

4-th year of Engineering School

PYTHON FOR DATA ANALYSIS PROJECT

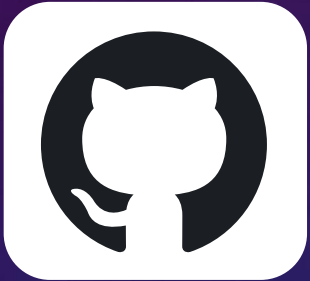
BY ESTELLE FIKET

01/05/2022

INTRODUCTION

For this project, I worked alone.

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



Create a Github
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 libraries:

- 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




10 Features

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



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

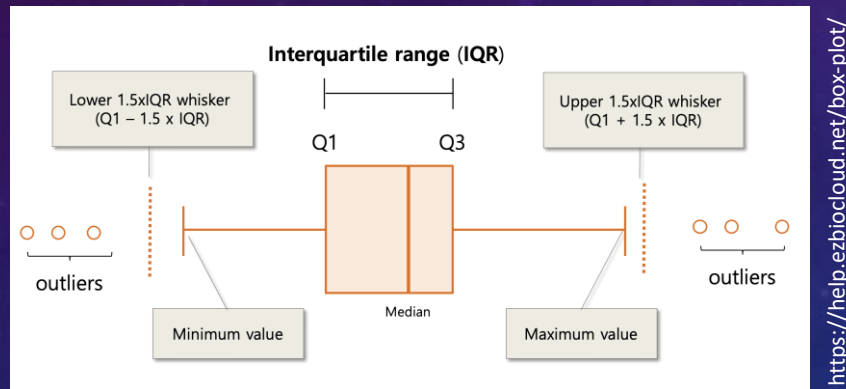


No categorical
features

BLOCKS CLASSIFICATION: DATA PREPROCESSING (2/2)

Outliers

→ Interquartile method:



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.

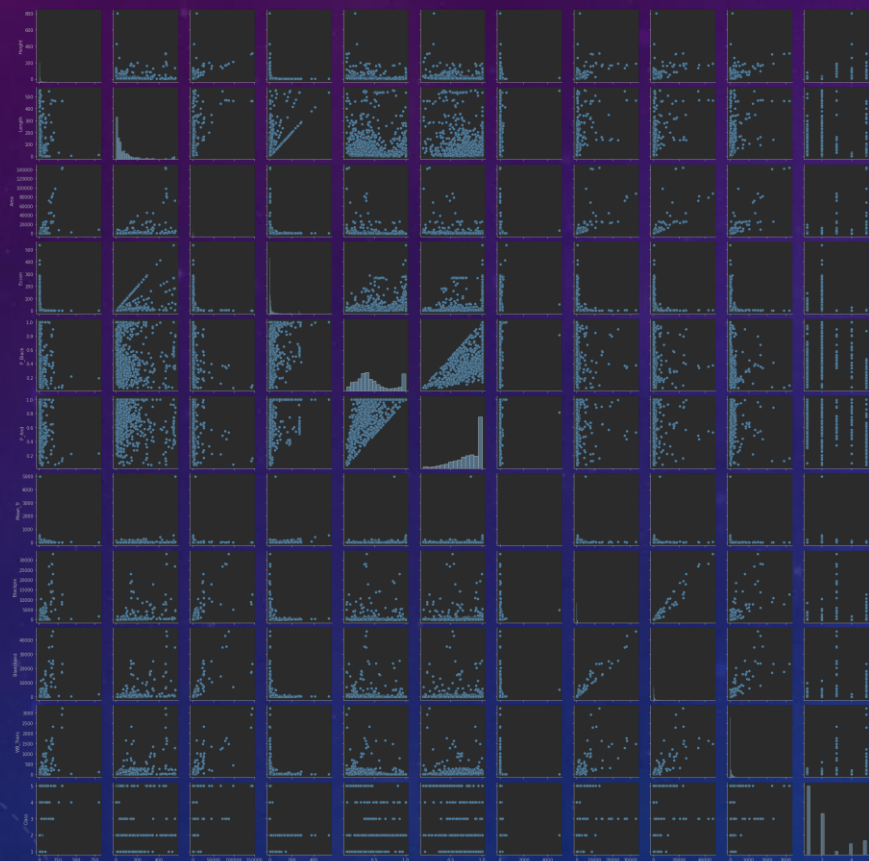
1	4394
2	329
5	115
4	88
3	28

→

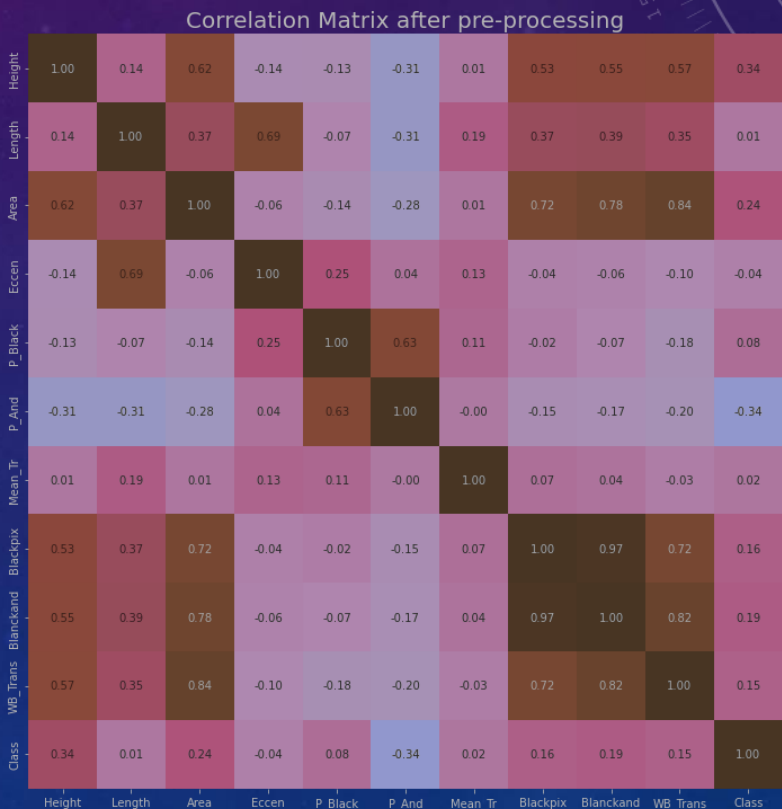
After outliers removing from the class 1, I treated the imbalanced classes problem

1	549
2	329
5	115
4	88
3	28

BLOCKS CLASSIFICATION: EDA – SUMMARY



Bivariate Analysis



Correlation Matrix

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 tuning:

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 does not impact the model (but another parameter impacts the scores, check the conclusion (2/2))

Design Document Template

Version X.X - 01 December 2016

Company Name - Address - Telephone - Email - www.website.com

© [year] [Company Name] or a [Company Name] affiliate company. All rights reserved.

[Company Name]

[Phone/Fax/Website]

Trademark

[Product Name] are registered trademarks of [Company]. All other trademarks or registered trademarks are the property of their respective owners.

Disclaimer

The information provided in this document is provided "as is" without warranty of any kind. [Company Name] disclaims all warranties, either expressed or implied, including the warranty of merchantability and fitness for a particular purpose. [Company Name] is liable for any damages, including direct, indirect, incidental, consequential, loss of business or other special damages, even if [Company Name] or its suppliers have been advised of the possibility of such damages.

Document History

[Company Name] may occasionally update online documentation to reflect changes in the related software. Consequently, if the document state not documented recently, it may not contain the most up-to-date information. Please refer to [www.\[website\].com](#) for the most current information.

Please refer to the table of contents for more information on the document if the document is updated, as indicated by changes to the table of contents.

Other legal help

[Company Name] support, product, and training information is obtained as follows:

Product information— Documentation, release notes, software updates, and other information [Company Name] products, services, training, and services are at [Company Name] website at [http://www.\[website\].com](#)

Technical support— [http://www.\[website\].com](#) and [http://www.\[website\].com](#) Support. Online support pages, you can get technical support, including product, training, and services request. Submit a request online or request, you must have a valid support agreement.

Your comments

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user interface. Please send your comments to the document to:

[DocumentName@Company Name.com](#)

If you have issues, comments, or questions about specific information or content, please include the file name, location, the page number, the version, the page number, and any other details that will help us locate the content that you are addressing.

[Company Name]

[Product Name]

[Company Name]

[Product Name]

Preface

Style Conventions

The following style conventions are used in this document:

Table

Names of commands, options, programs, processes, services, and utilities

Names of interface elements (such as windows, dialog boxes, menus, fields, and menus)

Interface elements (such as windows, dialog boxes, menus, fields, and menus)

Table

Publication date released in use

Example (for example a new name)

Table

System output, such as an error message or output

URLs, sample paths, filenames, programs, and options

Table

Variables in command line

User input variables

Example: `arg1` indicates an argument or variable value supplied by the user

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

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 problematic in professional real life project, so it was pretty interesting to discovered 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 biased models because of the first class which was the most important in this dataset.

Table of Contents

1 Introduction	1
1.1 Purpose	1
1.2 Background	1
1.3 Scope	1
1.4 Methodology	1
1.5 Evaluation Criteria	1
2 Executive Summary	2
2.1 Purpose of this document	2
2.2 Scope of this document	2
2.3 Methodology, Tools, and Techniques	2
2.4 Results and Discussion	2
2.5 Conclusion	2
3 Detailed Design	3
3.1 Introduction	3
3.2 System Architecture	3
3.3 Data Design	3
3.4 User Interface Design	3
3.5 Implementation Details	3
3.6 Testing and Validation	3
3.7 Deployment and Maintenance	3
3.8 Appendix	3
3.9 Glossary	3
3.10 Bibliography	3
3.11 Index	3
3.12 Revision History	3
3.13 Change Log	3
3.14 Document History	3
3.15 Document Information	3
3.16 Document Classification	3
3.17 Document Status	3
3.18 Document Version	3
3.19 Document Release	3
3.20 Document Approval	3
3.21 Document Review	3
3.22 Document Sign-off	3
3.23 Document Closure	3
3.24 Document Archiving	3
3.25 Document Retention	3
3.26 Document Disposal	3
3.27 Document Security	3
3.28 Document Access	3
3.29 Document Control	3
3.30 Document Management	3
3.31 Document Governance	3
3.32 Document Compliance	3
3.33 Document Audit	3
3.34 Document Monitoring	3
3.35 Document Reporting	3
3.36 Document Communication	3
3.37 Document Collaboration	3
3.38 Document Integration	3
3.39 Document Interoperability	3
3.40 Document Compatibility	3
3.41 Document Portability	3
3.42 Document Scalability	3
3.43 Document Flexibility	3
3.44 Document Adaptability	3
3.45 Document Resilience	3
3.46 Document Robustness	3
3.47 Document Reliability	3
3.48 Document Availability	3
3.49 Document Usability	3
3.50 Document Accessibility	3
3.51 Document Inclusion	3
3.52 Document Exclusion	3
3.53 Document Inheritance	3
3.54 Document Composition	3
3.55 Document Aggregation	3
3.56 Document Association	3
3.57 Document Relationship	3
3.58 Document Interaction	3
3.59 Document Communication	3
3.60 Document Collaboration	3
3.61 Document Integration	3
3.62 Document Interoperability	3
3.63 Document Compatibility	3
3.64 Document Portability	3
3.65 Document Scalability	3
3.66 Document Flexibility	3
3.67 Document Adaptability	3
3.68 Document Resilience	3
3.69 Document Robustness	3
3.70 Document Reliability	3
3.71 Document Availability	3
3.72 Document Usability	3
3.73 Document Accessibility	3
3.74 Document Inclusion	3
3.75 Document Exclusion	3
3.76 Document Inheritance	3
3.77 Document Composition	3
3.78 Document Aggregation	3
3.79 Document Association	3
3.80 Document Relationship	3
3.81 Document Interaction	3
3.82 Document Communication	3
3.83 Document Collaboration	3
3.84 Document Integration	3
3.85 Document Interoperability	3
3.86 Document Compatibility	3
3.87 Document Portability	3
3.88 Document Scalability	3
3.89 Document Flexibility	3
3.90 Document Adaptability	3
3.91 Document Resilience	3
3.92 Document Robustness	3
3.93 Document Reliability	3
3.94 Document Availability	3
3.95 Document Usability	3
3.96 Document Accessibility	3
3.97 Document Inclusion	3
3.98 Document Exclusion	3
3.99 Document Inheritance	3
3.100 Document Composition	3

1.5 Evaluation Criteria

Define the criteria used to evaluate software systems, such as organizational objectives, increased efficiency, and reduced operating costs.

2 Executive Summary

Provide a brief introduction to the system for which this design is being developed.

2.1 Purpose of this document

1 Introduction

Code Block Design is used to analyze and evaluate, from a cost and benefit perspective, potential solutions to a problem or opportunity. It also describes alternatives, benefits and risks, and the results of the analysis.

Note: A feasibility study may be required to verify the feasibility of the proposed solution. The level and complexity of the study will vary depending on the nature of the problem.

The Code Block Design process involves the following steps: 1. Define the problem. 2. Identify the requirements. 3. Analyze the requirements. 4. Develop the solution. 5. Implement the solution. 6. Evaluate the solution. 7. Maintain the solution.

1.1 Purpose

Introduce the business need that the Code Block Design intends to address. You may also want to report on the results of the business analysis that motivated the [Organization] to consider this solution. For example, the need to increase revenue, reduce costs, or improve the quality of the solution. The need to increase revenue, reduce costs, or improve the quality of the solution. The need to increase revenue, reduce costs, or improve the quality of the solution.

1.2 Background

Provide background information that places the Code Block Design in context. For example, previous decisions or projects that are relevant to understanding the current solution.

1.3 Scope

Define the scope of the Code Block Design. Make sure to highlight areas that were not included in the analysis and explain the reasons for these omissions. For example, budgetary constraints.

1.4 Methodology

Describe the methodology used to conduct the Code Block Design and how it aligns with the Software Development Life Cycle and other relevant standards. Also, describe the tools and techniques used to conduct the analysis. For example, the use of modeling tools, such as UML, to represent the system. Also, describe the tools and techniques used to conduct the analysis. For example, the use of modeling tools, such as UML, to represent the system.

2.5 Methodology, Tools, and Techniques

Describe the software tools (or techniques) required for performing design documents tasks.

Design Document Template

Version X.X - 01 December 2016

Company Name • Address • Telephone • Email • www.website.com

© [year] [Company Name] or a [Company Name] affiliate company. All rights reserved.

[Company]([address])

[Phone]([phone])

Table of Contents

[Product Name] are registered trademarks of [Company]. All other trademarks or registered trademarks are the property of their respective owners.

Disclaimer

The information provided in this document is provided "as is" without warranty of any kind. [Company Name] disclaims all warranties, either expressed or implied, including the accuracy of completeness and timeliness of the information. [Company Name] is liable for any damages arising from the use of this information, including direct, indirect, incidental, consequential, loss of business or other special damages, even if [Company Name] or its suppliers have been advised of the possibility of such damages.

Document History

[Company Name] may occasionally update online documentation to reflect changes to the product software. Consequently, if this document was not downloaded recently, it may not contain the most up-to-date information. Please refer to [www.\[company\].com](#) for the most current information.

Please refer to the table of contents for information on the document if it has been updated, as indicated by changes to the title (date).

Online legal help

[Company Name] support, product, and licensing information is obtained on [Company Name] website.

Product information—Documentation, release notes, software updates, and other information [Company Name] products, services, training, and services are at [Company Name] website at [www.\[company\].com](#).

Technical support—[Company Name] ([address]) and [Company Name] ([address]) support pages, you can get technical support, including product, training, and services request. [Company Name] website request, you must have a valid support agreement.

Your comments

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user interface. Please send your comments to the document to:

[\[Company Name\] \(\[address\]\)](#)

If you have issues, comments, or questions about specific information or content, please include the file name, location, the page number, the version, the page number, and any other details that will help us locate the content that you are addressing.

[Company Name]
[product name]

Table of Contents

1 Introduction	1
1.1 Purpose	1
1.2 Background	1
1.3 Scope	1
1.4 Methodology	1
1.5 Evaluation Criteria	1
2 Executive Summary	2
2.1 Purpose of this document	2
2.2 Background	2
2.3 Scope	2
2.4 Methodology	2
2.5 Evaluation Criteria	2
3 Design Overview	3
3.1 Background	3
3.2 Scope	3
3.3 Methodology	3
3.4 Evaluation Criteria	3
3.5 Design Overview	3
3.6 Design Overview	3
3.7 Design Overview	3
3.8 Design Overview	3
4 System Architecture	4
4.1 Background	4
4.2 Scope	4
4.3 Methodology	4
4.4 Evaluation Criteria	4
4.5 System Architecture	4
4.6 System Architecture	4
4.7 System Architecture	4
4.8 System Architecture	4
4.9 System Architecture	4
4.10 System Architecture	4
4.11 System Architecture	4
4.12 System Architecture	4
4.13 System Architecture	4
4.14 System Architecture	4
4.15 System Architecture	4
4.16 System Architecture	4
4.17 System Architecture	4
4.18 System Architecture	4
4.19 System Architecture	4
4.20 System Architecture	4
4.21 System Architecture	4
4.22 System Architecture	4
4.23 System Architecture	4
4.24 System Architecture	4
4.25 System Architecture	4
4.26 System Architecture	4
4.27 System Architecture	4
4.28 System Architecture	4
4.29 System Architecture	4
4.30 System Architecture	4
4.31 System Architecture	4
4.32 System Architecture	4
4.33 System Architecture	4
4.34 System Architecture	4
4.35 System Architecture	4
4.36 System Architecture	4
4.37 System Architecture	4
4.38 System Architecture	4
4.39 System Architecture	4
4.40 System Architecture	4
4.41 System Architecture	4
4.42 System Architecture	4
4.43 System Architecture	4
4.44 System Architecture	4
4.45 System Architecture	4
4.46 System Architecture	4
4.47 System Architecture	4
4.48 System Architecture	4
4.49 System Architecture	4
4.50 System Architecture	4
4.51 System Architecture	4
4.52 System Architecture	4
4.53 System Architecture	4
4.54 System Architecture	4
4.55 System Architecture	4
4.56 System Architecture	4
4.57 System Architecture	4
4.58 System Architecture	4
4.59 System Architecture	4
4.60 System Architecture	4
4.61 System Architecture	4
4.62 System Architecture	4
4.63 System Architecture	4
4.64 System Architecture	4
4.65 System Architecture	4
4.66 System Architecture	4
4.67 System Architecture	4
4.68 System Architecture	4
4.69 System Architecture	4
4.70 System Architecture	4
4.71 System Architecture	4
4.72 System Architecture	4
4.73 System Architecture	4
4.74 System Architecture	4
4.75 System Architecture	4
4.76 System Architecture	4
4.77 System Architecture	4
4.78 System Architecture	4
4.79 System Architecture	4
4.80 System Architecture	4
4.81 System Architecture	4
4.82 System Architecture	4
4.83 System Architecture	4
4.84 System Architecture	4
4.85 System Architecture	4
4.86 System Architecture	4
4.87 System Architecture	4
4.88 System Architecture	4
4.89 System Architecture	4
4.90 System Architecture	4
4.91 System Architecture	4
4.92 System Architecture	4
4.93 System Architecture	4
4.94 System Architecture	4
4.95 System Architecture	4
4.96 System Architecture	4
4.97 System Architecture	4
4.98 System Architecture	4
4.99 System Architecture	4
4.100 System Architecture	4

Interpretation of the document

Page number

[Company Name]
[product name]

Table of Contents

1 Introduction	1
1.1 Purpose	1
1.2 Background	1
1.3 Scope	1
1.4 Methodology	1
1.5 Evaluation Criteria	1
2 Executive Summary	2
2.1 Purpose of this document	2
2.2 Background	2
2.3 Scope	2
2.4 Methodology	2
2.5 Evaluation Criteria	2
3 Design Overview	3
3.1 Background	3
3.2 Scope	3
3.3 Methodology	3
3.4 Evaluation Criteria	3
3.5 Design Overview	3
3.6 Design Overview	3
3.7 Design Overview	3
3.8 Design Overview	3
3.9 Design Overview	3
3.10 Design Overview	3
3.11 Design Overview	3
3.12 Design Overview	3
3.13 Design Overview	3
3.14 Design Overview	3
3.15 Design Overview	3
3.16 Design Overview	3
3.17 Design Overview	3
3.18 Design Overview	3
3.19 Design Overview	3
3.20 Design Overview	3
3.21 Design Overview	3
3.22 Design Overview	3
3.23 Design Overview	3
3.24 Design Overview	3
3.25 Design Overview	3
3.26 Design Overview	3
3.27 Design Overview	3
3.28 Design Overview	3
3.29 Design Overview	3
3.30 Design Overview	3
3.31 Design Overview	3
3.32 Design Overview	3
3.33 Design Overview	3
3.34 Design Overview	3
3.35 Design Overview	3
3.36 Design Overview	3
3.37 Design Overview	3
3.38 Design Overview	3
3.39 Design Overview	3
3.40 Design Overview	3
3.41 Design Overview	3
3.42 Design Overview	3
3.43 Design Overview	3
3.44 Design Overview	3
3.45 Design Overview	3
3.46 Design Overview	3
3.47 Design Overview	3
3.48 Design Overview	3
3.49 Design Overview	3
3.50 Design Overview	3
3.51 Design Overview	3
3.52 Design Overview	3
3.53 Design Overview	3
3.54 Design Overview	3
3.55 Design Overview	3
3.56 Design Overview	3
3.57 Design Overview	3
3.58 Design Overview	3
3.59 Design Overview	3
3.60 Design Overview	3
3.61 Design Overview	3
3.62 Design Overview	3
3.63 Design Overview	3
3.64 Design Overview	3
3.65 Design Overview	3
3.66 Design Overview	3
3.67 Design Overview	3
3.68 Design Overview	3
3.69 Design Overview	3
3.70 Design Overview	3
3.71 Design Overview	3
3.72 Design Overview	3
3.73 Design Overview	3
3.74 Design Overview	3
3.75 Design Overview	3
3.76 Design Overview	3
3.77 Design Overview	3
3.78 Design Overview	3
3.79 Design Overview	3
3.80 Design Overview	3
3.81 Design Overview	3
3.82 Design Overview	3
3.83 Design Overview	3
3.84 Design Overview	3
3.85 Design Overview	3
3.86 Design Overview	3
3.87 Design Overview	3
3.88 Design Overview	3
3.89 Design Overview	3
3.90 Design Overview	3
3.91 Design Overview	3
3.92 Design Overview	3
3.93 Design Overview	3
3.94 Design Overview	3
3.95 Design Overview	3
3.96 Design Overview	3
3.97 Design Overview	3
3.98 Design Overview	3
3.99 Design Overview	3
3.100 Design Overview	3

Interpretation of the document

Page number

[Company Name]
[product name]

Table of Contents

1 Introduction	1
1.1 Purpose	1
1.2 Background	1
1.3 Scope	1
1.4 Methodology	1
1.5 Evaluation Criteria	1
2 Executive Summary	2
2.1 Purpose of this document	2
2.2 Background	2
2.3 Scope	2
2.4 Methodology	2
2.5 Evaluation Criteria	2
3 Design Overview	3
3.1 Background	3
3.2 Scope	3
3.3 Methodology	3
3.4 Evaluation Criteria	3
3.5 Design Overview	3
3.6 Design Overview	3
3.7 Design Overview	3
3.8 Design Overview	3
3.9 Design Overview	3
3.10 Design Overview	3
3.11 Design Overview	3
3.12 Design Overview	3
3.13 Design Overview	3
3.14 Design Overview	3
3.15 Design Overview	3
3.16 Design Overview	3
3.17 Design Overview	3
3.18 Design Overview	3
3.19 Design Overview	3
3.20 Design Overview	3
3.21 Design Overview	3
3.22 Design Overview	3
3.23 Design Overview	3
3.24 Design Overview	3
3.25 Design Overview	3
3.26 Design Overview	3
3.27 Design Overview	3
3.28 Design Overview	3
3.29 Design Overview	3
3.30 Design Overview	3
3.31 Design Overview	3
3.32 Design Overview	3
3.33 Design Overview	3
3.34 Design Overview	3
3.35 Design Overview	3
3.36 Design Overview	3
3.37 Design Overview	3
3.38 Design Overview	3
3.39 Design Overview	3
3.40 Design Overview	3
3.41 Design Overview	3
3.42 Design Overview	3
3.43 Design Overview	3
3.44 Design Overview	3
3.45 Design Overview	3
3.46 Design Overview	3
3.47 Design Overview	3
3.48 Design Overview	3
3.49 Design Overview	3
3.50 Design Overview	3
3.51 Design Overview	3
3.52 Design Overview	3
3.53 Design Overview	3
3.54 Design Overview	3
3.55 Design Overview	3
3.56 Design Overview	3
3.57 Design Overview	3
3.58 Design Overview	3
3.59 Design Overview	3
3.60 Design Overview	3
3.61 Design Overview	3
3.62 Design Overview	3
3.63 Design Overview	3
3.64 Design Overview	3
3.65 Design Overview	3
3.66 Design Overview	3
3.67 Design Overview	3
3.68 Design Overview	3
3.69 Design Overview	3
3.70 Design Overview	3
3.71 Design Overview	3
3.72 Design Overview	3
3.73 Design Overview	3
3.74 Design Overview	3
3.75 Design Overview	3
3.76 Design Overview	3
3.77 Design Overview	3
3.78 Design Overview	3
3.79 Design Overview	3
3.80 Design Overview	3
3.81 Design Overview	3
3.82 Design Overview	3
3.83 Design Overview	3
3.84 Design Overview	3
3.85 Design Overview	3
3.86 Design Overview	3
3.87 Design Overview	3
3.88 Design Overview	3
3.89 Design Overview	3
3.90 Design Overview	3
3.91 Design Overview	3
3.92 Design Overview	3
3.93 Design Overview	3
3.94 Design Overview	3
3.95 Design Overview	3
3.96 Design Overview	3
3.97 Design Overview	3
3.98 Design Overview	3
3.99 Design Overview	3
3.100 Design Overview	3

Interpretation of the document

Page number

[Company Name]
[product name]

Table of Contents

Define the criteria used to evaluate software systems, such as organizational objectives, increased efficiency, and reduced operating costs.

[Company Name]
[product name]

Table of Contents

Provide a brief introduction to the system for which this design is being developed.

2.1 Purpose of this document

[Company Name]
[product name]

Table of Contents

Define the software tools for development required for performing design documents tasks.

2.5 Methodology, Tools, and Techniques

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

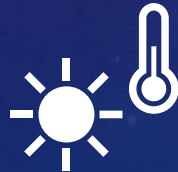


15 Features

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
- Functioning Day

→ No manipulation compared to the previous dataset

SEOUL BIKES: EDA – SUMMARY



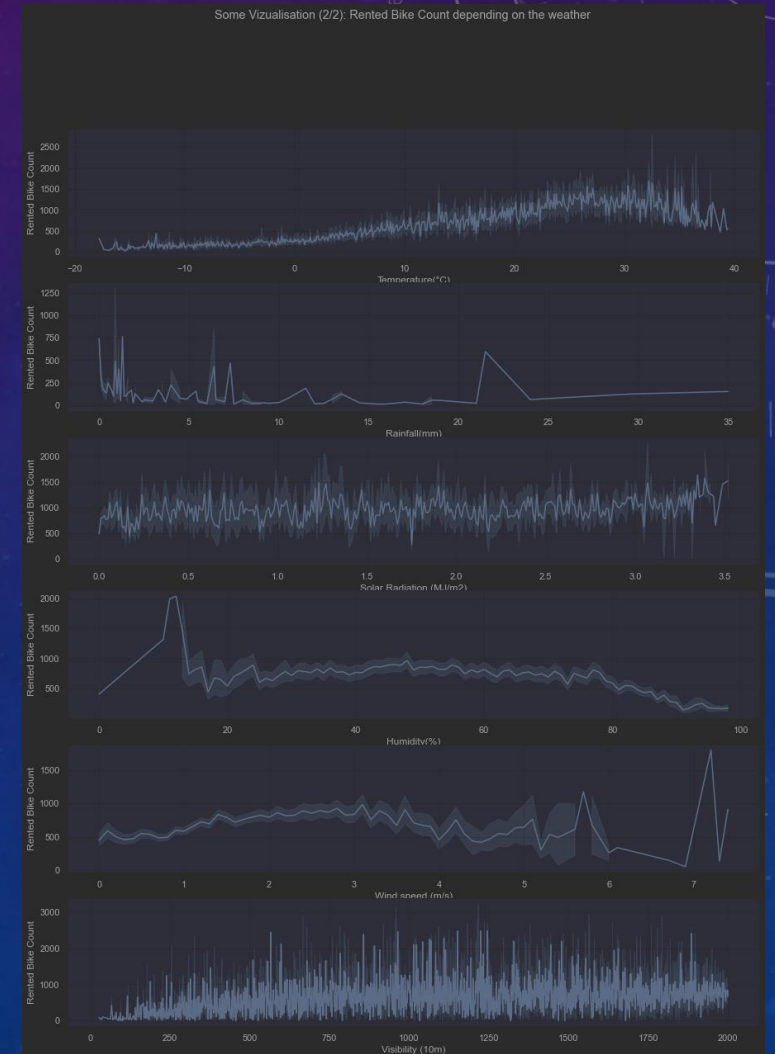
Univariate Analysis

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

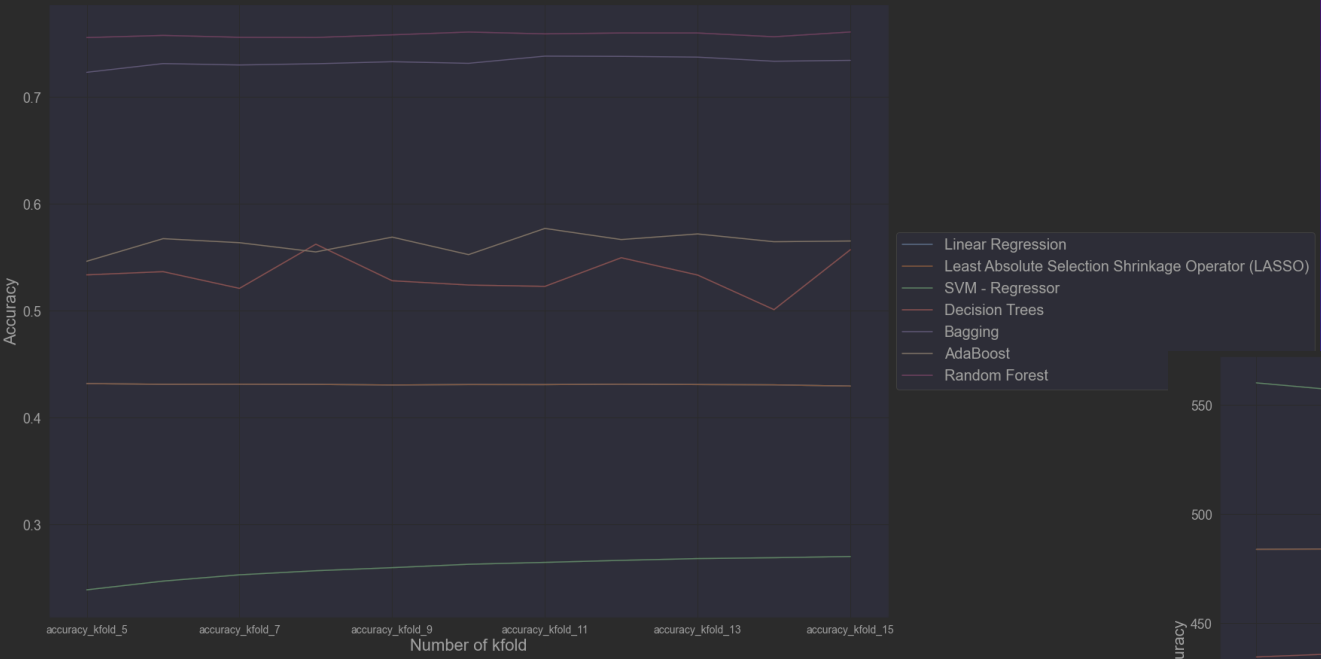
→ Linear Regression, Lasso, SVR, Decision Tree, Bagging, AdaBoost, Random Forest

Models for hyperparameters tuning:

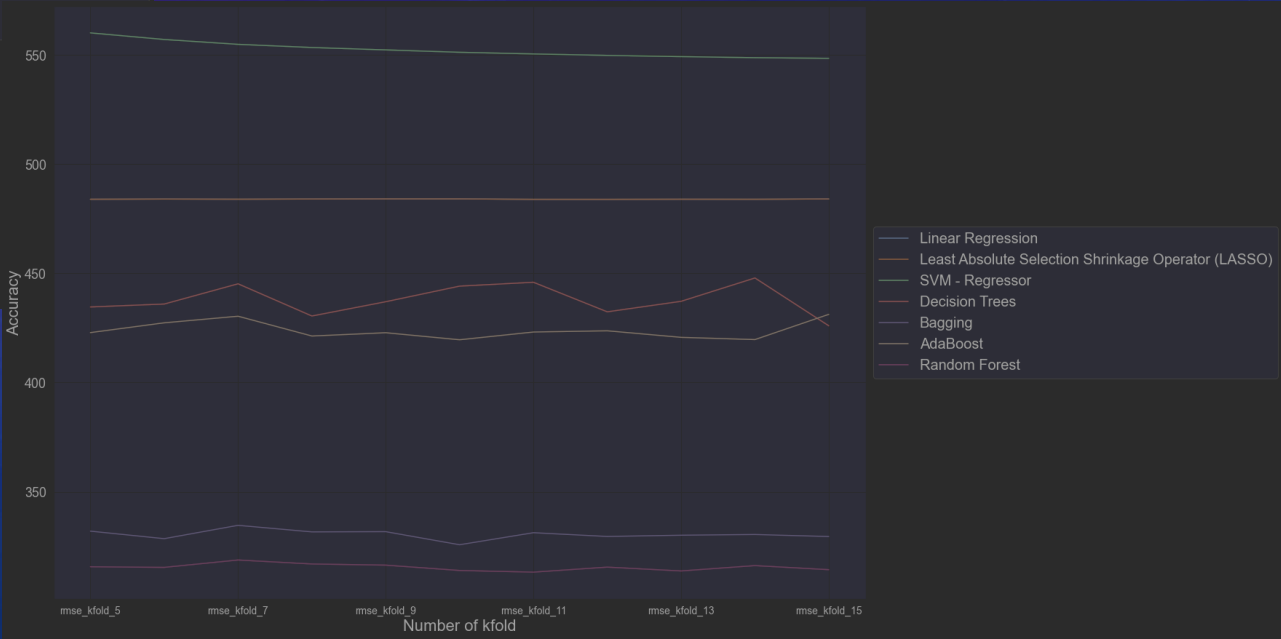
I took the best model from the step training, the middle one and an other which had bad performances

→ Random Forest, AdaBoost, Lasso

SEOUL BIKES: MAIN RESULTS - METRICS



Accuracy



RMSE

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 fonctionnality 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

blocks_classification

seoul_bike

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

Predict!

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: PERSPECTIVES

- 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.