4-th year of Engineering School

PYTHON FOR DATA ANALYSIS PROJECT

BY ESTELLE FIKET

01/05/2022

INTRODUCTION

For this project, I worked alone.

The objectives of this project are to implemented a python notebook to analyse a dataset. What we have to do:

Create a Gihtub Repository (to version some document on it)

Implemented a notebook with my work on the dataset assigned

Explaining the context, the problem and what I have done on the notebook

Implemented a Flask API

REQUIREMENTS

For each dataset, I used the language Python (version: 3.9 with respect to the PIP8 rules), and the following librairies:

- Pandas (Data Manipulation and Data Analysis)
- Seaborn and Matplotlib (Data Visualization and Data Analysis)
- Sklearn (Machine Learning)
- Flask (API)

I used DataSpell IDE to work on this project.

All the instructions to use the code implemented in the repository are written on the Read.me file.

OBJECTIVES FIXED FOR EACH DATASET

- To check if the information given is correct.
- To get other information in applying analysis on the different features such as:
    - Type of problem (regression or classification),
    - Which feature is the target,
    - Type of the different features + some statistics,
    - Checking the quality of the dataset.
- To fit several model
- To choose the best models and to search for the hyperparameters
- Discussion/Conclusion about the dataset, the results obtained, and the project in general.

WHAT WILL BE EXPLAIN IN THIS SLIDE SHOW?

Table of Contents

In this slide show, I will only explain the context, the problem for each dataset, and what I have done in the notebooks. Each step is already detailed in each notebook.

And then I will provide some explications about the main results and a discussion to conclude the analysis of each dataset.

Dataset

The dataset that has been assigned to me was the Blocks Classification one. During all the process, some things pushed me to choose another dataset in the end to work on, and to implement other method that we have seen in this course. My goal was to implement almost each subject that we have seen from each practical work.

I will first present my work on this first dataset. Then, I will detail what I have done on the second dataset (Seoul Bike), chosen randomly among the dataset from the Google Sheet shared to the class. To conclude this slideshow, a description of the almost finished implemented API will be given with a discussion around this project.

PAGE BLOCKS CLASSIFICATION DATASET

Link: https://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification

BLOCKS CLASSIFICATION: PRESENTATION

There are 5473 examples coming from 54 distinct documents and each observation concerns one block.

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. There are five classes:

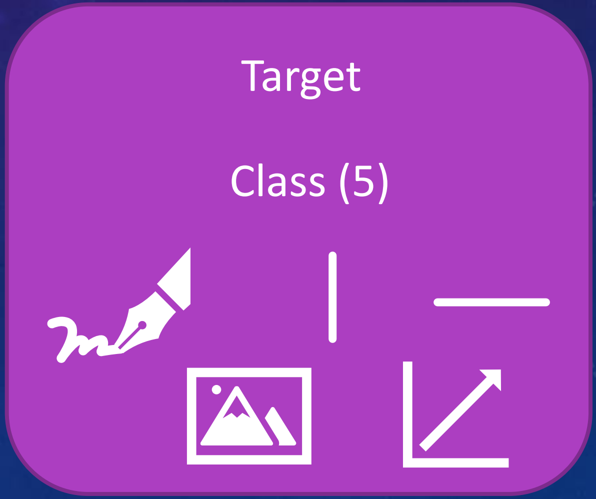
- Text (1),
- Horizontal line (2),
- Picture (3),
- Vertical line (4)
- Graphic (5).

BLOCKS CLASSIFICATION: PROBLEM

Supervised → Classification Problem

Classifying all the blocks of the page layout of a document:

10 Features



Target

Class (5)

BLOCKS CLASSIFICATION: STEPS FOLLOWED FOR THIS WORK

- Data Collection
- First Exploratory Data Analysis
- Data Preprocessing
    - Missing Values
    - Null & Zero Values
    - Outliers
    - Categorical Features
- Imbalanced Class & Outliers Management
- Second Exploratory Data Analysis
    - Univariate Analysis
    - Bivariate Analysis
    - Correlation
- Data Processing & Feature Engineering
- Models Fitting
    - Model Selection: Training and Cross Validation
    - Hyperparameters for the selected models

BLOCKS CLASSIFICATION: DATA PREPROCESSING (1/2)

No NaN
&
No Null or Zero
values

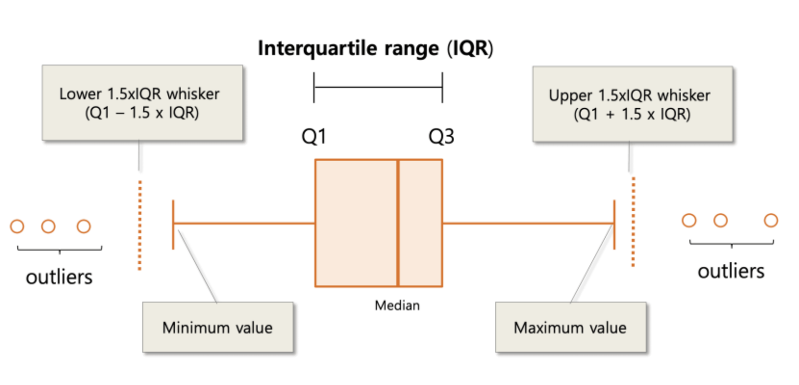
Almost all features
have outliers
→ Treatment to do

No categorical
features

BLOCKS CLASSIFICATION: DATA PREPROCESSING (2/2)

Outliers

→ Interquartile method:


Interquartile range (IQR)
Lower 1.5xIQR whisker (Q1 – 1.5 x IQR)
Upper 1.5xIQR whisker (Q1 + 1.5 x IQR)
Q1
Q3
outliers
outliers
Median
Minimum value
Maximum value

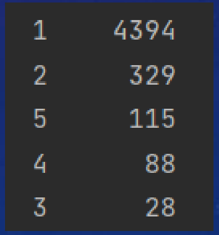
https://help.ezbiocloud.net/box-plot/

If the data is an outlier, I checked if it came from the class 1. If yes, then we delete, else we keep then since there is not so much outliers from the other class

Imbalanced Classes

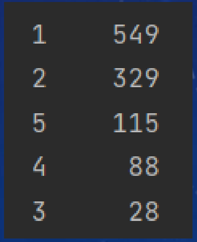
→ Random sample by fraction:

I decided to keep 1/8 of the class 1, this sample taken randomly.



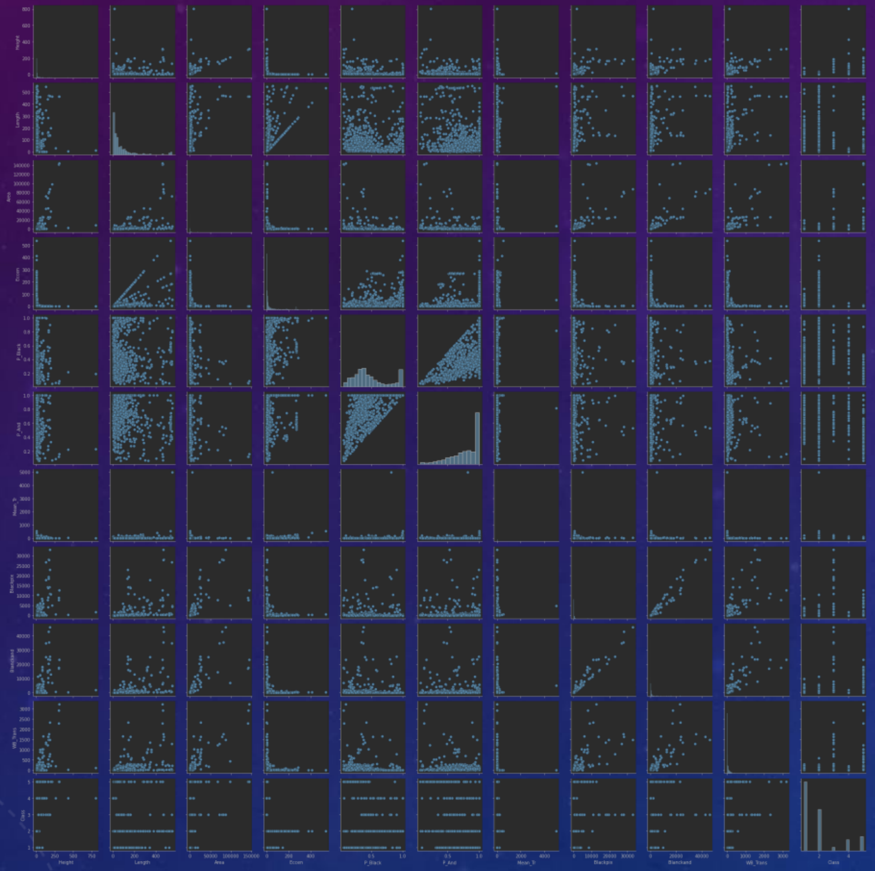
```
1    4394
2     329
5     115
4      88
3      28
```

After outliers removing from the class 1, I treated the imbalanced classes problem

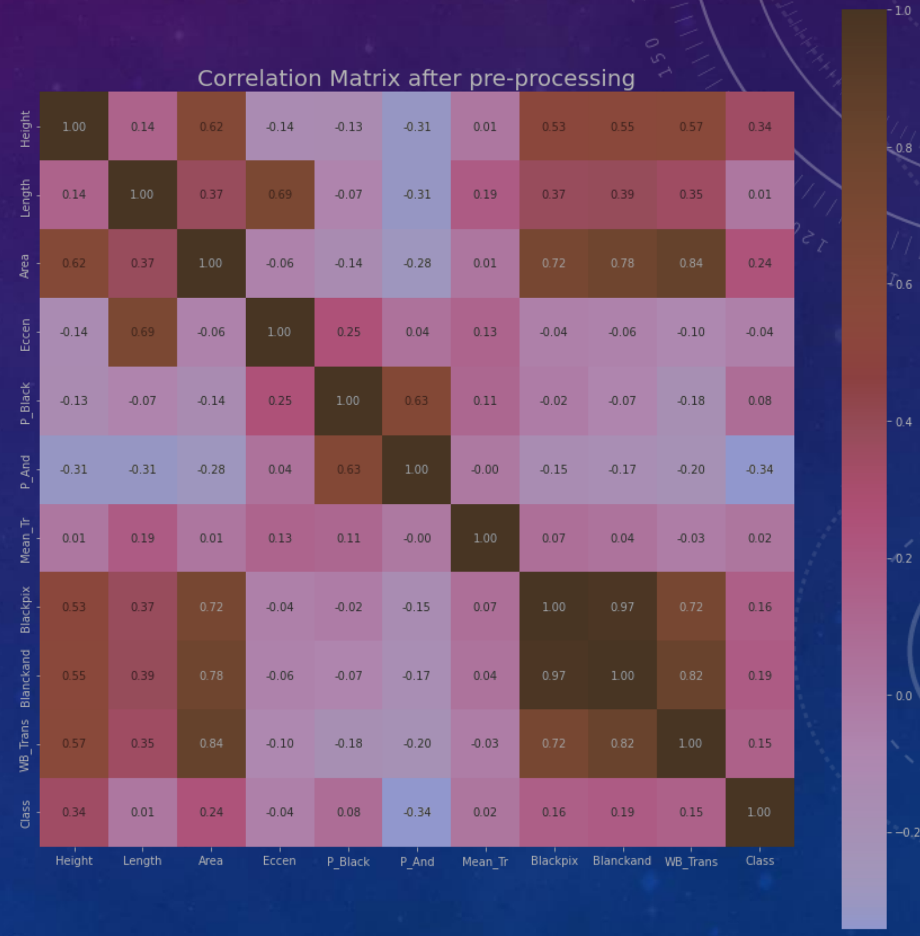


```
1    549
2    329
5    115
4     88
3     28
```

BLOCKS CLASSIFICATION: EDA – SUMMARY



Bivariate Analysis


Correlation Matrix after pre-processing

	Height	Length	Area	Eccen	P_Black	P_And	Mean_Tr	Blackpix	Blanckand	WB_Trans	Class
Height	1.00	0.14	0.62	-0.14	-0.13	-0.31	0.01	0.53	0.55	0.57	0.34
Length	0.14	1.00	0.37	0.69	-0.07	-0.31	0.19	0.37	0.39	0.35	0.01
Area	0.62	0.37	1.00	-0.06	-0.14	-0.28	0.01	0.72	0.78	0.84	0.24
Eccen	-0.14	0.69	-0.06	1.00	0.25	0.04	0.13	-0.04	-0.06	-0.10	-0.04
P_Black	-0.13	-0.07	-0.14	0.25	1.00	0.63	0.11	-0.02	-0.07	-0.18	0.08
P_And	-0.31	-0.31	-0.28	0.04	0.63	1.00	-0.00	-0.15	-0.17	-0.20	-0.34
Mean_Tr	0.01	0.19	0.01	0.13	0.11	-0.00	1.00	0.07	0.04	-0.03	0.02
Blackpix	0.53	0.37	0.72	-0.04	-0.02	-0.15	0.07	1.00	0.97	0.72	0.16
Blanckand	0.55	0.39	0.78	-0.06	-0.07	-0.17	0.04	0.97	1.00	0.82	0.19
WB_Trans	0.57	0.35	0.84	-0.10	-0.18	-0.20	-0.03	0.72	0.82	1.00	0.15
Class	0.34	0.01	0.24	-0.04	0.08	-0.34	0.02	0.16	0.19	0.15	1.00

Correlation Matrix

BLOCKS CLASSIFICATION: MAIN RESULTS - TRAINING

Feature Selection:

→ Height, Area, Blanckand, Blackpix, WB_Trans

Models Tested:

→ Kfold Cross-Validation

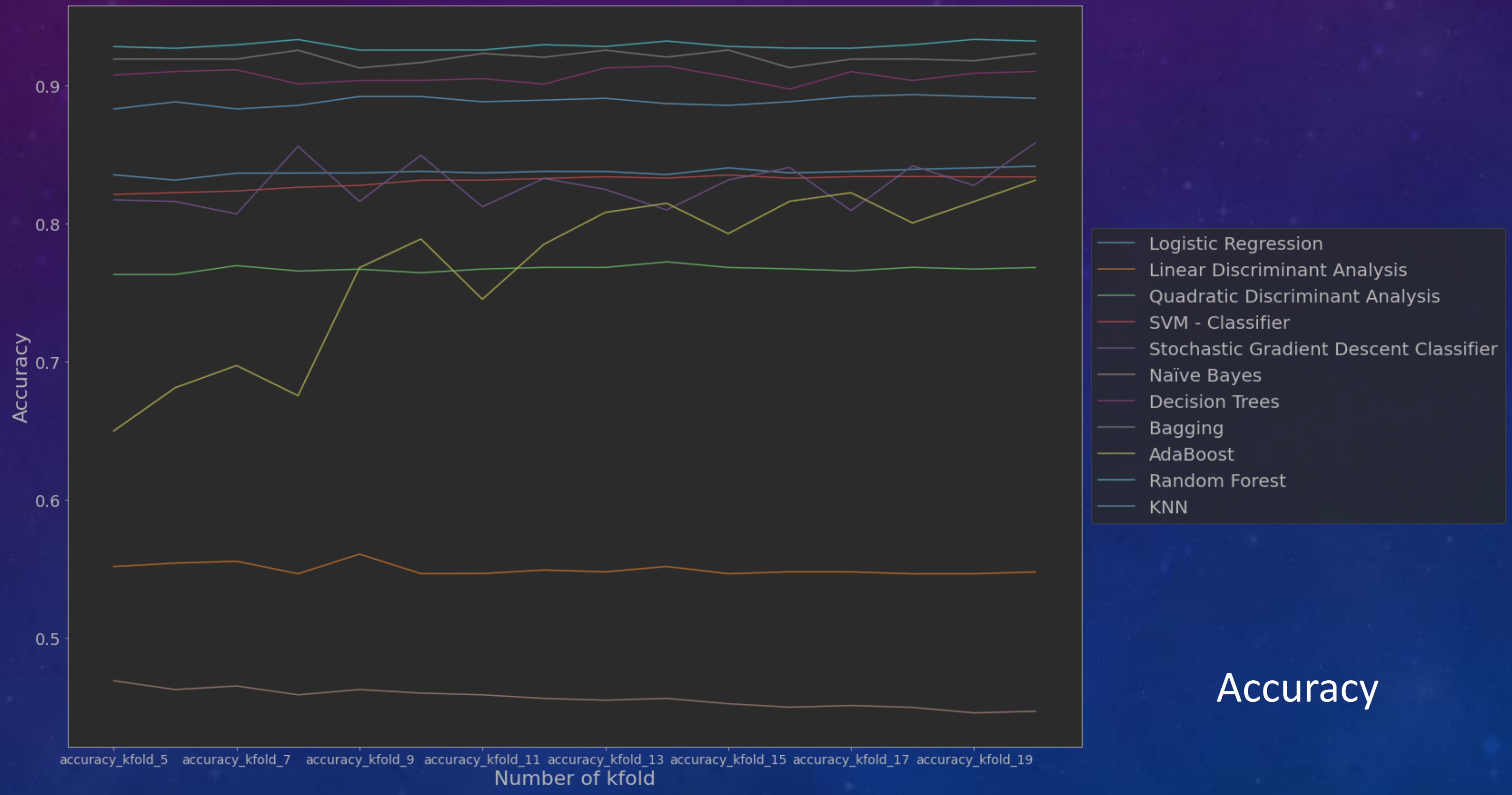
→ Logisitic Regression, LDA, QDA, SVC, Stochastic Gradient Descent Classifier, Naïve Bayes, Decision Trees, Bagging, AdaBoost, Random Forest, KNN

Models for hyperparameters tunning:

I took the best model from the step training, the middle one and an other which had bad performances

→ Random Forest, KNN, SVC

SEOUL BIKES: MAIN RESULTS - METRICS



Accuracy

BLOCKS CLASSIFICATION: MAIN RESULTS - HYPERPARAMETERS

Random Forest

- Initial Score: 0.5646
- Hyperparameters: {'criterion': 'entropy',
- 'max_depth': 60,
- 'min_samples_leaf': 1,
- 'min_samples_split': 2,
- 'n_estimators': 500}
- Final Score: 0.5976

KNN

- Initial Score: 0.7958
- Hyperparameters: {‘learning_rate’: 0.5, 'n_estimators': 100}
- Final Score: 0.8048

SVC

- Initial Score: 0.6728
- Hyperparameters {'C': 500, 'degree': 3, 'gamma': 1, 'kernel': 'linear’}
- Final Score: 0.7898

Problem here when I was searching for the hyperparameters. With only a max_depth of 60, I had the best score which was the same score. So the other parameters do not impact the model (but another one impacts the scores, check the conclusion (2/2))

BLOCKS CLASSIFICATION: DISCUSSION & CONCLUSION (1/2)

The first dataset allowed me to work on a classification problem.

Some aspects were challenging like the outliers management or the imbalanced classes to deal with. Those aspect are big problematics in professional real life project, so it was pretty interesting to discoverd some methods even if I didn’t have the time to implement them.

The treatments that I apply on my outliers and imbalanced classes have impacted the score of my models. But that allowed to avoid biaised models because of the first class which was the most important in this dataset.

Design Document

Template

Table of Contents

1 Introduction

1.1 Purpose

1.2 Background

1.3 Scope

1.4 Methodology

1.5 Evaluation Criteria

2 Executive Summary

2.1 Purpose of this document

2.5 Methodology, Tools, and Techniques

BLOCKS CLASSIFICATION: DISCUSSION & CONCLUSION (2/2)

Especially if we take a look at the imbalanced classes management, I took randomly samples (1/8 of the first class) to redo the first class and to have a better proportion between each class. Thus, when I run many time the notebook, I obtained different scores more or less better, depending on the sample that the algorithm took (since a part of the outliers was treated).

So, because of the fact that I didn’t succeed to do something that I can named as a ‘‘good job’’ with this dataset, I decided to take an other dataset to play with.

SEOUL BIKE SHARING DEMAND DATASET

Link: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

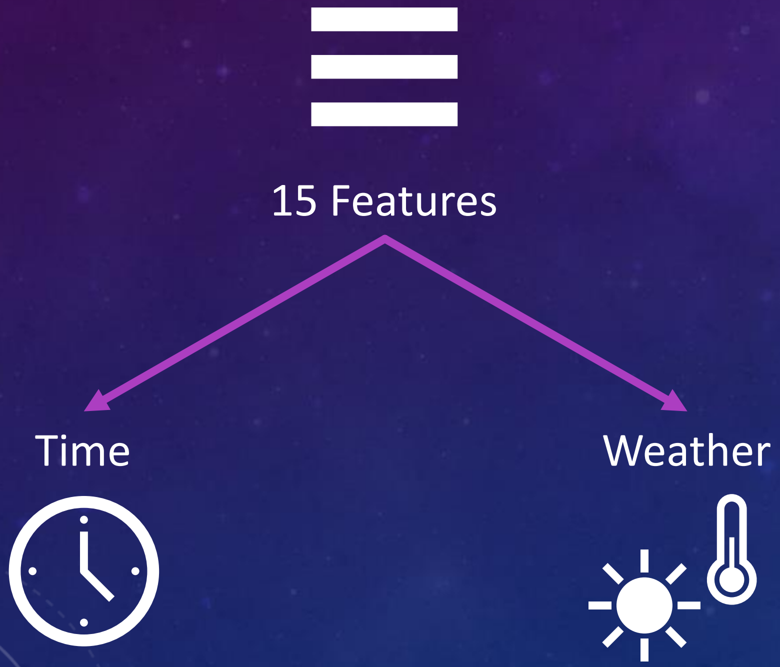
SEOUL BIKES: PRESENTATION

According to the responsible of this dataset, the dataset contains count of public bikes rented at each hour in Seoul Bike haring System with the corresponding Weather data and Holidays information

"Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information."

SEOUL BIKES: PROBLEM

Supervised → Regression Problem


15 Features

Time

Weather

Prediction of bike count required at each hour for the stable supply of rental bikes:

Target

Rented Bikes count

SEOUL BIKES: STEPS FOLLOWED FOR THIS WORK

- Data Collection
- First Exploratory Data Analysis
- Data Preprocessing
    - Missing Values
    - Null & Zero Values
    - Outliers
    - Categorical Features
- Second Exploratory Data Analysis
    - Univariate Analysis
        - Time
        - Weather
    - Bivariate Analysis
    - Correlation
- Data Processing & Feature Engineering
- Models Fitting
    - Model Selection: Training and Cross Validation
    - Hyperparameters for the selected models

SEOUL BIKES: DATA PREPROCESSING - SUMMARY

No NaN
&
almost no Null or
Zero values

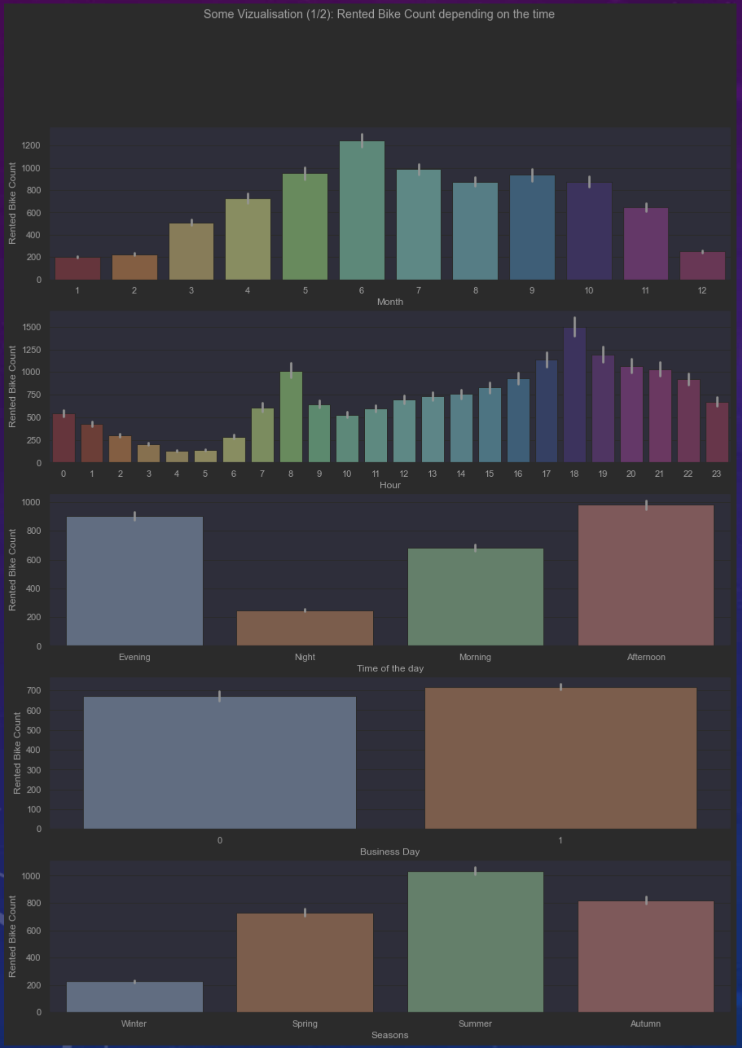
Some features have outliers but caused by the random variation of the weather

Three categorical features:

- Seasons
- Holiday
- Functionning Day

→ No manipulation compared to the previous dataset

SEOUL BIKES: EDA – SUMMARY



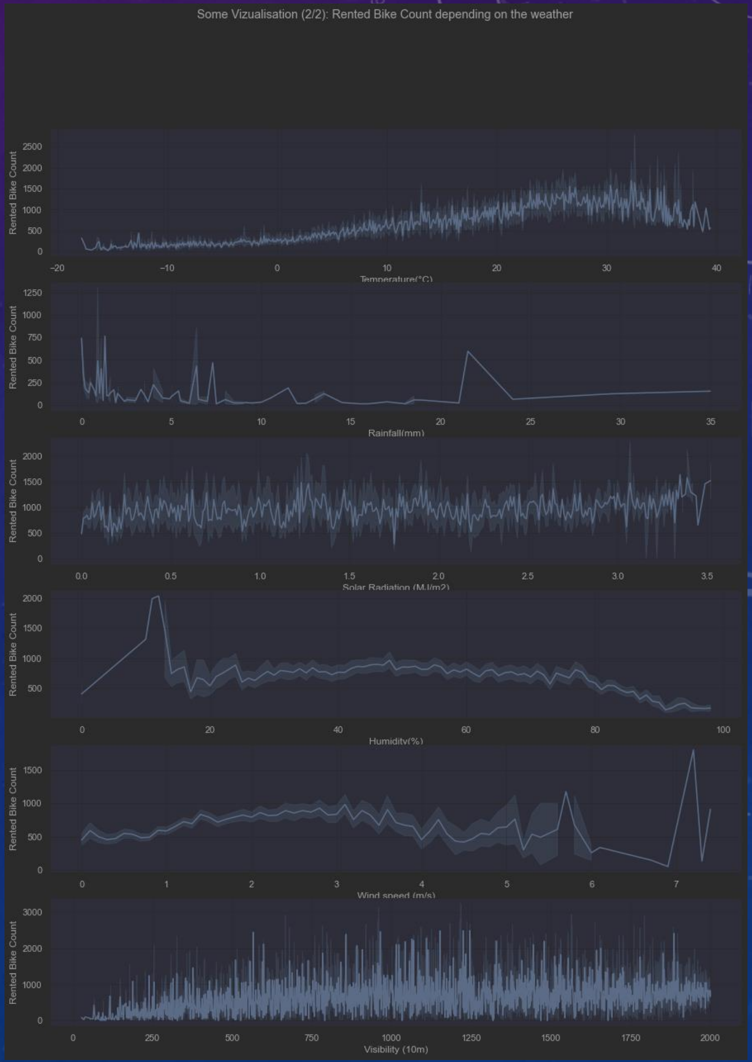
Univariate Analysis

Time
Month/Season and Hour
have an impact on the target

Weather
Temperature, visibility
and solar radiation have an
impact on the target for
the weather features

Bivariate Analysis

→ Dew point temperature has a big correlation with the temperature and the humidity



SEOUL BIKES: MAIN RESULTS - TRAINING

Feature Selection:

→ Temperature, Hour, Seasons, Solar Radiation, Visibility, Month, Wind speed, Business Day

Models Tested:

→ Kfold Cross-Validation

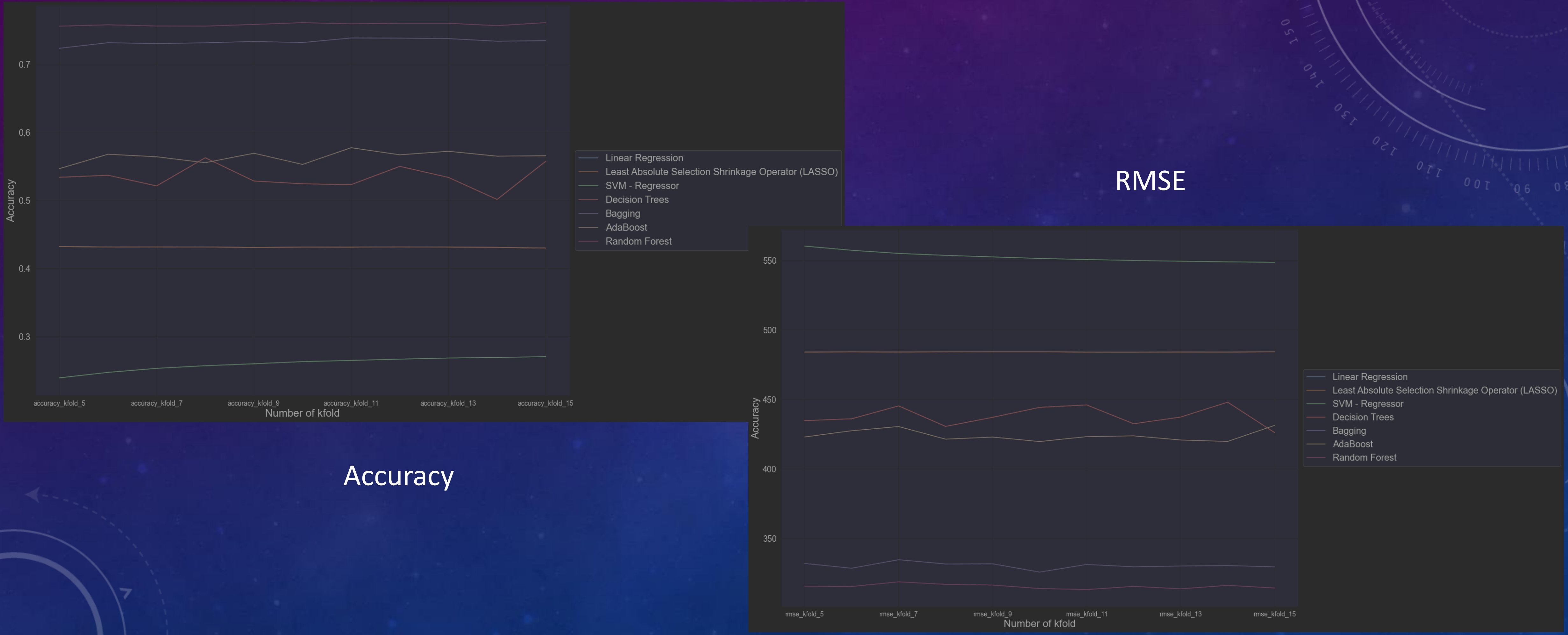
→ Linear Regression, Lasso, SVR, Decision Tree, Bagging, AdaBoost, Random Forest

Models for hyperparameters tunning:

I took the best model from the step training, the middle one and an other which had bad performances

→ Random Forest, AdaBoost, Lasso

SEOUL BIKES: MAIN RESULTS - METRICS



RMSE

Accuracy

SEOUL BIKES: MAIN RESULTS - HYPERPARAMETERS

Random Forest

- Initial Score: 0.7960
- Hyperparameters: {'criterion': 'squared_error', 'max_depth': 100, 'n_estimators': 1000}
- Final Score: 0.7961

AdaBoost

- Initial Score: 0.5976
- Hyperparameters: {‘learning_rate’: 0.5, 'n_estimators': 100}
- Final Score: 0.6210

Lasso

- Initial Score: 0.4506
- Hyperparameters: {‘alpha’: 0.1}
- Final Score: 0.4508

SEOUL BIKES: DISCUSSION & CONCLUSION

The second dataset was a regression problem.

I had the opportunity to work on another models we haven't seen in class. One of the advantages of this dataset was that it was pretty easy to do some Exploratory Data Analysis (EDA) compared to the other dataset, as much as to do some univariate analysis.

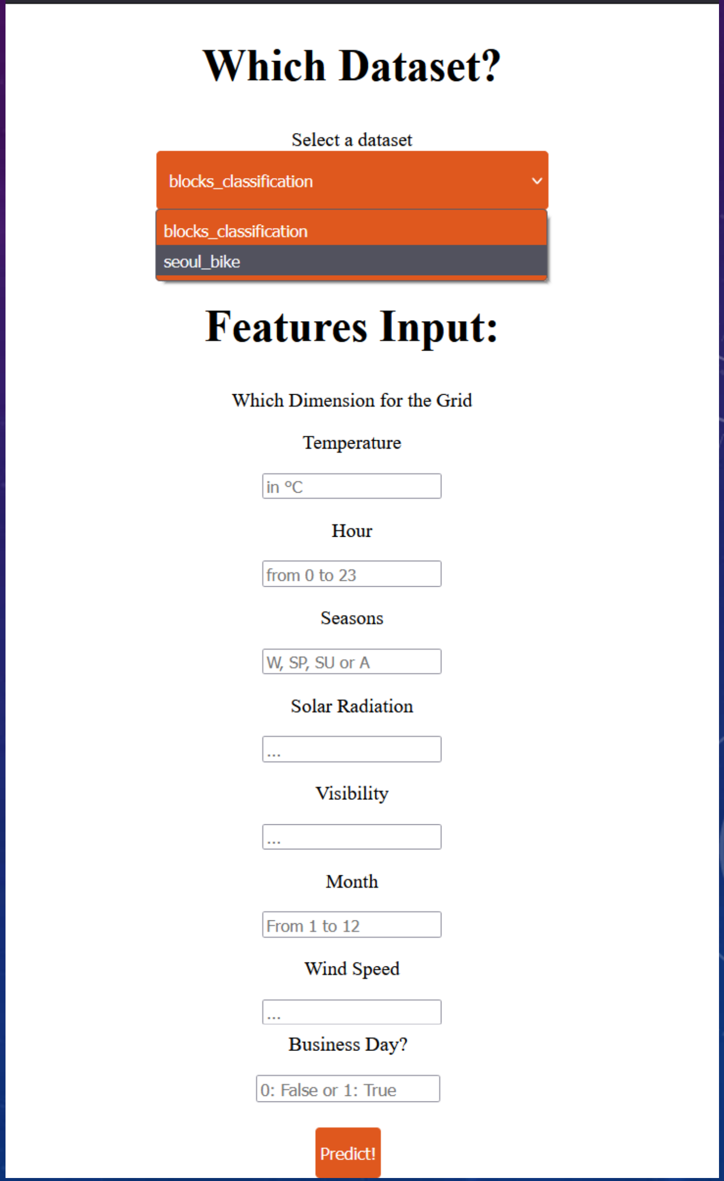
FLASK API IMPLEMENTED

An API has been implemented.

In fact, just the beginning of the API has been pushed. Currently, it is possible to select the dataset either the blocks classification one or the Seoul bike one. And then, the user can add a value for each significant feature find from the work which has been done on the notebooks. When all the forms are filled, we can push the button ‘’Predict!’’.

Normally, when this button is pushed, a prediction of either the class of the block or the number of rented bike, is made and appeared. This major functionnality has not been implemented yet (lack of time to do it properly).

Moreover, the design is too simple, it can be more worked on.



Which Dataset?

Select a dataset: blocks_classification

Features Input:

Which Dimension for the Grid

Temperature: in °C

Hour: from 0 to 23

Seasons: W, SP, SU or A

Solar Radiation: ...

Visibility: ...

Month: From 1 to 12

Wind Speed: ...

Business Day?: 0: False or 1: True

CONCLUSION: DISCUSSION ON THE PROJECT

For this project I wanted to use all we have seen during the practical works of this course.

I was able to work both on classification and regression problems, thus this project was complete with different aspects that we can run into from a supervised learning problem.

This project also pushes me to interpret each analysis I have made. It is a good thing to now how to solve problem in a practical manner but the interpretations are a big part of the job too. Even if we hadn’t seen interpretations that much during the course, I used my knowledge from another course I had this semester.

CONCLUSION: PERPECTIVES

- For the data preprocessing for each dataset, it is possible to do more than it was made for this project. First, the work on the outliers management could have been better. There are plenty of methods to treat them, not only the interquartile rule, and we can also treat them and not delete all of them (like using the median and so on). For the imbalanced classes, there are also some method that are better than taking randomly a define fraction of a class which is the more represented.

- More Data Visualizations could have been made. For this part, R is maybe more adequate but Python did the job with seaborn. I still have a lot to learn on this package.

- For the feature selection, I used only the correlation matrix. It is also possible to use the Principle Component Analysis (PCA). It is a good way to get the significant component and to support the interpretation of the previous method.

- Finally, the hyperparameters for each model can be improve considerably. I tried to start some hyperparameters running with large choice for different parameters of each model but I couldn't get anything in a good amount of time.

All the points describe above will be implemented after the evaluation of this project and the grade given.