].apply(le.fit_transform)
print(df.head())
```

Different Models

Discuss the different models that you implemented in this project with their parameters. Provide a detailed description of their parameters. Use tables as necessary.

5.1 Regression Models

OLS Model:

Here we have taken two variables X, Y. and imported a library,
import statsmodels.api as sm
X =
df[['IA1','IA2','IA3','IA4','IA5','IA6','AA1','AA2','AA3','AA4','AA5','AA6','TLA1','TLA2','TLA3','TLA4','TLA5','TLA6','TLA7','Satisfied']]
Y = df[['OR']]
Then we divided them into two parts testing part and training part. Then we have applied sm.ols() function on both training and testing part. Then we fit them using model.fit() function.
Then print the summary.

Linear regression:

In here, we exactly followed ols model for variables X, Y and we imported necessarily needed library.
from sklearn.linear_model import LinearRegression
Then we divided them into two parts testing part and training part. Then created an object called lr from LinearRegression(). Then we fit the training and testing both parts using lr.fit() function. Then we predict using lr.predict() function.
Then we calculate mean absolute error, mean squared error, r square. The result is:

Polynomial Regression

polynomial regression is a special case of multiple linear regression.
Parameters: default degree = 2
Degree: it represents the degree of the polynomial features. If an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features represent [1, a, b, a^2, ab, b^2].

Degree = 1 is equal to linear regression. If we apply degree = 3 or above in our dataset then the model's performance is 1. So, degree =1 makes the dataset overfitted and we get the worst test performance (<0).

Lasso Regression

Lesso is a regularization technique to panelize a model.
Parameter: alpha=0.01
Alpha: Constant that multiplies the L1 term.in our dataset, without alpha train r2 = 0; test r2 = -0.06

Ridge Regression

Redge is a regularization technique to panelize a model.
Parameter: alpha=0.01
Alpha: Constant that multiplies the L1 term.in our dataset, without alpha train r2 = 0; test r2 = -0.06

Own Implementation

```
In [60]: runcell(4, 'C:/Users/Noshin/OneDrive/Desktop
own method : The model performance for train data :
R2 score is 0.9601
MSE is 0.0230
MAE is 0.1316
```

5.2 Classification Models

Logistic regression:

We used GridSearchCV here to get the optimal parameter here.

Parameters: C=11.288378916846883, max_iter=180, penalty='l1', solver='liblinear'

C: it inverses the regularization strength, The value of c must be positive.

max_iter: it represents the maximum number of iterations

Penalty: it chooses different regularization technique to give penalty.

Solver: solver is used to optimize the problem.

Support Vector Machine

Support vector machine is a supervised algorithm.

Parameter: kernel='linear', C=9

Kernel: it represents the kernel type used in the algorithm. As our dataset is linear so linear kernel fit properly.

C: it inverses the regularization strength, The value of c must be positive.

K-Nearest Classifier using Backward Feature Elimination

Sequential Feature Selection:

For backward feature elimination, we use Sequential Feature Selection technique.

Parameters: k_features=16, forward=False, verbose=1, n_jobs=-1

k_features: it contains n number of best features to add or remove based on the cross-validation score of estimators. Here 16 is chosen after comparing 14 to 30's result.

Forward: it represents that the Sequential Feature Selection technique is going to follow forward feature selection or backward feature elimination. As forward is false so it will follow backward feature elimination method.

Verbose: verbose 1 is used to visualize the computation time for each parameter and their status.

n_jobs :This parameter is used to specify how many concurrent processes or threads should be used. N_jobs = -1 represent that all CPUs are going to be used. As this technique takes some times to compute so -1 is the best option to make it a little bit firster.

K-Nearest Classifier:

K- nearest neighbors is a regression and classification algorithm. We used it for classification purposes.

Parameters: n_neighbors = 10, p=1, weights='uniform'

n_neighbors: the number of nearest neighbors which are going to be used to get the predicted value. Euclidean distance between test data and chosen nearest neighbor's data is the lowest. Here, 10 is chosen after testing so multiple numbers.

P: P=1 contains power parameter for the Minkowski metric which actually equivalent to the Manhattan distance (l1). The formula of Minkowski distance is - sum (|x - y|^p) ^(1/p).

Weights: uniform is chosen by default. It means all the points in each neighborhood are weighted equally. If we use 'distance' (weight points will be inverse of their distance) than the r2 value is – 1.00. But in terms of testing, r2 value is 92. So, we can say that the model gets overfitted.

Performance Evaluation

Regression Models

Write the performance evaluation of Regression Models here. **(Task 1)**

Linear Regression

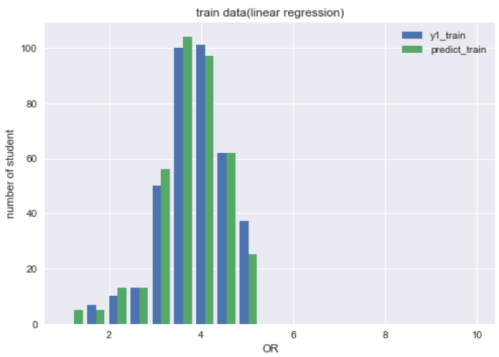


Figure-1.1

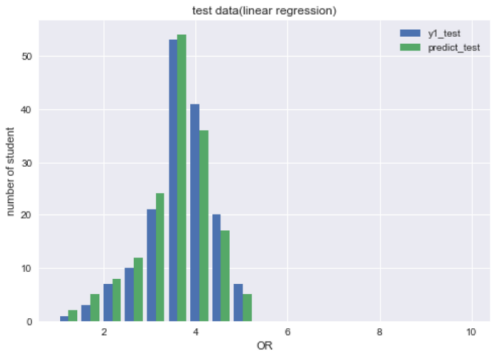


figure-1.2

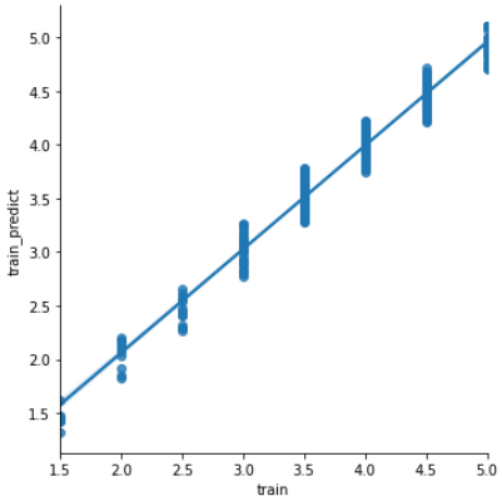


Figure- 1.3

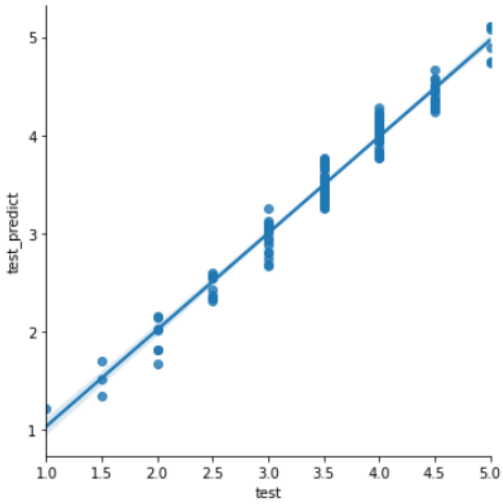


figure- 1.4

From figure 1.1 & 1.3, we can visualize that the model's performance is good. From figure 1.2& 1.4, we also can see that the prediction of test dataset is also good.

The value of R-square, MSE, MAE for training and testing dataset are -

```
In [65]: runcell(5, 'C:/Users/Noshin/OneDrive/Desktop/faric
LinearRegression : The model performance for training set
--------------------------------------
train MAE = 0.1214
train MSE = 0.0200
train R2 score = 0.9654
LinearRegression : The model performance for testing set
--------------------------------------
test MAE = 0.1260
test MSE = 0.0224
test R2 score = 0.9614
```

Visualizing the graphs and scores, we can say that this model has low bias and low variance.

Lasso Regression

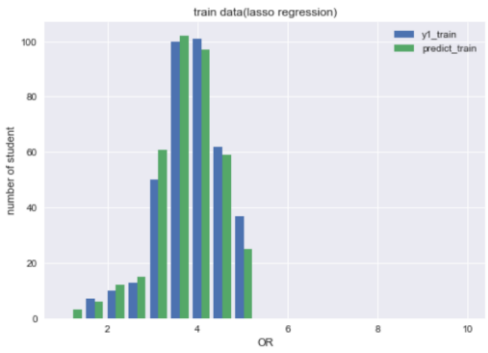


Figure-2.1

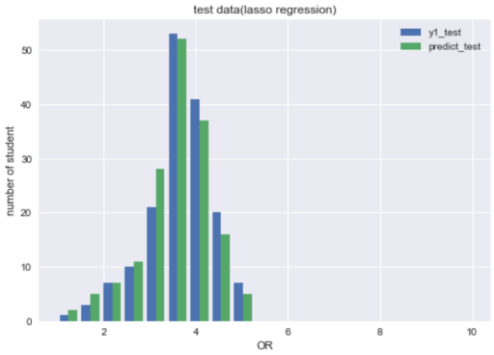


figure-2.2

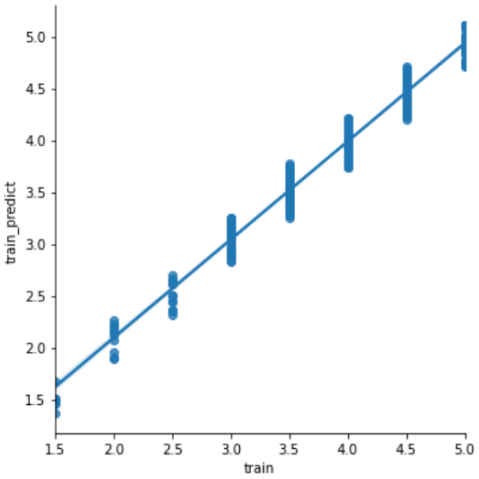


Figure- 2.3

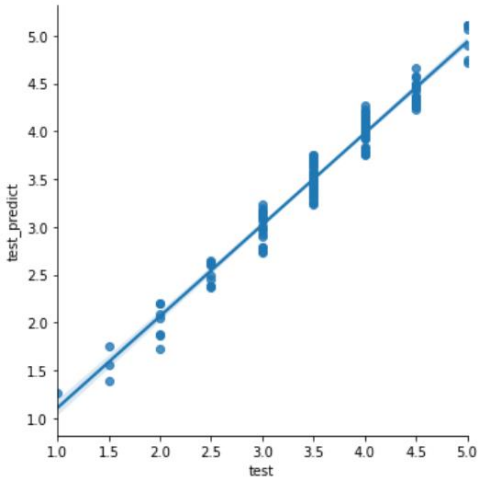


figure- 2.4

From figure 2.1 & 2.3, we can visualize that the model's performance is good. From figure 2.2& 2.4, we also can see that the prediction of test dataset is also good.

The value of R-square, MSE, MAE for training and testing dataset are -

```
In [70]: runcell(6, 'C:/Users/Noshin/OneDrive/Desktop/faric
lasso regression : The model performance for training set
--------------------------------------
train MAE = 0.1251
train MSE = 0.0208
train R2 score = 0.9641
lasso regression : The model performance for testing set
--------------------------------------
test MAE = 0.1291
test MSE = 0.0229
test R2 score = 0.9605
```

Visualizing the graphs and scores, we can say that this model has low bias and low variance.

Ridge Regression

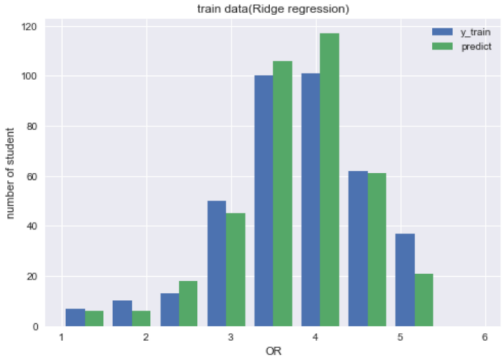


Figure-3.1

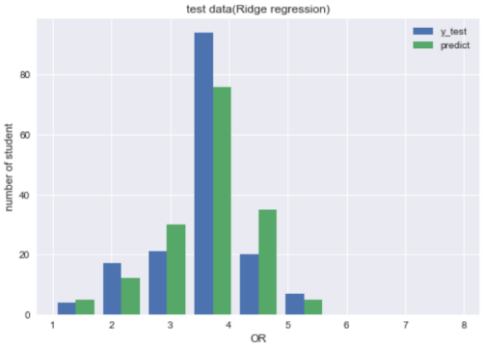


figure-3.2

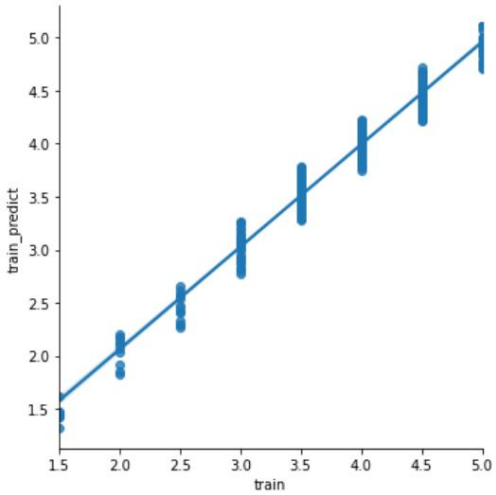


Figure- 3.3

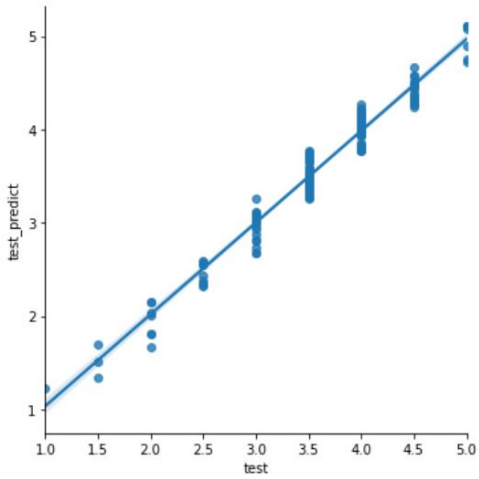


figure- 3.4

From figure 3.1 & 3.3, we can visualize that the model's performance is good. From figure 3.2& 3.4, we also can see that the prediction of test dataset is also good.

The value of R-square, MSE, MAE for training and testing dataset are -

```
In [71]: runcell(7, 'C:/Users/Noshin/OneDrive/Desktop/fari
Ridge regression : The model performance for training set
--------------------------------------
train MAE = 0.1215
train MSE = 0.0200
train R2 score = 0.9654
Ridge regression : The model performance for testing set
--------------------------------------
test MAE = 0.1260
test MSE = 0.0224
test R2 score = 0.9614
```

Visualizing the graphs and scores, we can say that this model has low bias and low variance.

Polynomial Regression

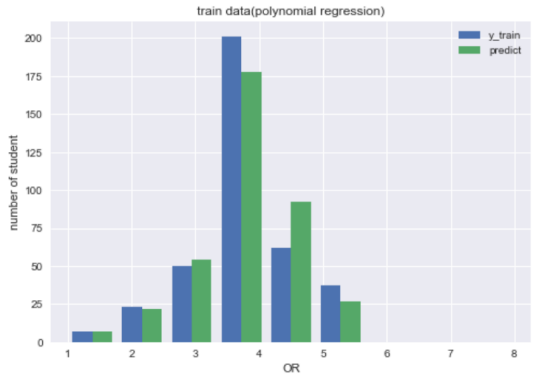


Figure-4.1

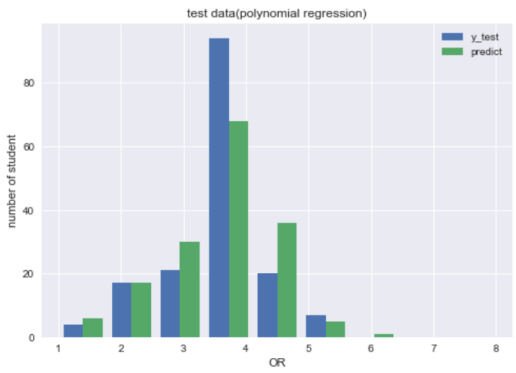


figure-4.2

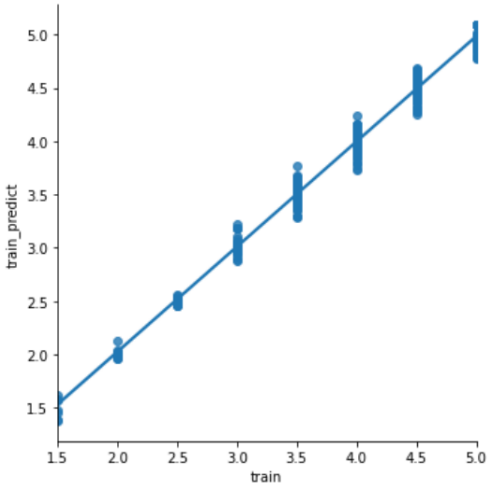


Figure- 4.3

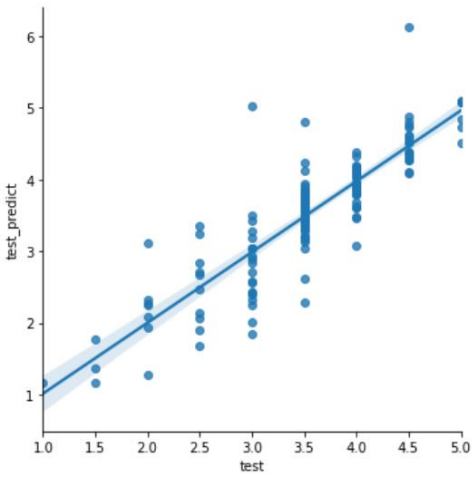


figure- 4.4

From figure 4.1 & 4.3, we can visualize that the model's performance is good. From figure 4.2& 4.4, we can see that the prediction of test dataset is worst.

The value of R-square, MSE, MAE for training and testing dataset are -

```
In [74]: runcell(8, 'C:/Users/Noshin/OneDrive/Desktop/faria/cse30
polynomial regression : The model performance for training set
--------------------------------------
train MAE = 0.0695
train MSE = 0.0083
train R2 score = 0.9857
polynomial regression : The model performance for testing set
--------------------------------------
test MAE = 0.2874
test MSE = 0.1749
test R2 score = 0.6982
```

Visualizing the graphs and scores, we can say that this model has low bias and high variance. It's an overfitted model.

Classification Models

Write the performance evaluation of Classification Models here. **(Task 2)**

Logistic Regression

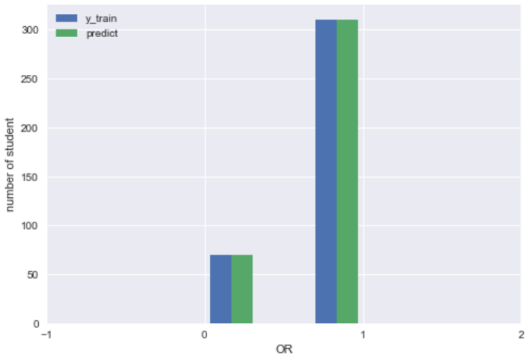


Figure-5.1

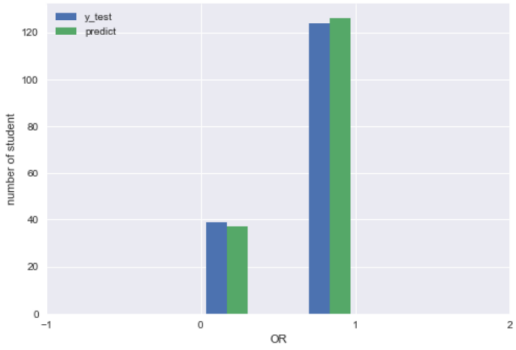
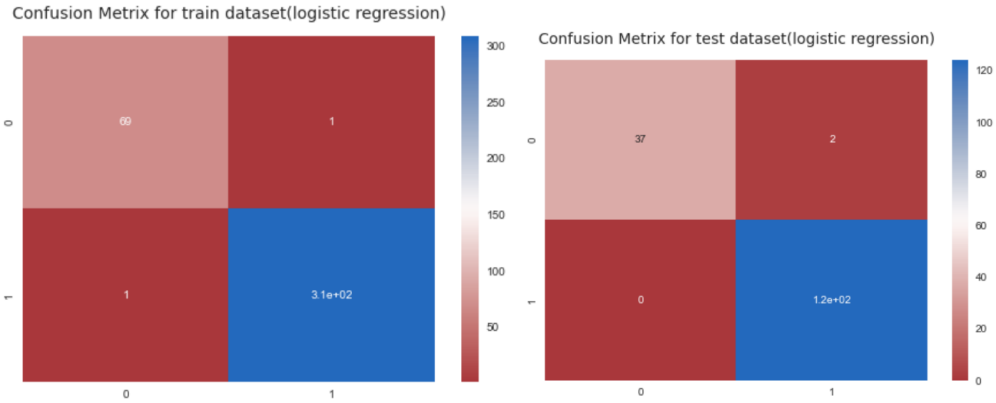


figure-5.2

From figure 5.1, we can visualize that the model's performance is so good. This model has a high change to be overfitted. But from figure 5.2, we can see that the prediction of test dataset is also good.



From the confusion matrix we can say that the accuracy of training data and testing data is almost same.

```
In [87]: runcell(12, 'C:/Users/Noshin/OneDrive/Desktop/fo
Logistic regression :
train accuracy = 0.9947
test accuracy = 0.9877
confusion matrix of train data :
[[ 69   1]
 [  1 309]]
confusion matrix of test data :
[[ 37   2]
 [  0 124]]

In [88]: runcell(14, 'C:/Users/Noshin/OneDrive/Desktop/fo
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        39
           1       0.98      1.00      0.99       124

    accuracy                           0.99       163
   macro avg       0.99      0.97      0.98       163
weighted avg       0.99      0.99      0.99       163
```

Visualizing the graphs and scores, we can say that this model has lowest bias and lowest variance.

Support vector machine

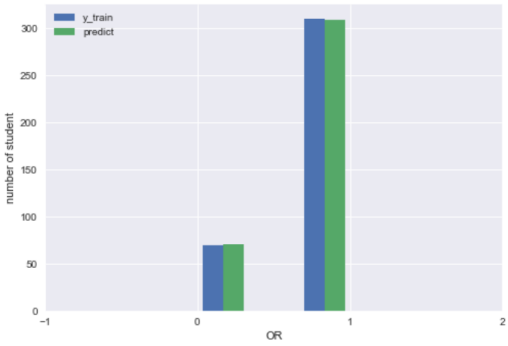


Figure-6.1

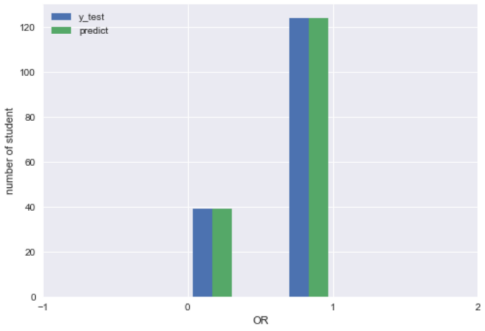
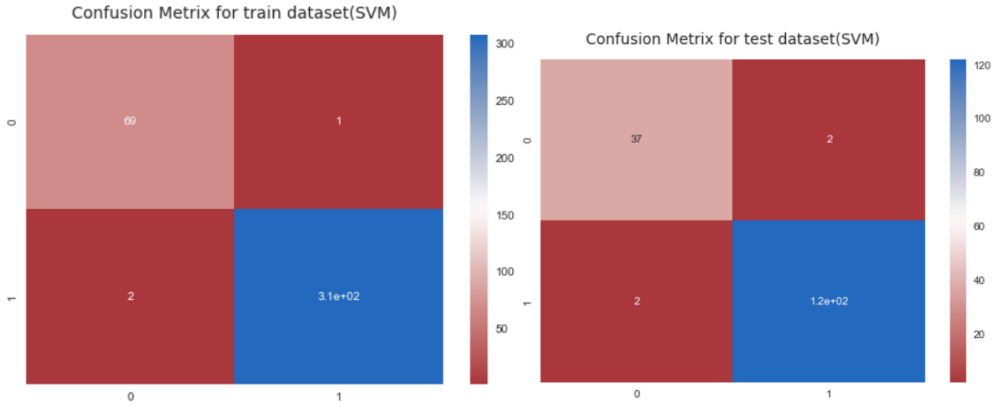


figure-6.2

From figure 6.1, we can visualize that the model's performance is so good. This model has a high change to be overfitted. But from figure 6.2, we can see that the prediction of test dataset is also good.



From the confusion matrix we can say that the accuracy of training data and testing data is almost same.

```
In [97]: runcell(17, 'C:/Users/Noshin/OneDrive/Desktop/f
svm = train accuracy = 0.9921
svm = test accuracy = 0.9755
svm = confusion matrix of train data :
[[ 69   1]
 [  2 308]]
svm = confusion matrix of test data :
[[ 37   2]
 [  2 122]]

In [98]: runcell(19, 'C:/Users/Noshin/OneDrive/Desktop/f
              precision    recall  f1-score   support

           0       0.95      0.95      0.95        39
           1       0.98      0.98      0.98       124

    accuracy                           0.98       163
   macro avg       0.97      0.97      0.97       163
weighted avg       0.98      0.98      0.98       163
```

Visualizing the graphs and scores, we can say that this model has lowest bias and lowest variance.

K-Nearest Classifier using Backward Feature Elimination

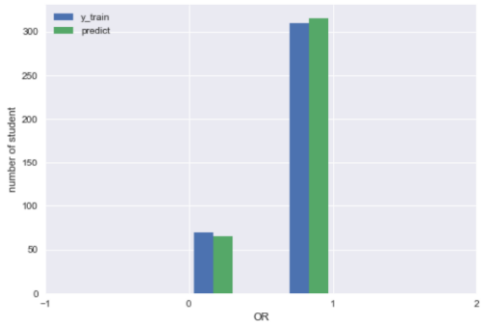


Figure-7.1

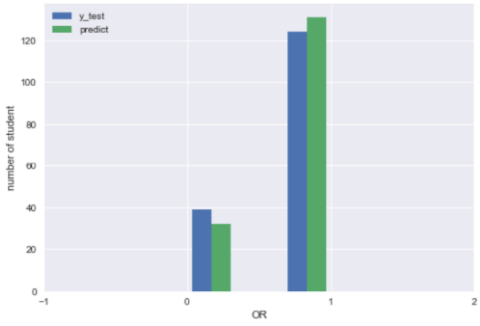
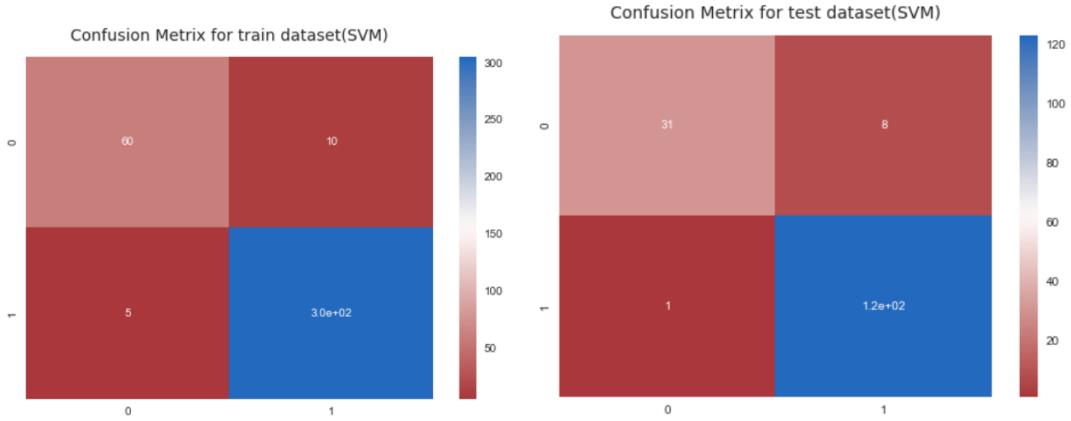


figure-7.2

From figure 7.1, we can visualize that the model's performance is good. From figure 7.2, we can see that the prediction of test dataset is also good.

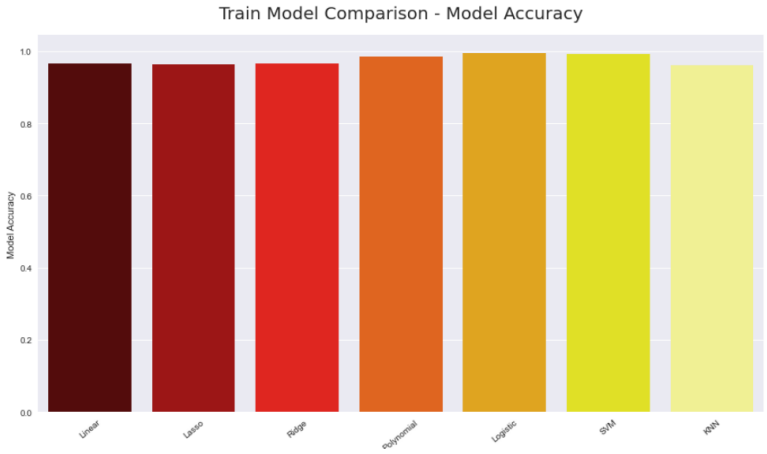


From the confusion matrix we can say that the accuracy of training data and testing data is good.

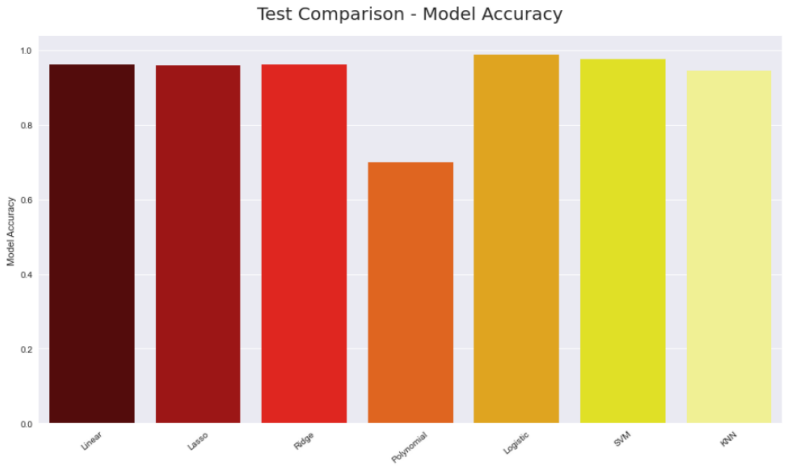
```
In [100]: runcell(21, 'C:/Users/Noshin/OneDrive/Desktop/far
Backward Feature Elimination K-Nearest Neighbours :
train accuracy = 0.9605
test accuracy = 0.9448
confusion matrix of train data :
[[ 60  10]
 [  5 305]]
confusion matrix of test data :
[[ 31   8]
 [  1 123]]

In [101]: runcell(22, 'C:/Users/Noshin/OneDrive/Desktop/far
              precision    recall  f1-score   support

           0       0.97      0.79      0.87        39
           1       0.94      0.99      0.96       124

    accuracy                           0.94       163
   macro avg       0.95      0.89      0.92       163
weighted avg       0.95      0.94      0.94       163
```

Visualizing the graphs and scores, we can say that this model has low bias and low variance.

Discussion



From this “Train model comparison” we can see, all the model's performance is good. Among them, “Polynomial”, “logistic” and “SVM” gives the best score.



“Polynomial”, “logistic” and “SVM” gives the best score in terms of training dataset. But the score of polynomial regression is worst in terms of testing dataset. So, “logistic” and “SVM” are the best model for both training and testing dataset.

All the scores of these two models -

```
In [87]: runcell(12, 'C:/Users/Noshin/OneDrive/Desktop/fo
Logistic regression :
train accuracy = 0.9947
test accuracy = 0.9877
confusion matrix of train data :
[[ 69   1]
 [  1 309]]
confusion matrix of test data :
[[ 37   2]
 [  0 124]]

In [88]: runcell(14, 'C:/Users/Noshin/OneDrive/Desktop/fo
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        39
           1       0.98      1.00      0.99       124

    accuracy                           0.99       163
   macro avg       0.99      0.97      0.98       163
weighted avg       0.99      0.99      0.99       163
```

Logistic regression

```
In [97]: runcell(17, 'C:/Users/Noshin/OneDrive/Desktop/f
svm = train accuracy = 0.9921
svm = test accuracy = 0.9755
svm = confusion matrix of train data :
[[ 69   1]
 [  2 308]]
svm = confusion matrix of test data :
[[ 37   2]
 [  2 122]]

In [98]: runcell(19, 'C:/Users/Noshin/OneDrive/Desktop/f
              precision    recall  f1-score   support

           0       0.95      0.95      0.95        39
           1       0.98      0.98      0.98       124

    accuracy                           0.98       163
   macro avg       0.97      0.97      0.97       163
weighted avg       0.98      0.98      0.98       163
```

SVM

Comparing all the scores, we see **“Logistic Regression”** is the best model for this dataset.

Hypothesis

- "If a student is satisfied through the teacher's interactive behavior, supportive teaching-learning, and appropriate assessments then that student will satisfy on the university."

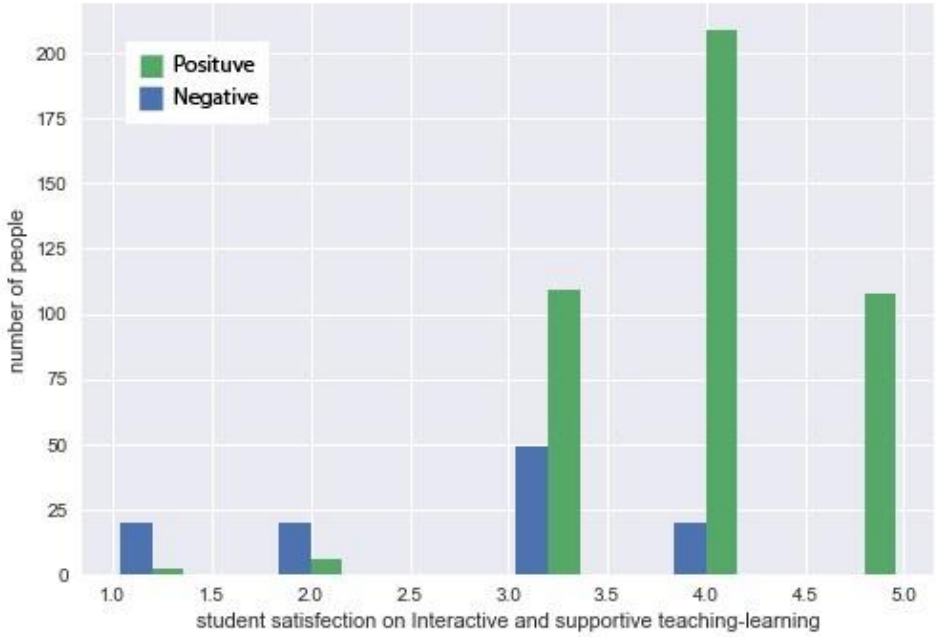


Figure-1

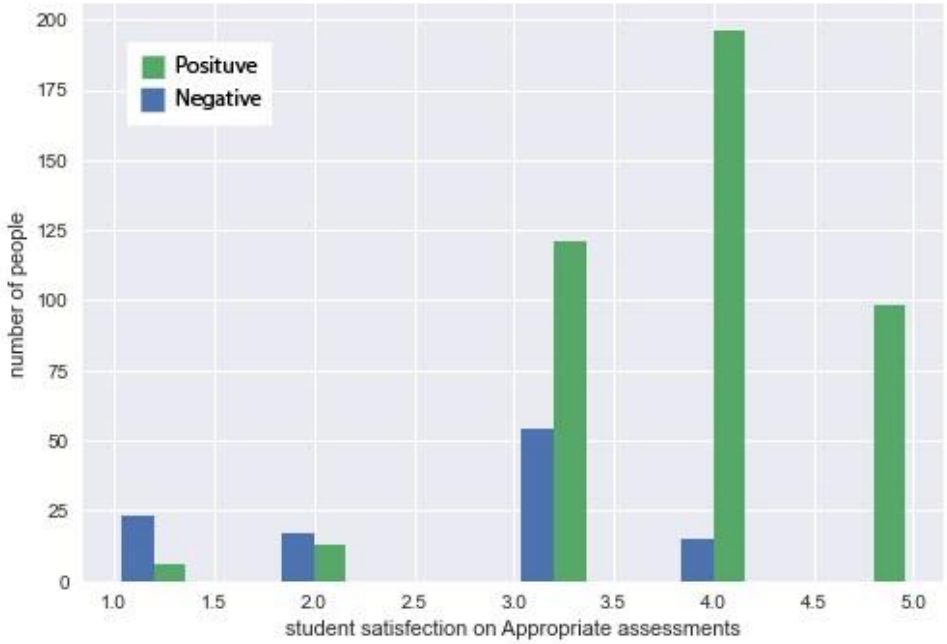
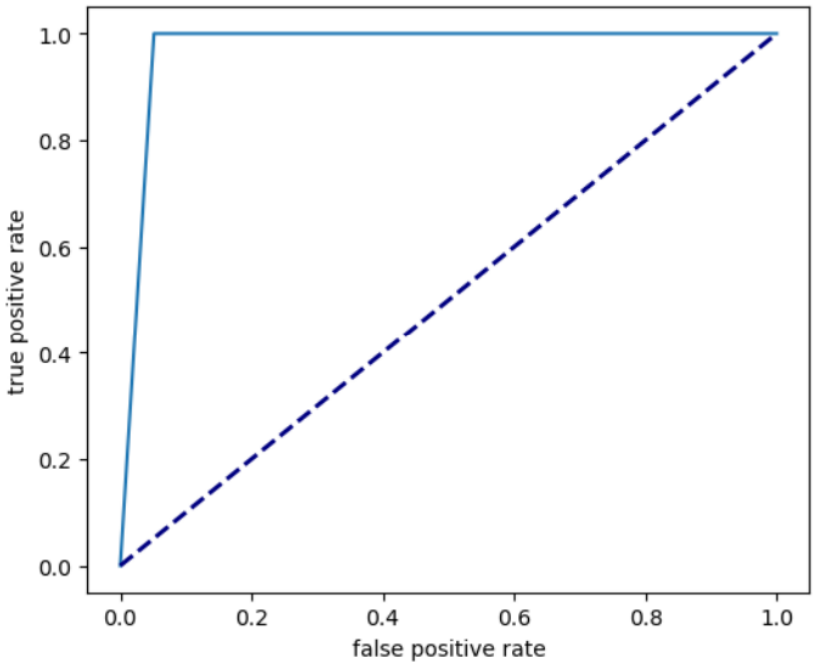


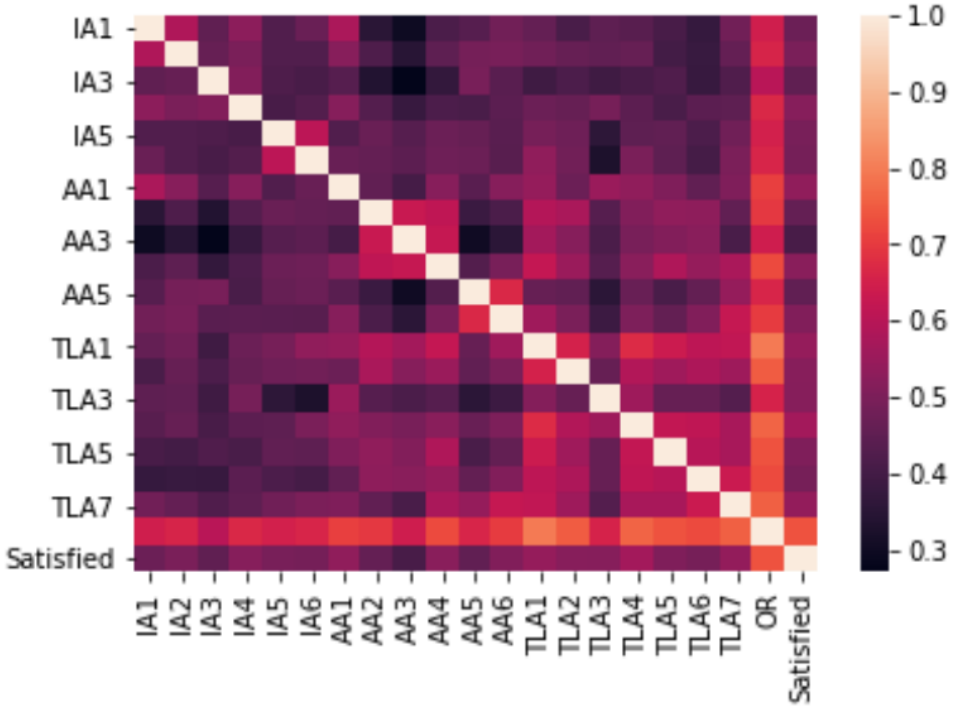
Figure- 2

From the above figure, we can see that student's satisfaction level depend on appropriate assessment, interactive behavior and supportive teaching-learning.

ROC curve using Logistic Regression



So, what should university do based on this analysis -



The correlation among all kind of university activities with Satisfied column is 40-50 %. In this case, university have to work on all kind of activities to increase the satisfaction level.

Conclusion

Learning is not going to be fulfilled if you don’t apply them. This project helps us to fulfill that gap. I believe that a project is more useful than 100 exams. It gives us an opportunity to visualize the real-life problems. Many of us knew about these algorithms and have enough knowledge about them. But this project gives us that opportunity to apply them to a particular problem and get the best output. Now it will be easier for us to learn new algorithms in an efficient way and handle any machine learning problem.