Prediction of student performance in secondary education.

Name: AHMED SHMELS MUHE

REG_NO:GE22M009

M.Tech in Data Science

Indian institute of technology Madras

Objectives

- To find out what features most affect student achievement
- Descriptive analysis, and correlation studies of the given data.
- To classify the data using classification algorithms(SVM,KNN,logistic).
- Compare the three ML models
- Find the best algorithm with high accuracy

Data processing

- Feature Engineering:

Datasets came up with non-numerical values and it is impossible to give them to any classifier. So, I convert non-numeric values to numerical ones/ **Digitization of values.**

- Feature scaling: is a method used to normalize the range of independent variables or features of data, This will help our model to converge quickly. Just by calling :
- Data Normalization
- z-score normalisation/Standardization
- **Missing** value check: **No missing** values in the data, The shape of our data set is (**395 rows × 31 columns**).

$$\frac{col - mean(col)}{std(col)}$$

$$\frac{col - mean(col)}{max(col)},$$

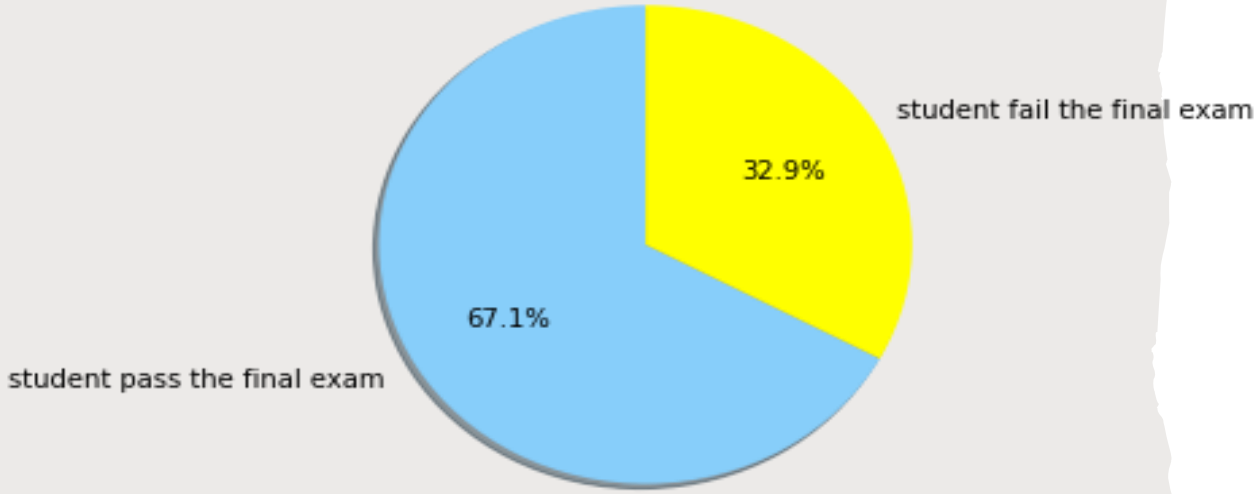
Correlation of the Data

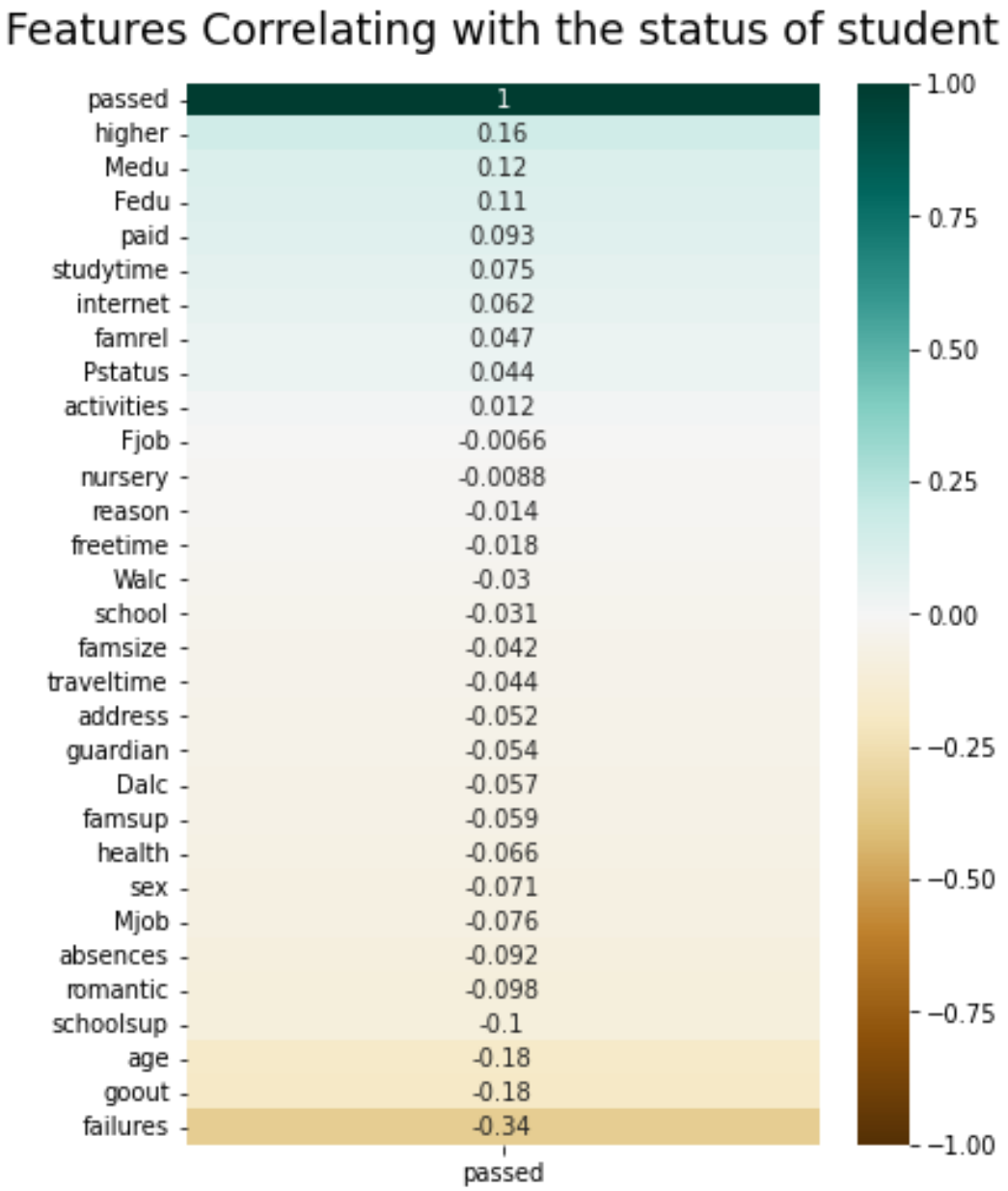
- This correlation could be positive or negative.
- Positive correlation: This means the variable ↑, the related variable also ↑,

↑StudyTime↑FinalGrades

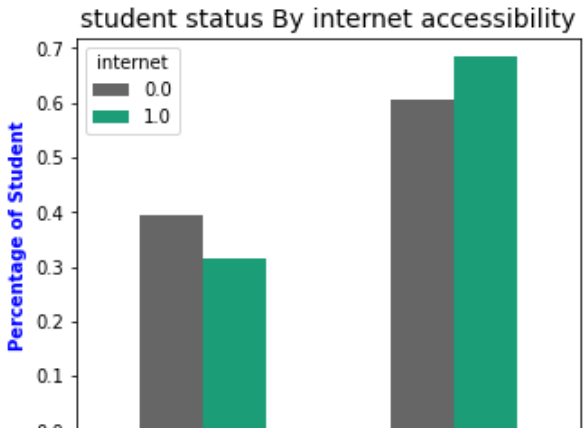
- Negative correlation: This means that as the variable ↑, the related variable ↓,

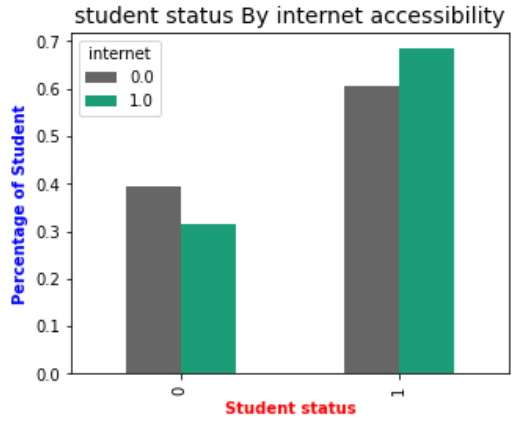
↑Failures↓FinalGrades

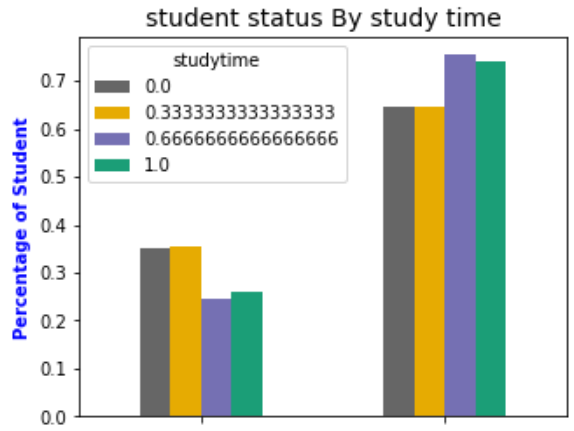


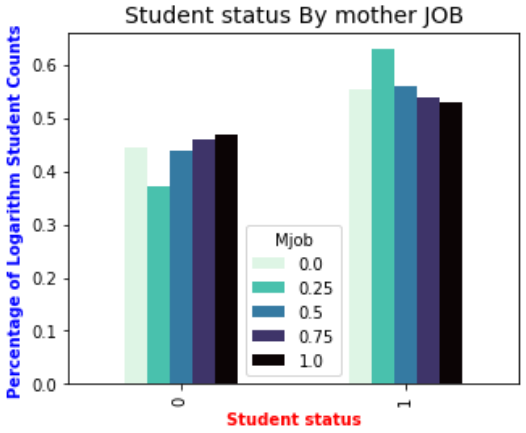


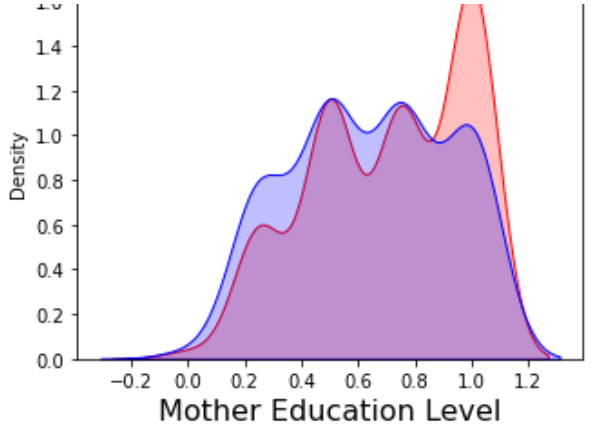
Features have a +ve impact on academic performance

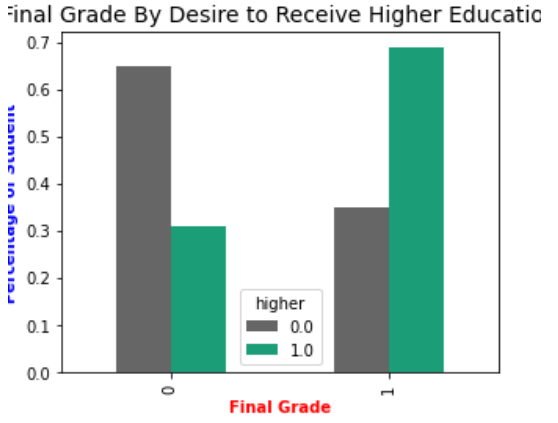




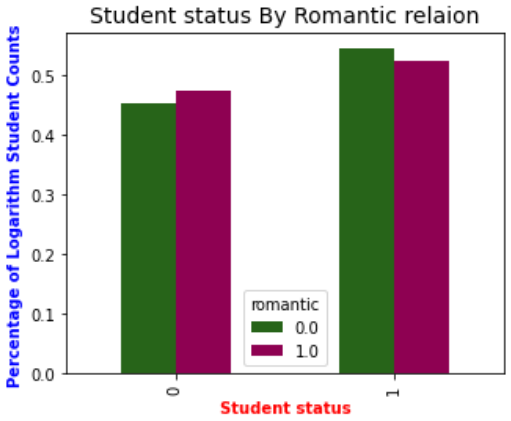


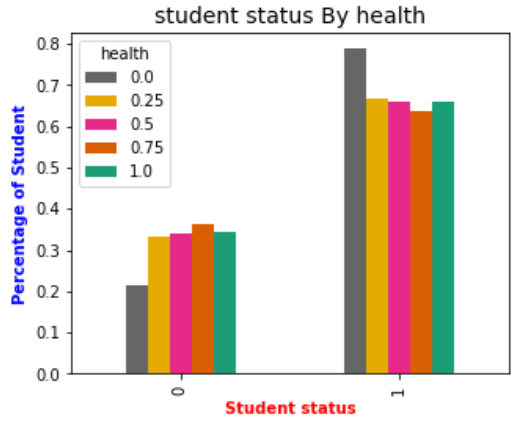


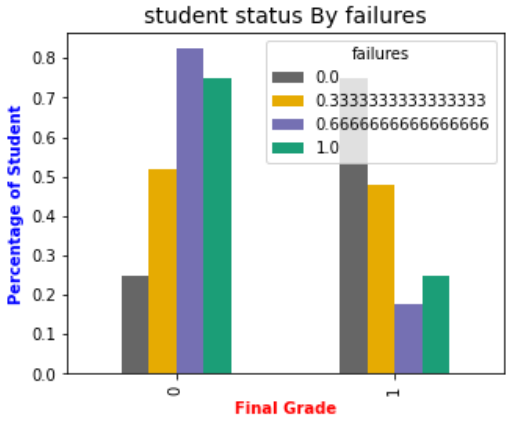


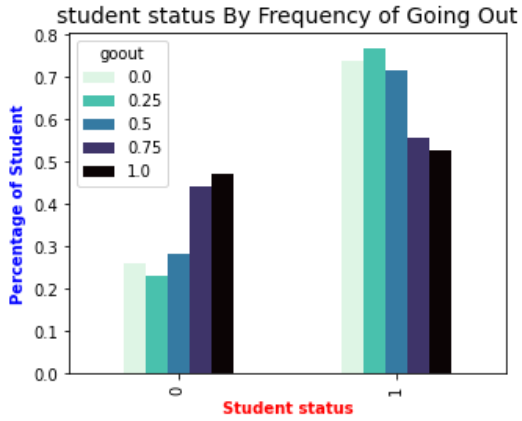


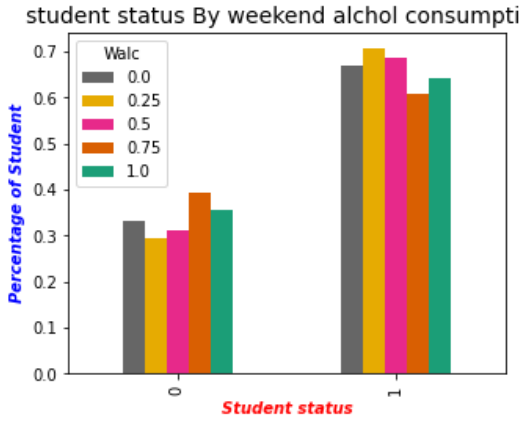
Features have a -ve impact on academic performance

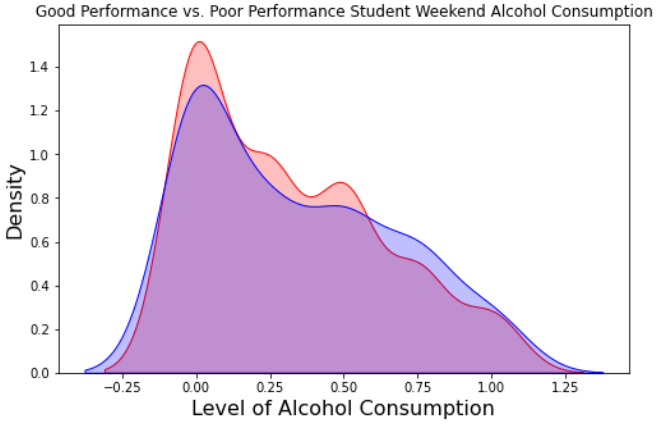






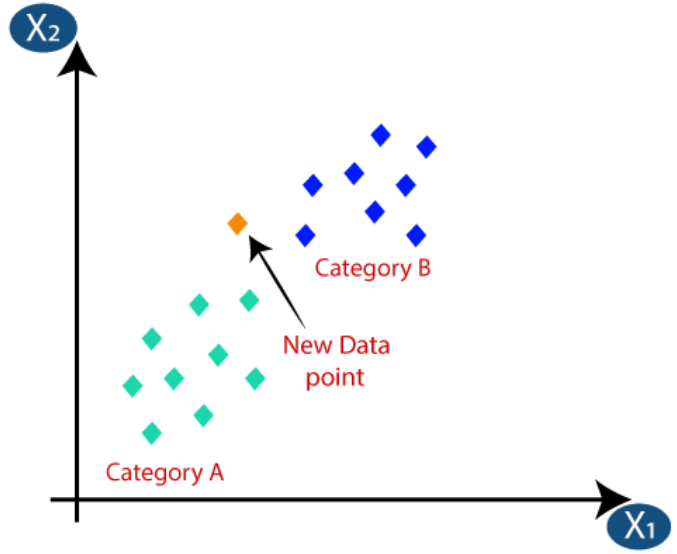


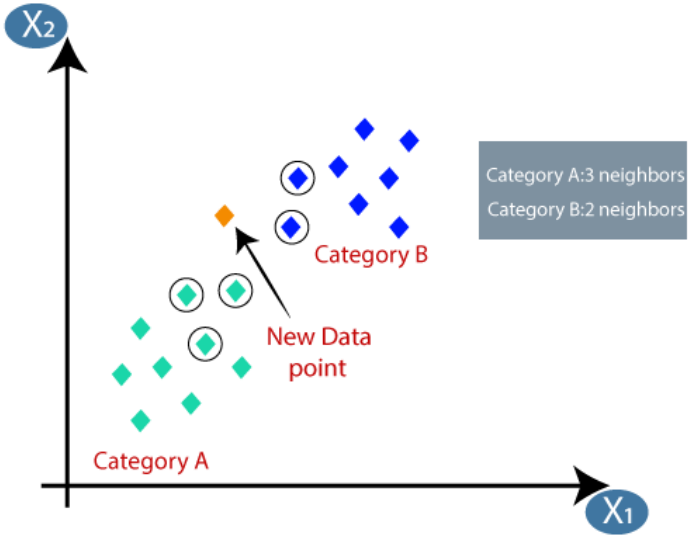




KNN

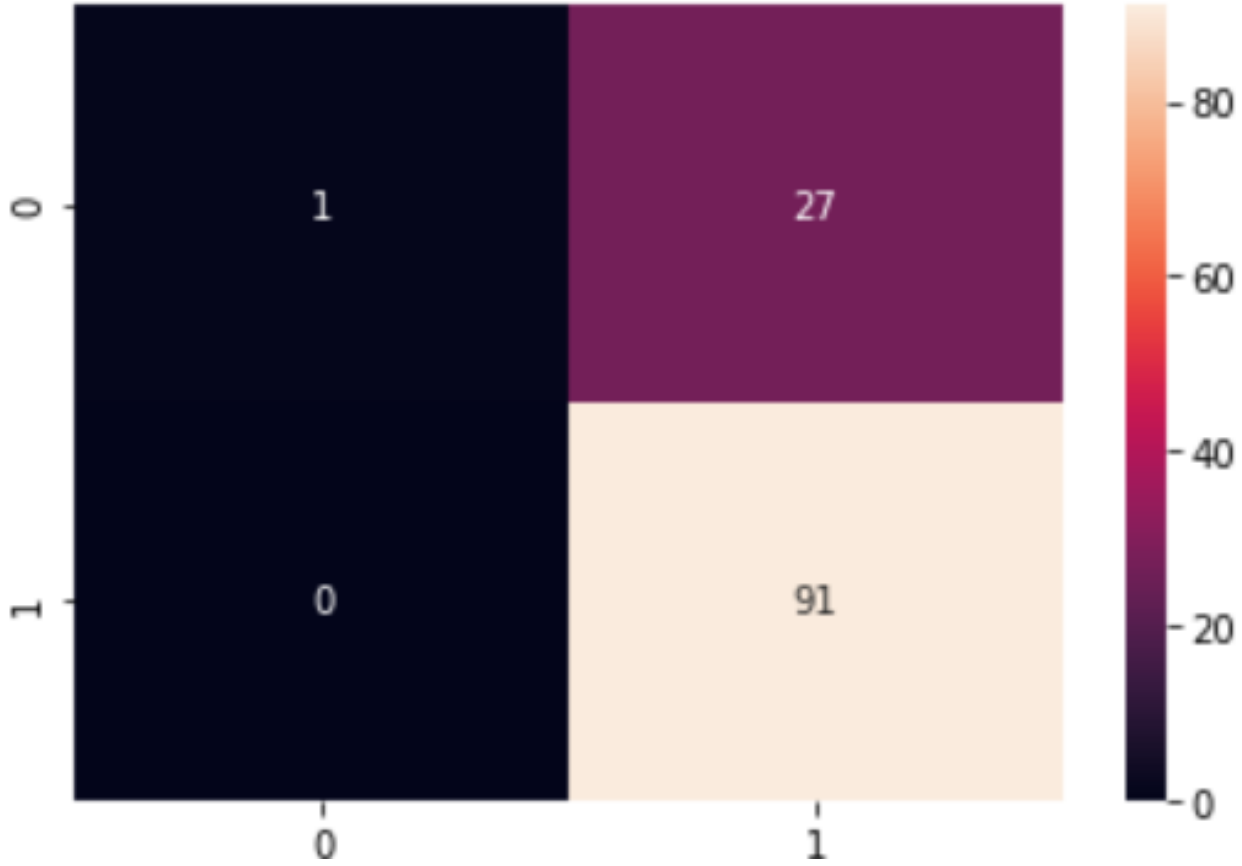
- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean(or any other type of distances) distance of K number of neighbors
- Step-3: Among the k nearest neighbors, count the number of the data points in each category.
- Step-4: Assign the new data points to that category for which the number of the neighbor is maximum.
- Suppose we have a new data point and we need to put it in the required category. Consider the below image:



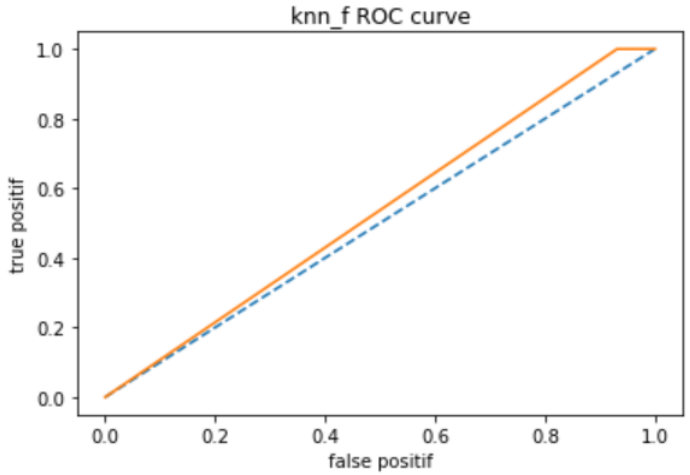


KNN classification result

Accuracy is: 0.773109243697479



the ROC curve:



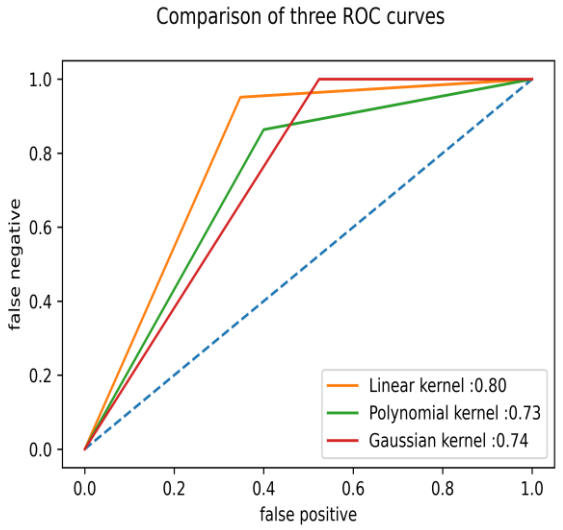
2)**classification_report**

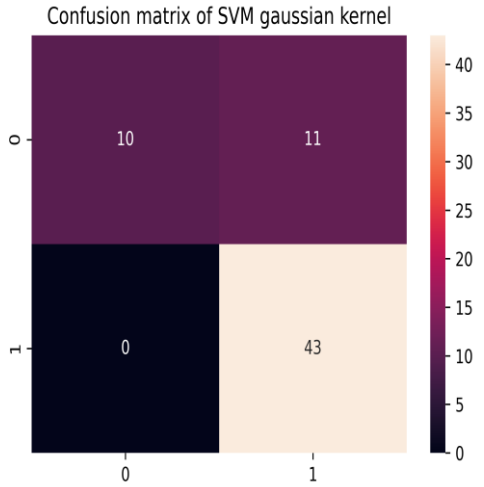
```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0.0	1.00	0.04	0.07	28
1.0	0.77	1.00	0.87	91
accuracy			0.77	119
macro avg	0.89	0.52	0.47	119
weighted avg	0.83	0.77	0.68	119

SVM classification result

- SVM uses the kernel trick technique to transform The data and then finds an optimal boundary between the possible outputs based on these transformations. We will use three kernels:
- Linear, polynomial and Gaussian kernel.
- Gaussian kernel performed better





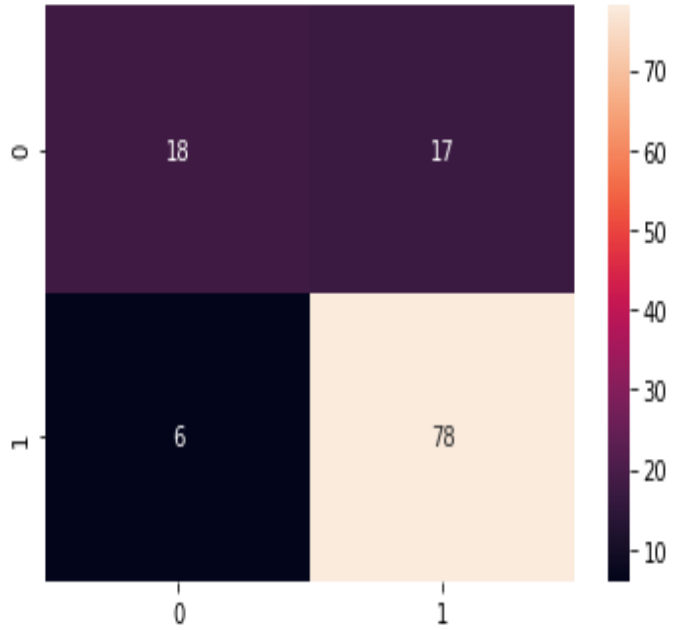
Metric	Linear kernel	polynomial kernel	gaussian kernel
training time	11ms	7ms	3ms
accuracy %	84.375	78.125	82.8125
confusion matrix	[15 8] [2 39	[12 8] [6 38]	[10 11] [0 43]
f1 score	0.82	0.74	0.77
roc_auc_score	0.80	0.73	0.74

Logistic Regression

- The algorithm gives different accuracy each time we change the data split:
- accuracy should not vary too much depending on the random state
- **Improving model accuracy:**
- instead of using the values "0" and "1" for the random state, we will choose the value "optimal_state" this maximizes the accuracy and the F1 score.

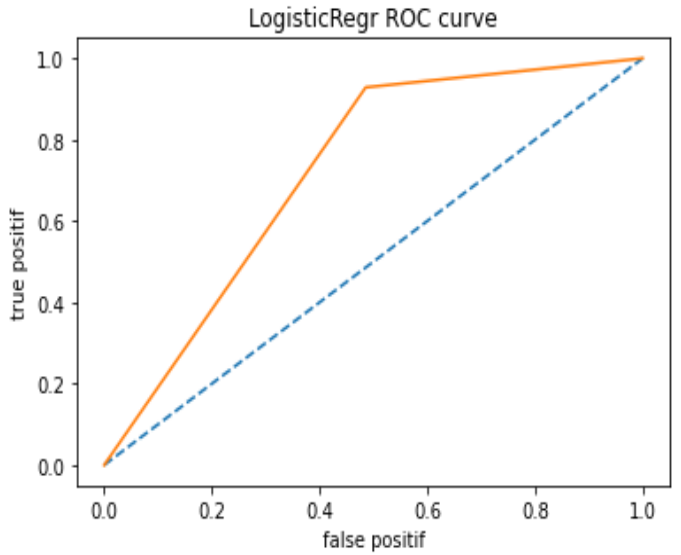
```
#import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0.0	0.71	0.24	0.36	50
1.0	0.63	0.93	0.75	69
accuracy			0.64	119
macro avg	0.67	0.58	0.55	119
weighted avg	0.66	0.64	0.58	119

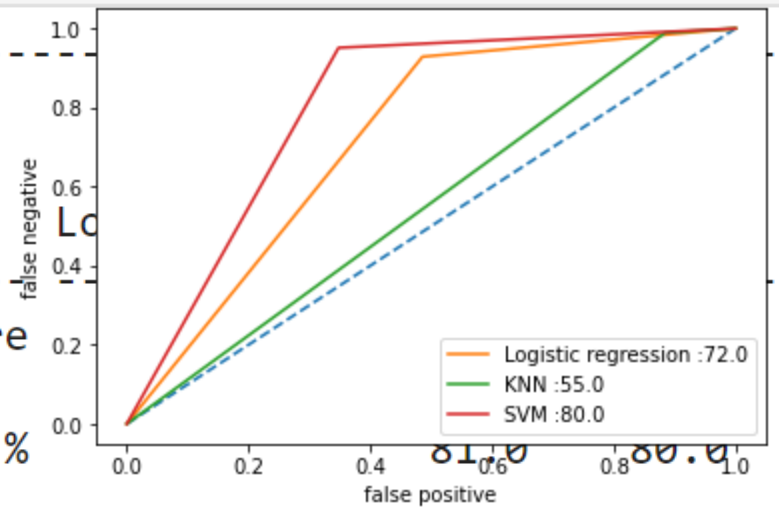


*Accuracy is: 80.67226890756302

*f1 score is: 0.7408389357068459



```
compare_lg_knn_svm(yt_knn,yp_knn,yt_lg,yp_lg,yt_svm,yp_svm)
```



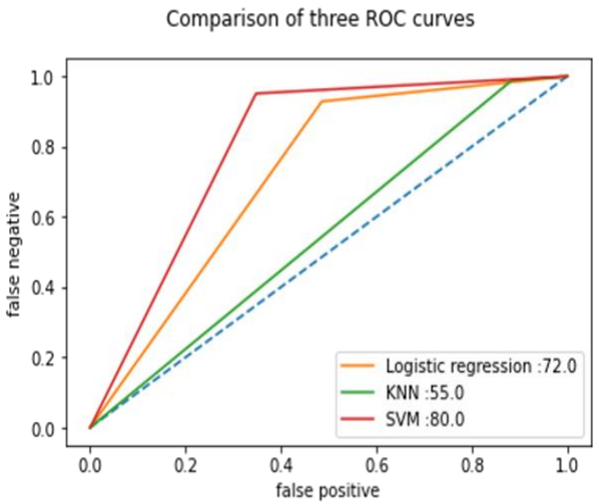
metric	Lo		SVM
f1 score			82.0
accuracy %	81.0	80.0	84.0
confusion matrix	[18 17] [6 78]	[3 23] [1 92]	[15 8] [2 39]
ROC score	72.0	55.0	80.0

Comparison of the three algorithms

- The metrics that we will be used to compare those three algorithms are :

1. F1 score.
2. Accuracy score.
3. Confusion matrix.
4. ROC curve.
5. ROC score.

- SVM classifier with high accuracy of **84%**, the confusion matrix is roughly diagonal which indicates that this classifier is able to label data correctly.




```
compare_lg_knn_svm(yt_knn,yp_knn,yt_lg,yp_lg,yt_svm,yp_svm)

----------------------------Table of metrics----------------

    metric       Logistic regression   KNN     SVM
--------------- -------------------- ------- -------
        f1 score                 74.0    54.0    82.0

      accuracy %                 81.0    80.0    84.0

confusion matrix           [18 17] [ 3 23] [15  8]
                           [ 6 78] [ 1 92] [ 2 39]

       ROC score                 72.0    55.0    80.0
```

```
    metric      Learning algorithm winnig
-------------- ---------------------------
  max f1 score                         SVM

max accuracy %                         SVM

 max ROC score                         SVM
```

Conclusions

Factors helping students succeed :

father's education
guardian
wants to take higher education
studytime
father's job

Factors leading students to failure :

age
health
going out with friends
absences
failures

metric	Logistic regression	KNN	SVM
f1 score	74	48	82
accuracy %	81	70	84
confusion matrix	[18 17] [6 78]	[3 23] [13 80]	[15 8] [2 39]
ROC score	72	49	80

THANK YOU!