National Graduate School of Arts and Crafts
ENSAM Meknes

Industrial Engineering : Artificial Intelligence Data Science

Project of Data Science Foundation

Analysing students performance

June 20, 2022

Submitted by :

FAYTOUT Achraf

IZEM Mourad

Supervised by:

M. Abdelkader FASSI FIHRI
M. Mustapha EL OSSMANI

Contents

1 Abstract

Academic performance is among the several components of academic success. Many factors, including socioeconomic status, student temperament and motivation, type of social facilities, peer, and parental support influence academic performance. Moreover, International surveys such as PISA have attracted considerable attention from the media and policy makers. In particular, focus has been on the relative rankings of countries on the basis of students' average achievement scores. The question that could be asked is whether we can compare academic performance of different countries, what we are trying to do next in this project, applying machine learning in a US-dataset that could reflect some of the factors that influence this target.

"Success is not the key to happiness. Happiness is the key to success. If you love what you are doing, you will be successful." – Herman Cain

---

2 Project presentation

Our main objective is to visualize, interpret the data to better understand it and make the necessary preprocessing before moving on to the Machine Learning approach by exploiting the algorithms of regression to predict the grade of the students in the exam according to different features that affect it .

2.1 DATASET

It is an American data-set that deals with the different variables that influence performance college students.

- Gender : female, male
- State : New York, Florida, California, Minnesota, Texas
- Parental level of education : Master's degree, High school, College,...
- Lunch : Standard, Free/Reduced
- Test Preparation : None , Completed
- Math, reading, writing score : Numerical variables

3 DATA Visualization

First, in order to better understand the distribution of our data, we tried to visualize the different features and the relationships between them. So we can certainly merge the three features later into a single feature.

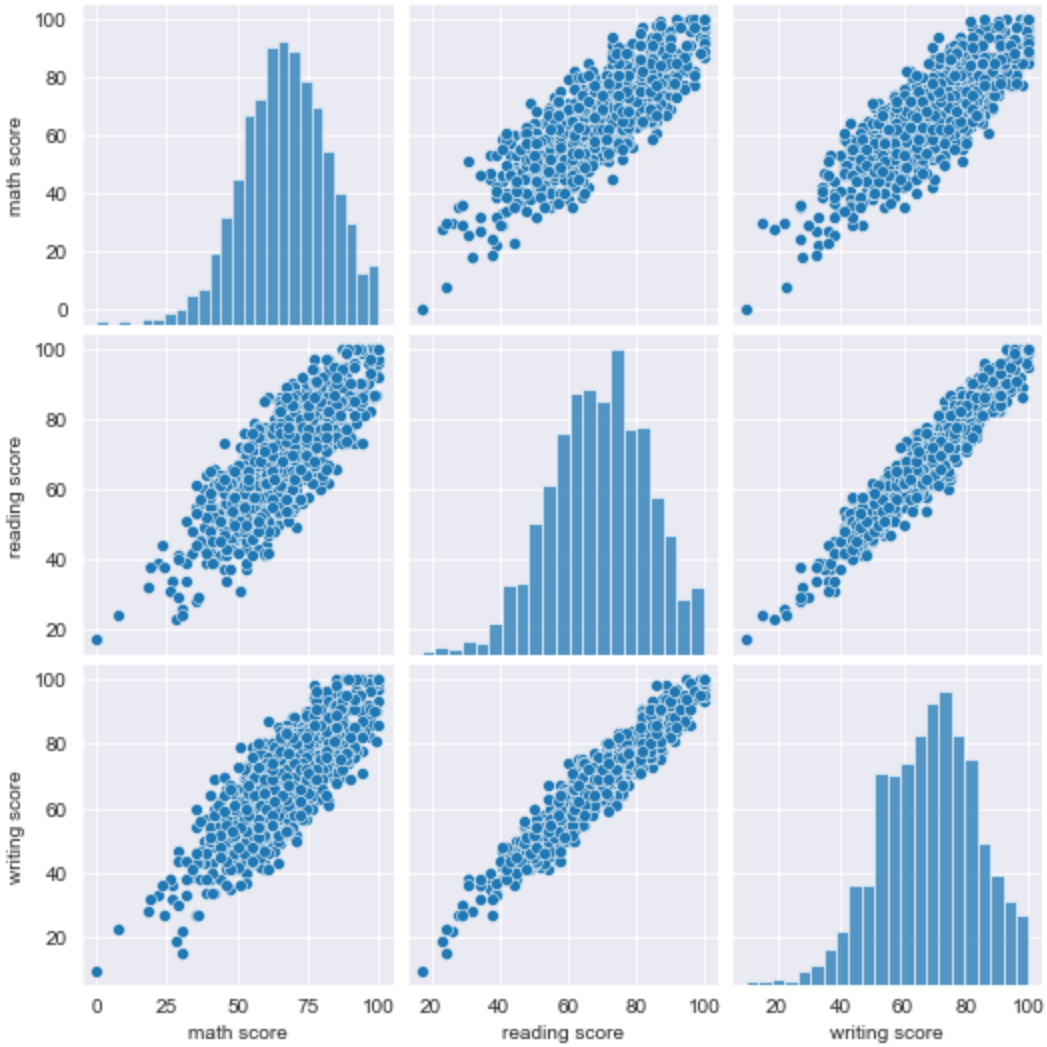

Figure 1: Correlation between numerical variables

We can clearly see that there is a strong correlation between the reading score and writing score and a less correlation between the latter and math score.

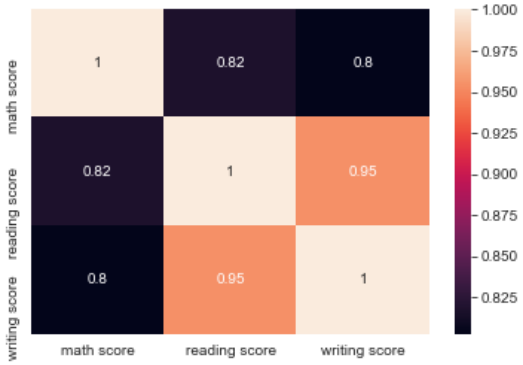

Figure 2: The HeatMap between numerical variables

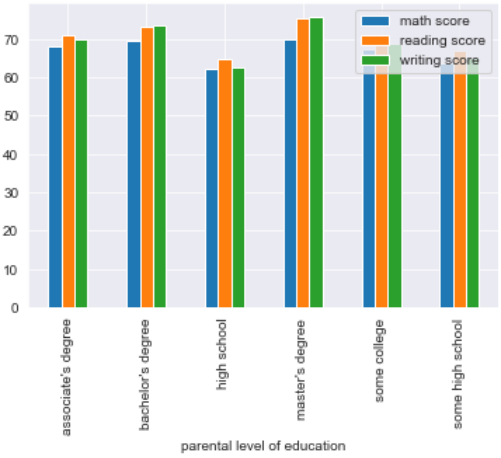

Figure 3: the grades in function of parental educational level

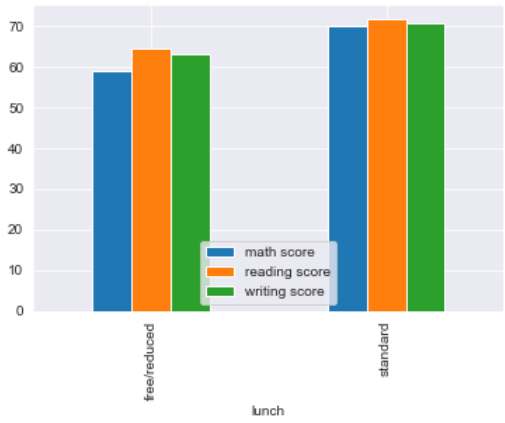

Figure 4: the grades in function of Lunch

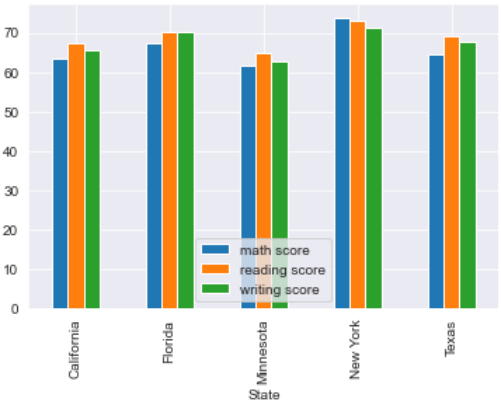

Figure 5: the grades in function of the state

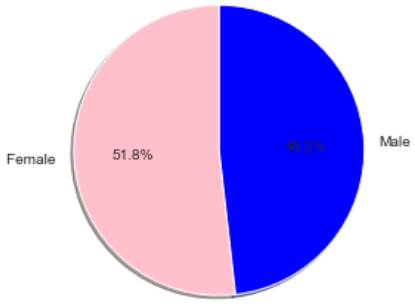

Figure 6: Distribution Male/Female

According to graphigues, we notice that:

- Students with parents with "master's degree" have high grades in maths, reading and writing.
- The majority of students take the standard lunch, which may affect their results.
- Students in New York have the highest scores.
- Women present the majority of students.

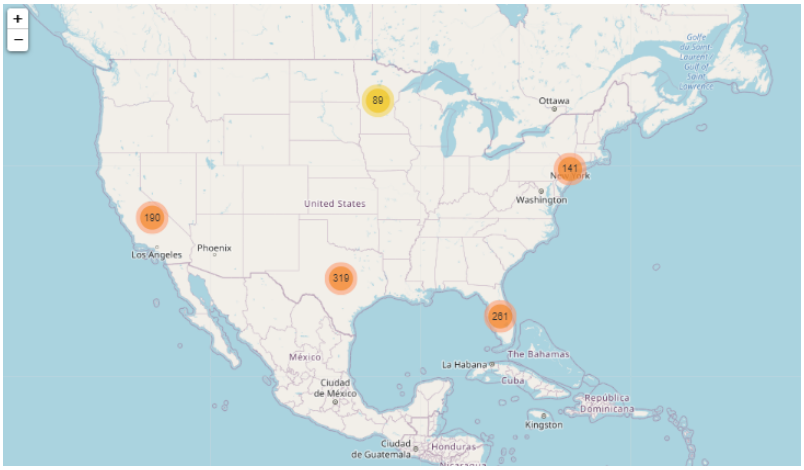

Figure 7: Visualization of students by state using Folium

Using Marker cluster of Folium and geocode, we also remark that the most of students in this study case are from Texas.
As the threescores are correlated with each other, we decided to consider only the average score afterwards.And we plotted the box plot of the average grade according to the other features.

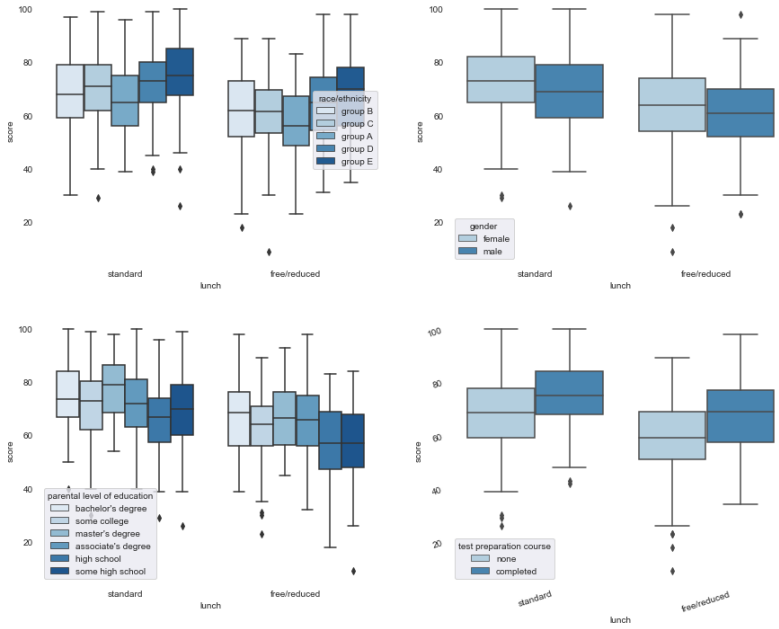

Figure 8: The average grade in function of other features

students who take a standard lunch, belonging to New York State, Women, whose parental educational level is master's degree, who have well prepared of course :) have a higher average score.

4 DATA Preprocessing

4.1 Removing the outliers

Before we started the training of machine learning models on our dataset, we have removed the outliers of the mean score by adopting the interquartile method for better performance and in order to get a normal distribution.

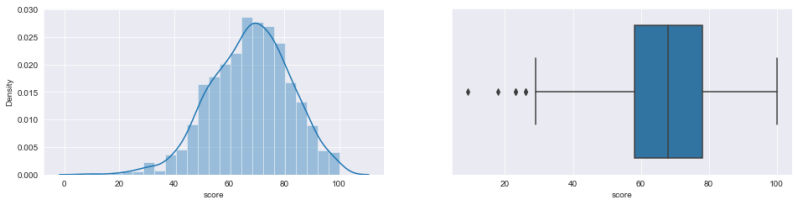

Figure 9: Before removing the outliers

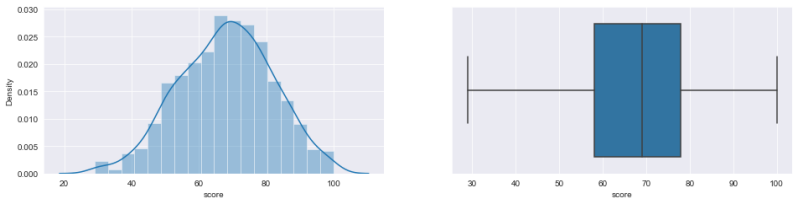

Figure 10: After removing the outliers"

We can see clearly that after removing outliers we got a normal distribution of the mean score with P—value = 0.027 < 0.05.

4.2 Missing Values

We have filled the missing values in numerical variables by mean and in categorical variables by mode.

```
df['score'].fillna(df['score'].mean(), inplace=True)
cols = cat_vars.tolist()
for col in cols :
    df[col] = df[col].fillna(df[col].mode()[0])
df.isnull().sum()
```

```
gender                        0
parental level of education   0
lunch                         0
test preparation course       0
State                         0
score                         0
dtype: int64
```

Figure 11: Filling the missing values.

4.3 Data libelling

we labeled the categorical features (Launch, Parental level of education, State and Gender).

Feature	label	elements
State	0	Minnesota
	1	California
	2	Texas
	3	Florida
	4	New York
parental level of education	0	associate's degree
	1	bachelor's degree
	2	high schools
	3	master's degree
	4	some college
	5	some high schoole
Lunch	0	Standard
	1	Free/reduced
Gender	0	Female
	1	Male
test preparation course	0	Completed
	1	None

Table 1: Data libelling

5 Prediction of students perfomance

In this part, we will exploit the RandomForest Classifier model to predict whether the student will succeed or not.

5.1 RandomForest Classifier

Decision trees are used for regression and classification problems. They present themselves visually as trees, hence their name, and in the case of regression, they start with the root of the tree and follow divisions based on the results of the variables until a leaf node is reached and the result is given.
We trained sklearn's RandomForest Classifier model on our dataset split into 20% test and 80% for training and testing it afterwards, we obtained an Accuracy of 85.353%.
The graph of important features during the training is as follows:

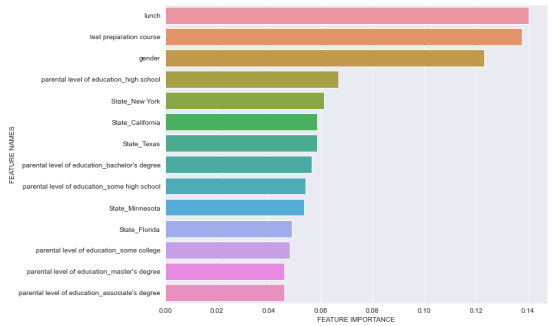

Figure 12: Features Importance.

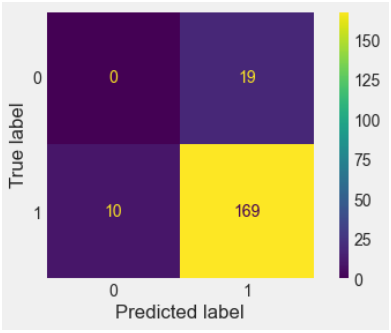

Figure 13: Confusion matrix of RFC model.

We notice for our model: "Lunch", "test preparation race" and "Gender" are the first 3 most important features.

5.2 Other used Models

We tried using SVM (support vector machine) model for better accuracy, and why not increasing the specificity of the classification, after training our model we measure the quality of the classification and we obtained an accuracy equal to 90.404%.

```python
from sklearn.svm import SVC
# Training the model
model_svm = SVC()
model_svm.fit(x_train,y_train)
# predition
y_pred = model_rfc.predict(x_test)
# Model quality
accuracy = accuracy_score(y_pred, y_test)
print('Model Quality : \nAccuracy= ',accuracy * 100)
plot_confusion_matrix(model_rfc, x_test, y_test)
plt.grid()
plt.show()
```

Figure 14: Implementation of SVC model.

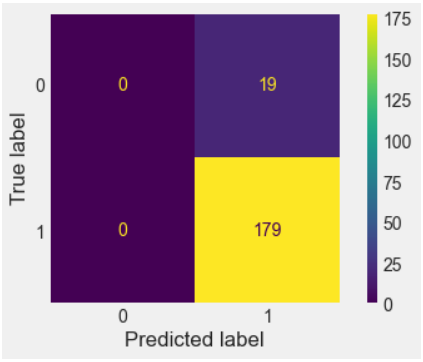

Figure 15: Confusion matrix of SVC model.

6 Model deployment

Our project uses a SVC model (has higher accuracy), but can be modified to fit any model we want to deploy. We did use for the deployment flask framework.

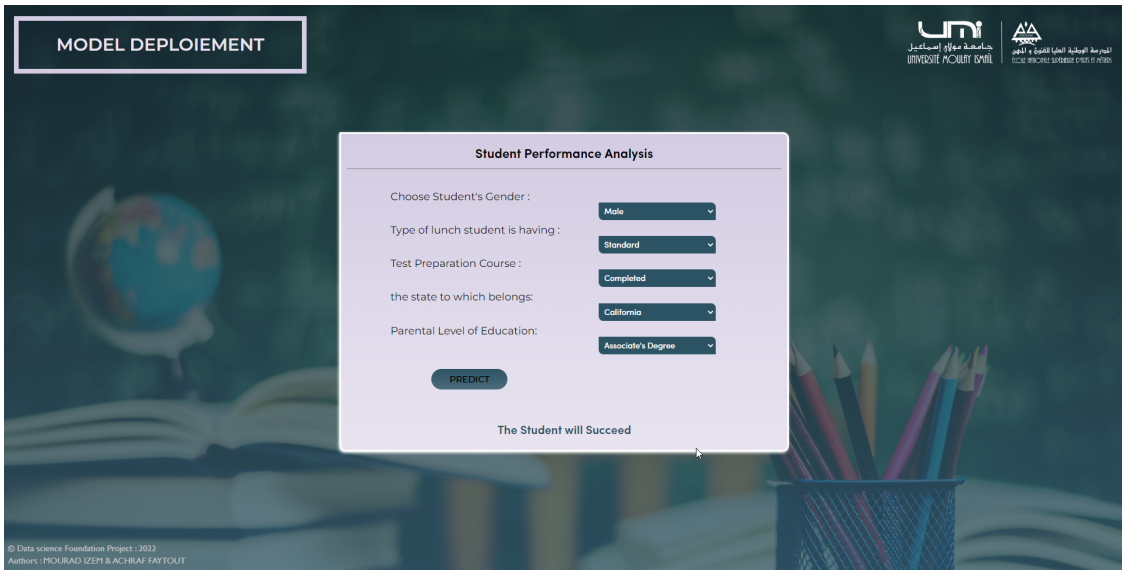

Figure 16: Deployment using Flask.

7 Conclusion

In this project, we had the opportunity to use the various methods of data visualization and preprocessing as well as machine learning algorithms such as RandomForest Classifier, on the other hand, we deployed our program in a web application, where the user can have a clear idea of this data set and choose which of the models to use to predict performance. Finally, by choosing each value of a feature, the user can simply know the prediction whether or not the student will pass the final exam.