



ECOLE INTERNATIONALE
DES SCIENCES DU
TRAITEMENT DE L'INFORMATION

EISTI

ADEO1 PROJECT

Geographical statistics of engineering internships

Authors:

Gustavo FLEURY
Cylia BERKANE
Fagnuo CAI
INDURAJ PR
Quoc Viet PHAM
Yen Chu CHEN

Advisers:

Esma TALHI
Rachid CHELOUAH

Cergy - June 3, 2019

Contents

I	INTRODUCTION	4
I.1	Background of the Project	4
I.2	About EISTI	4
I.3	General Objectives	4
I.3.1	Specific Objective of the Project	4
I.4	Scope of the project	4
II	TECHNICAL REPORT	6
II.1	Specifications	6
II.2	Functional and detailed specifications	6
II.3	TECHNICAL SOLUTION	6
II.3.1	SQLite	7
II.3.2	Python	7
II.3.3	Django	8
II.3.4	Chart.js	8
II.3.5	Bootstrap	8
II.3.6	Responsive	8
II.3.7	Libraries and Essential Tools	8
II.4	Overleaf	9
III	PROJECT PLANNING REPORT	10
III.1	Project management tools	10
III.2	Organization of the team and cutting into tasks	10
III.3	Initial Gantt charts	11
III.4	Update of the Status of the project in each sprint	13
III.5	Conclusion	13
IV	REALIZATION	14
IV.1	Data Source	14
IV.1.1	PRG_STUDENT_SITE_2017_2018_DATA_TABLE	14
IV.1.2	ADR_STUDENTS_2017_2018_DATA_TABLE	14
IV.1.3	STUDENT_INTERNSHIP_2013_2018_DATA_TABLE	14
IV.2	Data Model	15
IV.3	Implementation	15
V	MACHINE LEARNING SOLUTION	17
V.1	ETL - Extract Transform and Load	17
V.2	Forecast Predict - Random Forest	18
V.3	Descriptive statistics	20
V.3.1	Filters	20
V.3.2	Descriptive Statistics	20
V.3.3	Heatmap	22
V.4	Geographical statistic of engineering internships	24
VI	GENERAL CONCLUSION AND PERSPECTIVES	25
	Appendices	26
A	USER INTERFACE	26
A.1	Home	26
A.2	Login	26
A.3	Reset Password	26
A.4	Contact Us	27
A.5	Register	27

A.6	Main Page	28
A.7	CSV Files Load	29
A.8	ETL	29

List of Figures

1	Technical Solution of the project.	7
2	Example of tasks organization	10
3	Schedule and current status of the final grade project in first sprint.	11
4	Tasks for "Database 1st Version", "Application 1st Version", "Data Science Concepts", "Functional Sprints" and "Final Presentation.	12
5	Schedule of the final grade project in Sprint 2.	13
6	Schedule and current status of the final grade project in the last Sprint.	13
7	CSV files First Tables	15
8	First Data Model.	15
9	Here it's show how it work with Django	16
10	How ETL Works.	17
11	ETL - Missing Values	18
12	Random Forest	18
13	Update Weights	19
14	Forecast - Predict Internships	19
15	Forecast - Enterprise	19
16	Descriptive statistics - Filters	20
17	Descriptive statistics	21
18	Top 10 of Companies	22
19	Descriptive statistics -Heatmap	23
20	Distance between the students homes and campus	24
21	Home Screen Interface	26
22	User Login Interface	26
23	Interface for Password Reset	27
24	Contact Us Interface	27
25	Register Interface	27
26	Register Interface password restrictions	28
27	Main Page for Admin	28
28	Main Page for simple user	29
29	CSV Files Load	29
30	ETL Merged Tables	30
31	ETL	30
32	Data base Tables	31

I INTRODUCTION

Data mining is the process of analyzing data from different sources and summarizing it into relevant information that can be used to help increase revenue and decrease costs. Its primary purpose is to find correlations or patterns among dozens of fields in large databases.

This project consists in developing a solution of Geographical statistics of engineering internships, The dataset that we have been assigned is EISTI students sites, internships and addresses. Our team choose to develop a system that could help the school gain more knowledge on the geographical statistics of the engineering internships chosen by student of both Cergy and Pau campus. We will be using all the methodes seen during this year in ADEO1, Method like SIXO or Agile Method, Data Exploration method and programming languages.

I.1 Background of the Project

Every Educational institution have students going from the university to different parts of the world for internships to gain experience and explore career path. In EISTI's campus both Cergy and Pau, at least few hundred students relocate every semester to different geographical locations for the internships. But it is unclear as what's the choice of the students in choosing companies while applying for internships. It is thus crucial to analyze the internship preference of the students from the available data so as to get thorough insight on if the student preferred companies based on the distance between their home and campus or between their home and location of the company or between campus and internships location! The thorough insight on the student's preference and about the companies that hire the EISTIEN'S will help the university in many ways such as in adapting even more dynamic on-demand competitive edge curriculums, partnering with the companies for giving on-campus training sessions and seminars, etc. These insights can also be used to define a model which would predict as where the student is most likely to end up as intern!

I.2 About EISTI

EISTI - As an engineering school officially recognized by the State and made competent by the CTI (French Engineering accreditation institution) the EISTI's vocation is to trains future engineers in Mathematics and Software engineering in accordance with the technological world in which we live and its constant evolution, so as to be able to meet the needs of companies and thus offer its students a gateway to the professional world.

I.3 General Objectives

The general objective of the project is to build a tool by which we can visualize as what factor governs the most while the student decides to apply and go for the internship.

I.3.1 Specific Objective of the Project

- Study and analyze the existing data-set;
- Perform data cleaning and transformation of raw data;
- Design the front end (web based) and back-end system;
- Incorporating the front end and the back end with the database;
- Implementing the designed system.

I.4 Scope of th project

To properly define the scope of this project we must understand the needs of the client, mainly the development of a desktop/web application that will allow users study and extract relevant information from a geographical database of EISTI students.

This application will have the following functionalities :

- Upload CSV file and store it within a database. Upload CSV file with some restrictions such as the number of attributes and the column names must match the specific format;
- Automatically clean the data;
- A menu to perform univariate and bivariate analysis and statistics that are defined in advance and can be configured;
- Distance Calculation between the different locations « home - campus », « home – internships location » and « campus – internships location »;
- User management (Login/Password /roles).

II TECHNICAL REPORT

II.1 Specifications

Authentication and authorization

- CRUD operations with users
 - Define the format of the login (e-mail,Username...)
 - The password must be at least 8 characters long including a number and a non-alphanumeric characters
- CRUD operations with Roles
- Grant Roles for user (Use & extend authentication and authorization component of Django)

The upload of the csv file and the storage of the data within a database

Before the data is stored, you have to analyze it and check its integrity/consistency (for exmaple to highlight whether there is missed data, forbidden characters, ...).

Statistic functionality

- Perform univariate and bivariate analysis
- Indicate the field of each class : ing3 (bi, erp, inem, ifi, i3, ..), ing1/ing2 (gi, gm, mi, sie,gi, mf),
- Cluster the statistics for a campus (Cergy, Pau)

Distances calculation between the different locations cited above using Web APIs

II.2 Functional and detailed specifications

Admin	User
1- The upload of the csv file and the storage of the data within a database. Before the data is stored, you have to analyze it and check its integrity/consistency (for exmaple to highlight whether there is missed data, forbidden characters, ...).	
2- A menu (bar) to perform univariate and bivariate analysis that are defined in advance and can be configured. To this end, you have to calculate the distances between the different locations cited above using Web APIs. As well, remember to use existing APIs for the vizualisation of the statsitics like OpenStreetMap, GoogleMap, D3.js,... It is up to you to decide which statistics to perform and which parameters are relevant. For instance, you can choose to : a) Indicate the field of each class : ing3 (bi, erp, inem, ifi, i3, ..), ing1/ing2 (gi, gm, mi, sie,gi, mf), b) cluster the statistics for a campus (Cergy, Pau) , c) ..	2- A menu (bar) to perform univariate and bivariate analysis that are defined in advance and can be configured. To this end, you have to calculate the distances between the different locations cited above using Web APIs. As well, remember to use existing APIs for the vizualisation of the statsitics like OpenStreetMap, GoogleMap, D3.js,... It is up to you to decide which statistics to perform and which parameters are relevant. For instance, you can choose to : a) Indicate the field of each class : ing3 (bi, erp, inem, ifi, i3, ..), ing1/ing2 (gi, gm, mi, sie,gi, mf), b) cluster the statistics for a campus (Cergy, Pau) , c) ..
3- Access to the application via login credentials (count) : login/password. You have to define the format of the login (e-mail...). The password must be at least 8 characters long including a number and a non-alphanumeric characters. You have to find a way in which the admin can validate the count.	

II.3 TECHNICAL SOLUTION

There are three major developing software that we use in our project, Python, SQLite and Django. In the beginning of the research, the raw data is present in text file (CSV). Then, we import the data to python to do data cleaning and a general data profiling. After, we insert the data to the Oracle SQL. During the developing procedure, we use python again as data analysis software. At last, we choose Django do be our visualizing tool. With Django linking with SQLite, our final goal is to present the client dynamically analyzed data model.

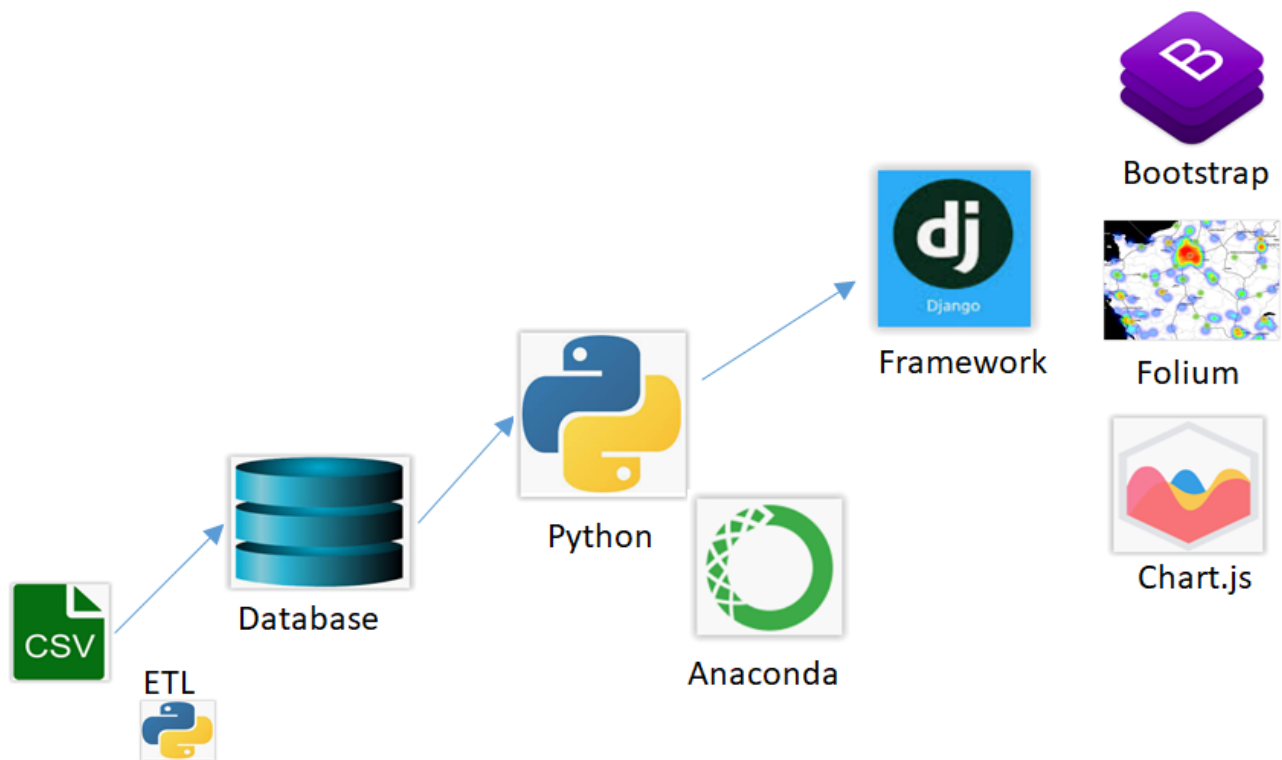


Figure 1: Technical Solution of the project.

II.3.1 SQLite

SQLite is a library written in C language that offers a relational database engine accessible by the SQL language. SQLite largely implements the SQL-92 standard and ACID properties. Our choice was made following some problems encountered during the preparation of the environment with Oracle, we finally opted for a lighter and faster RDBMS suitable for small projects. In addition, Django automatically creates a SQLite database for every project.



II.3.2 Python

In this project, our team choose Python over Java. For the first and the most important reason is that Python is the mostly used software in the data science field. The essential of the Python is that it simplified the complex coding grammar with fewer lines, and it is an open source that popular in business model. With more and more packages created, Python has an comprehend function to analyze data. The the other reason for using python over Java is that Python has a great flexibility that can easily connected with the web.



II.3.3 Django

Django is a web framework, which permit use Model-View-Template (MVT) architectural pattern. We decided to select Django for development de system because it is the most common Python web framework, allowing reusability of components, rapid development and some user control interface.

django

II.3.4 Chart.js

Chart.js is a JavaScript library that allows you to draw different types of charts by using the HTML5 canvas element.



Chart.js

Simple yet flexible JavaScript charting for designers & developers



II.3.5 Bootstrap

Bootstrap is an open source toolkit for developing with HTML, CSS, and Java Script.

II.3.6 Responsive

With the increase in Smartphone usage, the demand for responsive websites has increased tremendously. A responsive layout and 12-column grid system are present in Bootstrap that aid in website adjustment according to screen size.

II.3.7 Libraries and Essential Tools

NumPy

In scikit-learn, NumPy tables are the basic data structure. In fact, scikit-learn takes its data in the form of NumPy tables. All the data we use must be converted to a NumPy table. The central functionality of NumPy is the ndarray class, which is a multidimensional array (a n dimesions). All the elements of the table must be of the same type.

Pandas

Pandas is a Python library for data manipulation and analysis. It is built around a data structure called DataFrame. Pandas provides a large number of methods to modify and process this table. Another important interest of pandas is that it is able to work with a large number of file formats and databases, such as SQL, Excel files or CSV. The most common data structures that we use in this project are pandas.Series (One-dimensional ndarray with axis labels) and pandas.DataFrame. Folium

Is another data visualization library in Python that was built primarily to help visualize geospatial data. Sklearn.Preprocessing

The sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

The most common function we use is sklearn.preprocessing.LabelEncoder, which encodes labels with value between 0 and n_classes -1.

```
1 le = preprocessing.LabelEncoder()
2 le.fit(["ADEO", "ING3", "QFRM", "ADEO"])
3 list(le.classes_)
4 ['ADEO', 'ING3', 'QFRM']
5 le.transform(["ADEO", "ING3", "ADEO"])
6 array([1, 2, 1]...)
```

Sklearn.Ensemble.RandomForestRegressor

A random forest regressor

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

We use random forest regressor to fix missing values of attribute *RENUMERATION* in table *STUDENT INTERNSHIP*

Feature_Selector.FeatureSelector

In this project we will walk-through using the FeatureSelector class for selecting features to remove from a dataset. This class has methods for finding features as:

Features	Description
Missing values	Find any columns with a missing fraction greater than a specified threshold
Single Unique value	Find any features that have only a single unique value
Collinear (highly correlated) Features	This method finds pairs of collinear features based on the Pearson correlation coefficient.

II.4 Overleaf

For our documentations and reports we used Overleaf, it's a web-based freemium academic writing environment. Although the system is based on LaTeX, Overleaf provides a mature rich text editor that you can easily use. In addition it lets you git clone, push and pull changes and manage all versions of your documentations.

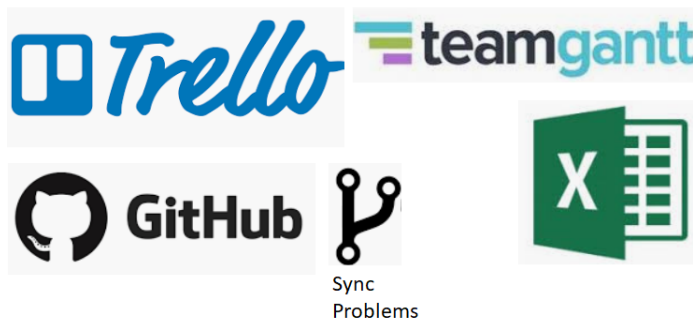


III PROJECT PLANNING REPORT

To organize the distribution of tasks and to meet deadlines of the project, we decided to use Scrum, an agile framework for managing the work. The idea of Scrum is to divide the project in a list of actions (backlog) that can be completed within time boxed iterations (sprints). Because of small time of all project, the sprints intervals are at maximum one week, and the daily scrums occurs before or after some normal class.

III.1 Project management tools

The tools chosen to assist in the execution of the scrum and in the follow-up of the project execution were Trello and TeamGantt. The Gantt graph is shown in the next topics.



To facilitate the exchange of information between the team, a group uses common chat tool and the code is shared in Github and documentation was written using an online latex framework.

III.2 Organization of the team and cutting into tasks

The project was divided in some groups, to evidence the deadlines of sub products and the application. Each group has some tasks, that could be linked to sub product. The following figures shows this groups, estimated time and deadline. The tasks will be chosen freely by team members, and dynamically allocated.

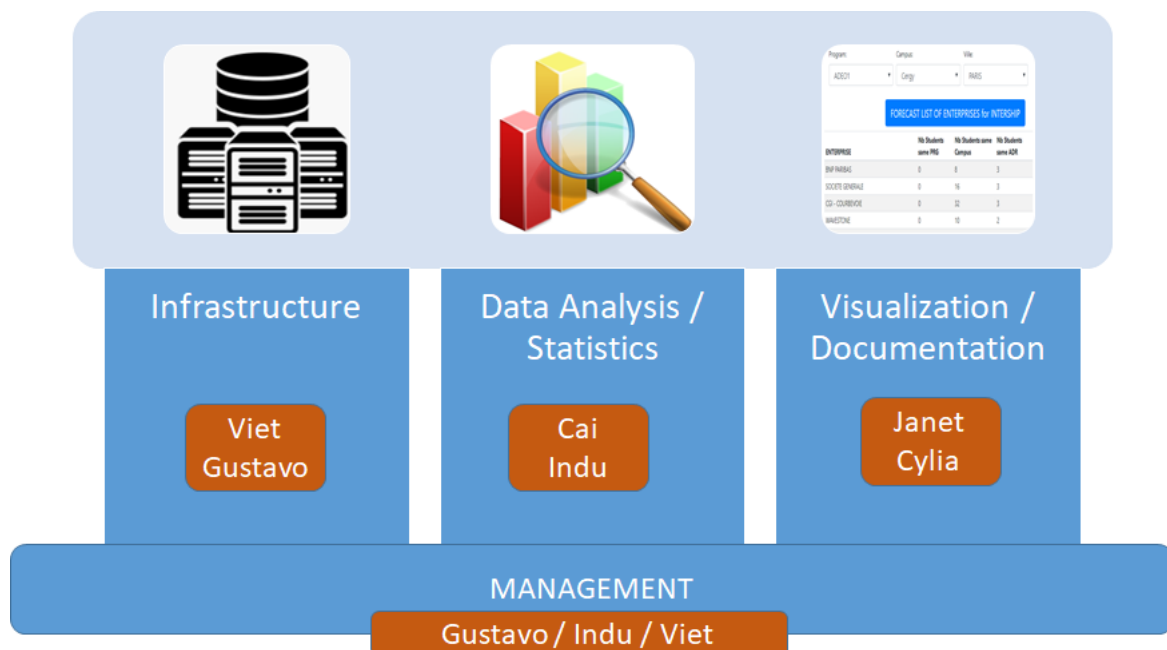


Figure 2: Example of tasks organization

These initial tasks are just to estimate the functions we can add in the application, considering the deadline, number of team member and experience of team in the chosen technology. Later we define the additional functionalities, the backlog will be filled by the new tasks at the beginning of each sprint.

III.3 Initial Gantt charts

The second column in the Gantt figures shows the percentage of the execution of the product.

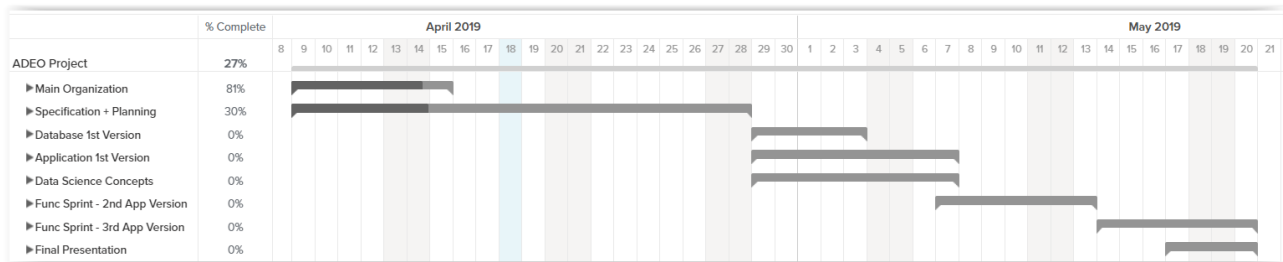


Figure 3: Schedule and current status of the final grade project in first sprint.

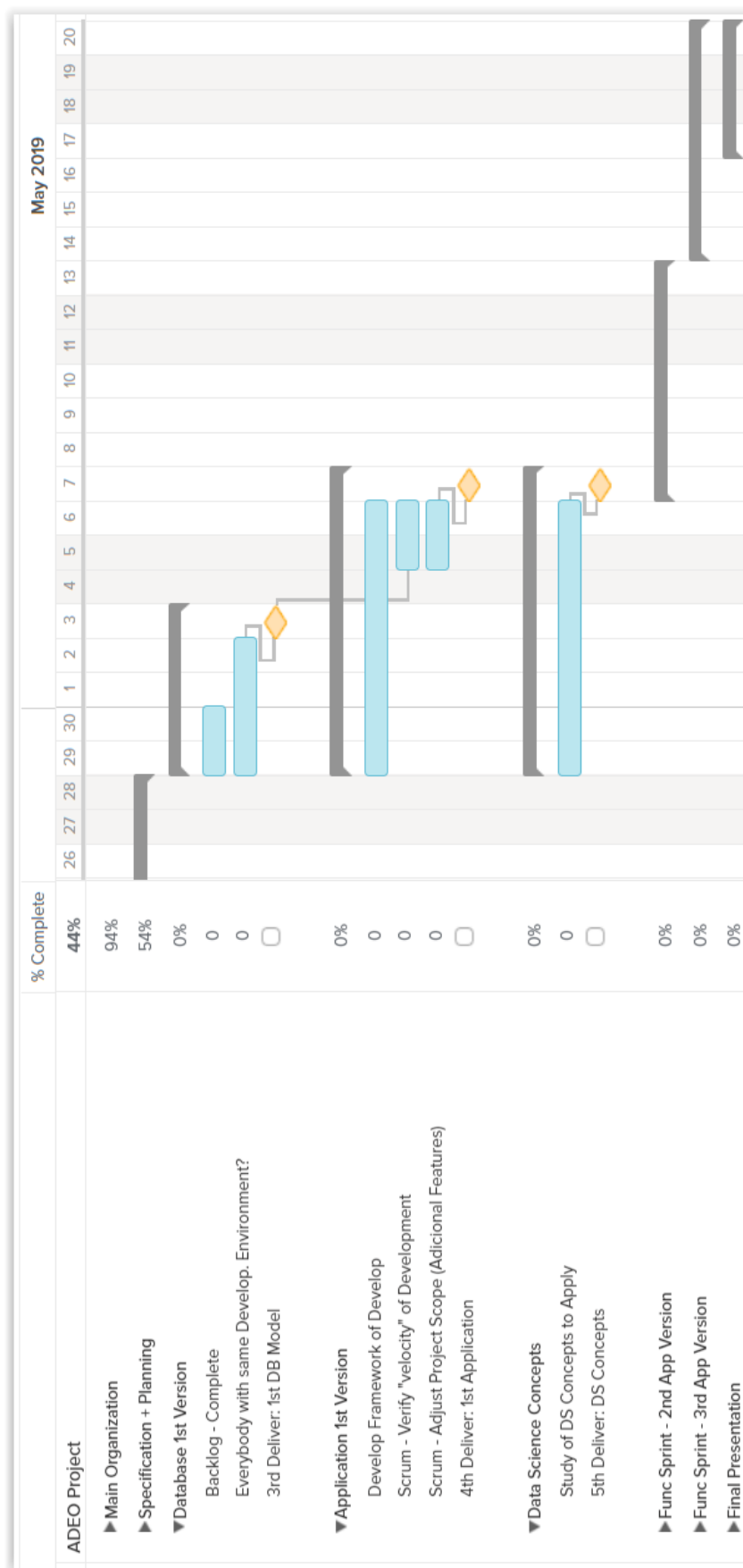


Figure 4: Tasks for "Database 1st Version", "Application 1st Version", "Data Science Concepts", "Functional Sprints" and "Final Presentation".

III.4 Update of the Status of the project in each sprint

The project steps done are showed in figure below. We update the date of Final Presentation and add one more Sprint for development. For the previous activities.

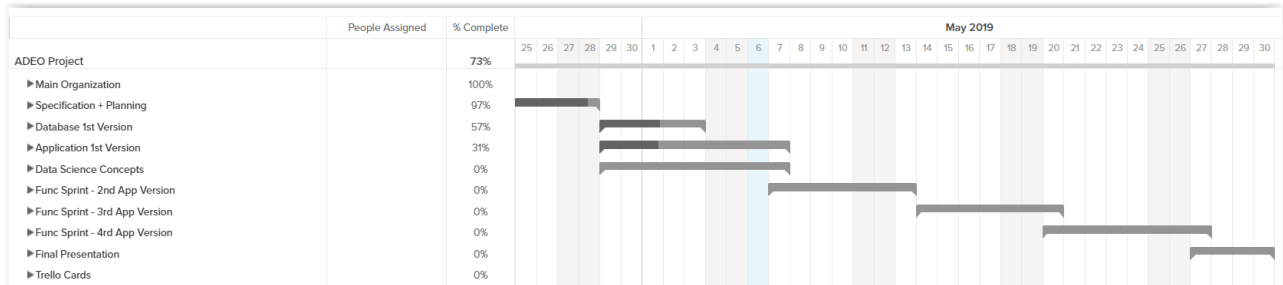


Figure 5: Schedule of the final grade project in Sprint 2.

	Progress
ADEO Project	100%
► Main Organization	100%
► Specification + Planning	100%
► Database 1st Version	100%
► Application 1st Version	100%
► Data Science Concepts	100%
► Func Sprint - 2nd App Version	100%
► Func Sprint - 3rd App Version	100%
► Func Sprint - 4rd App Version	100%
▼ Final Presentation	100%
Prepare Documentation	100%
Prepare Presentation	100%
Defense	<input checked="" type="checkbox"/>
► Trello Cards	0%

Figure 6: Schedule and current status of the final grade project in the last Sprint.

III.5 Conclusion

Overall, the tools used throughout this project allowed us to work with the agile method by implementing different aspects, such as the scum master, who managed the team, as well as the application release performed in each Sprint, We were able to meet the client's deadline and improve our solution at each step.

IV REALIZATION

IV.1 Data Source

In this part, we will show our main sources of data, consisting of three tables with which we will work. The following tables describe all their attributes and meanings.

IV.1.1 PRG_STUDENT_SITE_2017_2018_DATA_TABLE

VARIABLE	FORMAT	DEFINITION
ID_ANO	Long	Student's ID
PRG	Text	The student's curriculum
Annee_scolaire	Text	College year
Site	Text	University campus

Table 1: PRG_STUDENT_SITE_2017_2018_DATA_TABLE

IV.1.2 ADR_STUDENTS_2017_2018_DATA_TABLE

VARIABLE	FORMAT	DEFINITION
ADR_CP	integer	The student's address zip code
ADR_VILLE	Text	The student's home city address
ADR_PAYS	Text	Indicate the login of user
ID_ANO	Long	Student's Id

Table 2: ADR_STUDENTS_2017_2018_DATA_TABLE

IV.1.3 STUDENT_INTERNSHIP_2013_2018_DATA_TABLE

VARIABLE	FORMAT	DEFINITION
ANNEE	Integer	Year of internship
ANNEE_SCOLAIRE	Text	College year
ENTERPRISE	Text	Enterprise name
CODE_POSTAL	Integer	Zip code of the enterprise
VILLE	Text	Company city
PAYS	Text	Company's country
SUJET	Text	Description of the intership subject
REMUNERATION	Integer	Remuneration of the students
ID_ANO	Long	Student's ID

Table 3: STUDENT_INTERNSHIP_2013_2018_DATA_TABLE

IV.2 Data Model

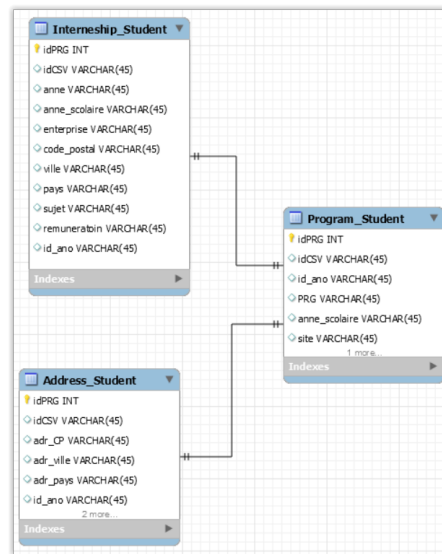


Figure 7: CSV files First Tables .

After ETL we merged like showed below :

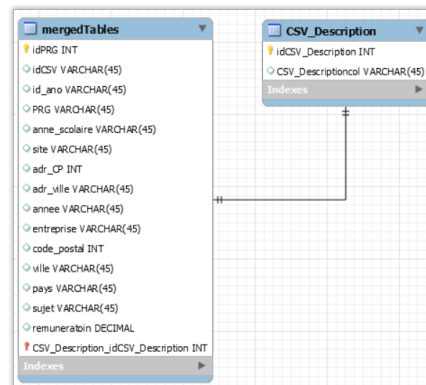


Figure 8: First Data Model.

IV.3 Implementation

In this document we will briefly describe the logic behind the framework used. An overview of the user interface and the different features implemented is detailed in appendice A.

In this project we are using Django, a Web Framework of Python with batteries included. It has everything which need to built Robust Framework. The main features or batteries are:

- The Model Layer
- The Views Layer
- The Template Layer
- Forms
- The Development Process

- The Admin
- Security
- Internationalization and Localization
- Performance and Optimization
- Python compatibility
- Geographic framework
- Common Web Application Tools

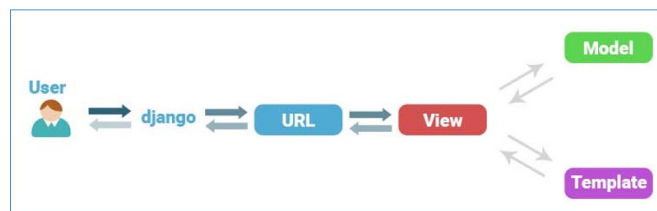


Figure 9: Here it's show how it work with Django

The Model-View-Template (MVT) is slightly different from MVC. Django itself takes care of the Controller part (Software Code that controls the interactions between the Model and View), leaving us with the template. The template is a HTML file mixed with Django Template Language (DTL).

V MACHINE LEARNING SOLUTION

V.1 ETL - Extract Transform and Load

The role of an ETL software is to collect relevant data from both files or systems, transform them to make them compatible with the Data Model, and finally load them into the Data Base.

The operation of the ETL platform is divided into three phases. The Extraction phase consists of collecting data from one or more sources.

The transformation phase consists of reformatting and transforming the data. Finally, the loading phase consists of transferring the transformed data to the target database.

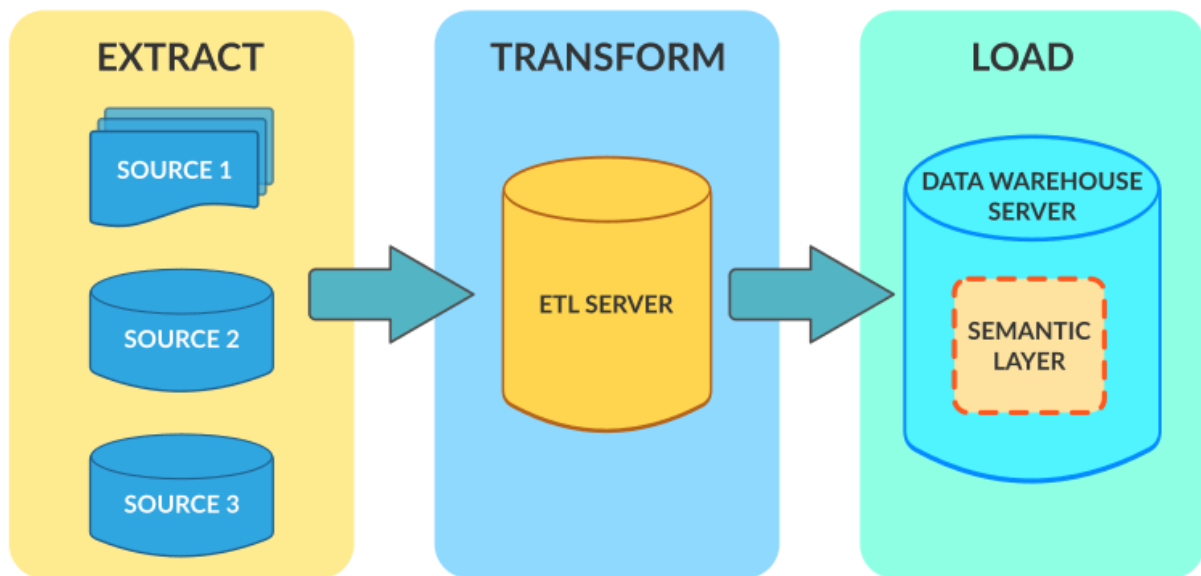


Figure 10: How ETL Works.

During the cleaning process of the data we went through several steps, one of them was the discovery of the missing values. All the steps are found in the Jupyter File. We allow the users to Delete Null Values and Merge tables or Filling with RandomForest and Merge Tables.

ETL

Select RAW DATA Version: 2

Missing Values - Data Version: 2

PRG_STUDENT_SITE	
COLUMN	NULL COUNT
id	0
ID_ANO	0
PRG	0
ANNE_SCOLAIRE	0
SITE	0
idCSV	0

ADR_STUDENTS	
COLUMN	NULL COUNT
id	0
ADR_CP	1722
ADR_VILLE	1513
ADR_PAYS	0
ID_ANO	0
idCSV	0

STUDENT_INTERNSHIP	
COLUMN	NULL COUNT
id	0
ANNEE	1
ANNEE_SCOLAIRE	3
ENTREPRISE	0
CODE_POSTAL	61
VILLE	6
PAYS	0
SUJET	9
REMUNERATION	353
ID_ANO	0
idCSV	0

Delete Null Values && Merge Tables

Filling with RandomForest && Merge Tables

Figure 11: ETL - Missing Values

V.2 Forecast Predict - Random Forest

Decision tree forests (or random forest classifier) are part of machine learning techniques. This algorithm combines the concepts of random subspaces and bagging and is used here to predict

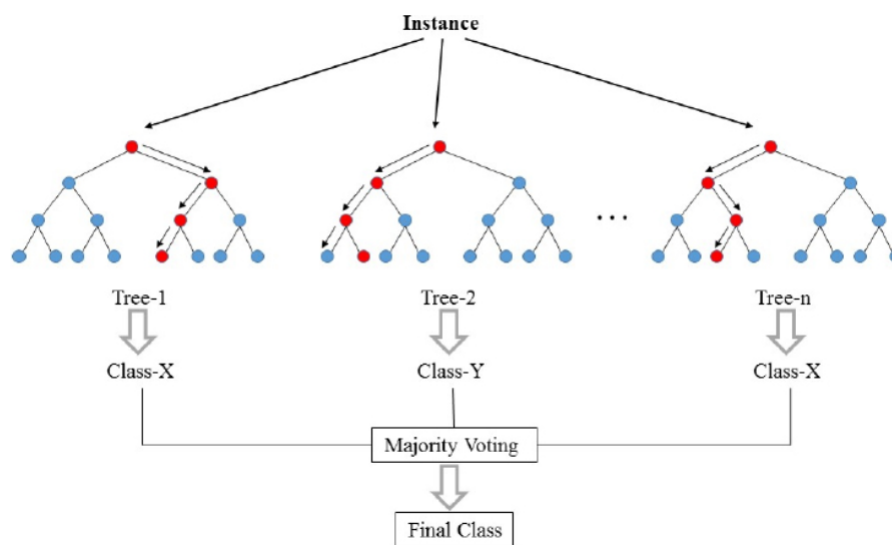


Figure 12: Random Forest

Forecast - Predict - Update Weight Table

Merged Table - Last Version 4

W0	W1	W2	ID MergedTable
0.6411390851115564	0.00995699529048115	0.2993355787543132	1
0.03784225332611805	0.6674696436704373	0.5210054703384425	2
0.22956508874324447	0.327188780656073	0.5813378124667714	3
0.16121312269137758	-0.039828951425608244	0.631103480177645	4

UPDATE

* This update could take several minutes (more than 10 minutes)

- Update Weights.PNG

Figure 13: Update Weights

Forecast the list of enterprises for internship, depending on three attributes, the program, the campus and the City.

Forecast - Predict

Version: 4

Program: Campus: Ville:

FORECAST LIST OF ENTERPRISES for INTERSHIP

ENTERPRISE	Nb Students same PRG	Nb Students same Campus	Nb Students same ADR
BNP PARIBAS	0	8	3
SOCIETE GENERALE	1	16	3
CGI - COURBEVOIE	0	32	3
SOPRA GROUP - LA DEFENSE	1	15	2
WAVESTONE	0	10	2
SOCIETE GENERALE - PARIS 18	0	12	2
VAL'EISTI	10	21	1
BKBIET - BIRLA INSTITUTE OF ENGINEERING AND TECHNOLOGY	4	6	1
ORANGE	2	6	1
THALES SERVICES - VELIZY VILLACOUBLAY	2	7	1

- Predict Internship.PNG

Figure 14: Forecast - Predict Internships

Enterprise Evolution

RAW DATA Version: 2 Apply

Enterprises
Number of
1404

Students
Number of
1393

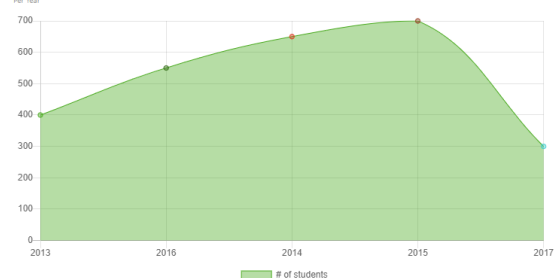
Enterprises 2016
Number of
710

Salary
Average of
€ 898.80

Biggest Number Students



Most Growing Qtd Internships



- Enterprise.PNG

Figure 15: Forecast - Enterprise

V.3 Descriptive statistics

This page allow the user to generate graphs of descriptive statistics for the EISTI database, using filters on any attribute.

V.3.1 Filters

Descriptive Statistics

Version: 4 Apply

Feel free to choose any option and if a option doesn't interest you, leave it as it is

Program's: Choose... Ville: Choose... Code Postal: Choose... (Open list: 92287, 75003, 75116, CH-6264, 95011, 75010, 75450, 93111, 92937, 91551, 75015, 20200, 93526, 75009, 92300, 69160, 75017, 75008, 93430)

Remuneration: Choose... Year's from: Choose...

☐ Check me out

Search

Figure 16: Descriptive statistics - Filters

V.3.2 Descriptive Statistics

Many graph types can be realized as curves, bar graphs, subdivided bar graphs or pie graphs. This part show us all the descriptive statistics of the data set, as the number of records, students, enterprises, and average salary, the student distribution between campuses (cergy and pau). The Salary and amout of student as well as their evolution during this past years.

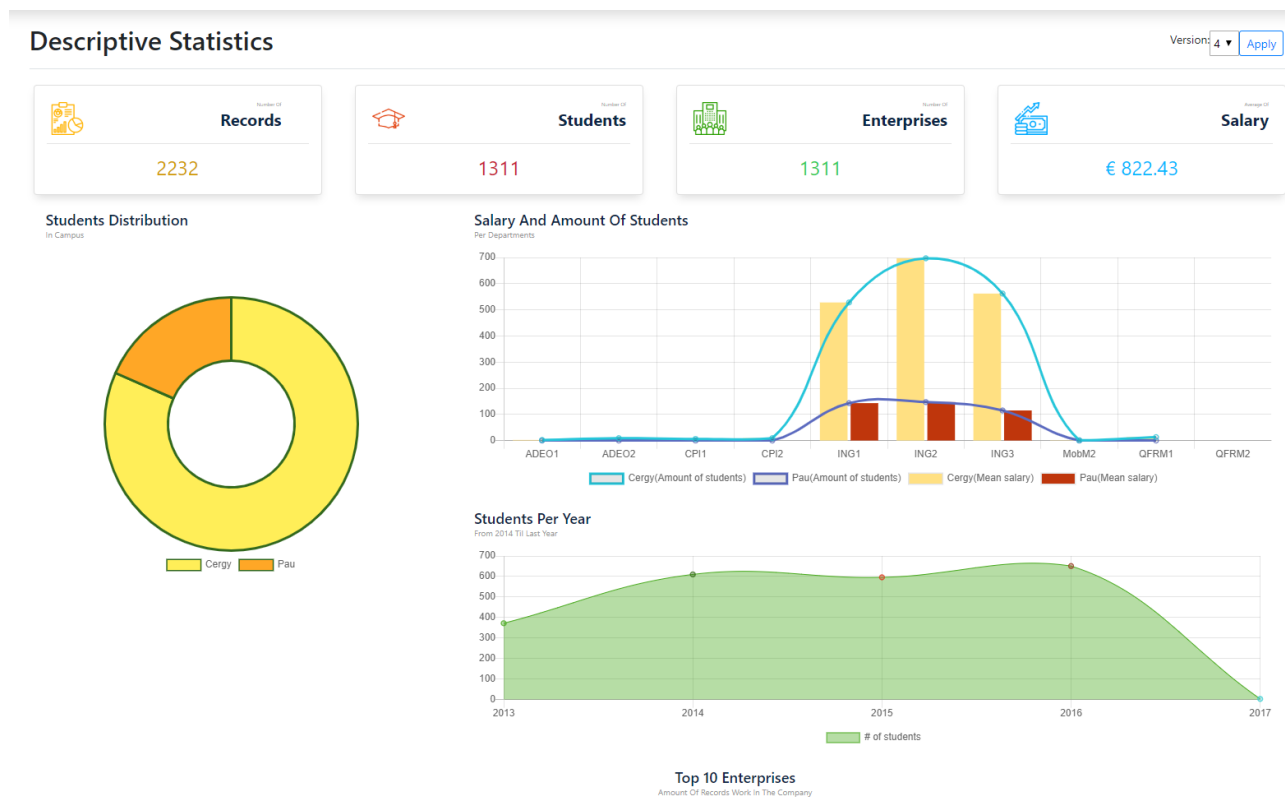


Figure 17: Descriptive statistics

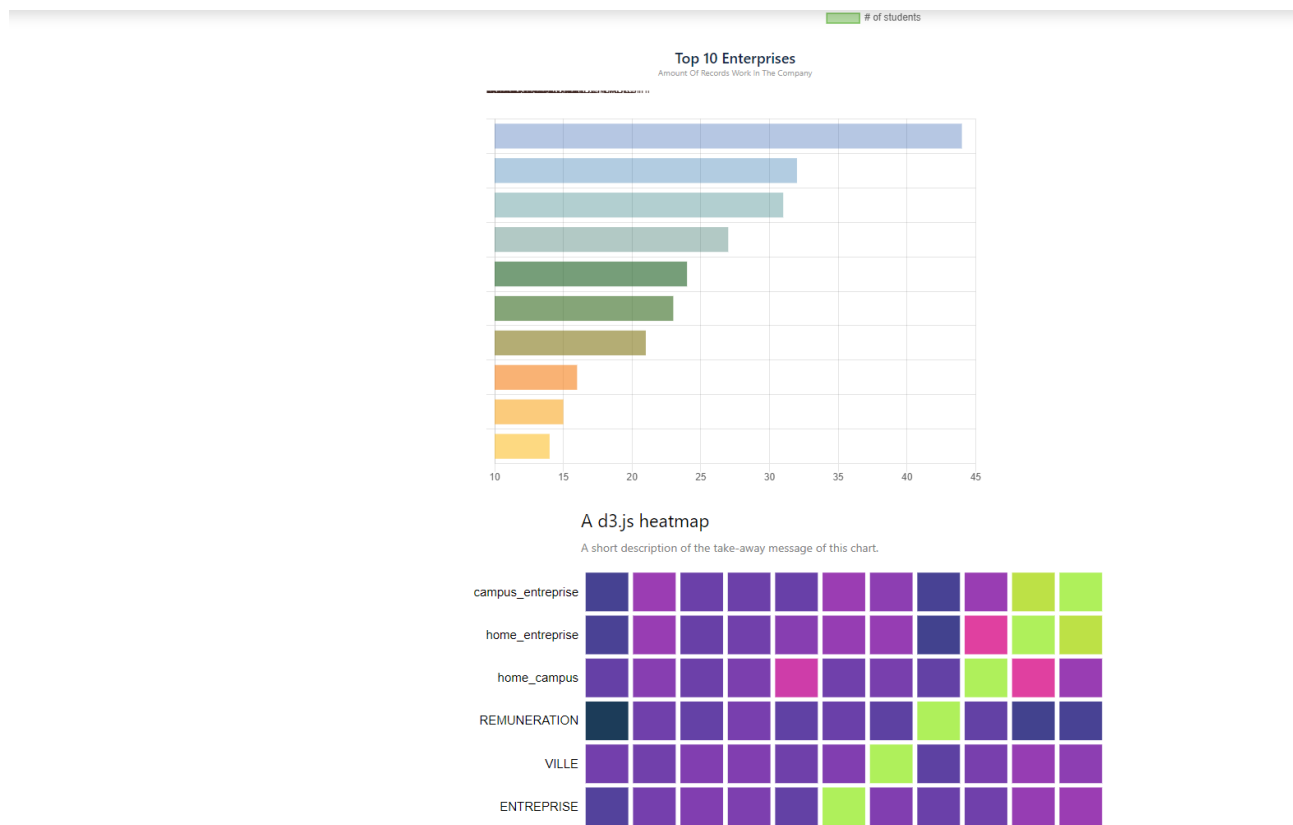


Figure 18: Top 10 of Companies

V.3.3 Heatmap

This heatmap shows the correlation between the different attributes of the database, such as the remuneration that a student can earn according to the program he or she follows. We can detect behaviors, predict when we have a strong correlation.

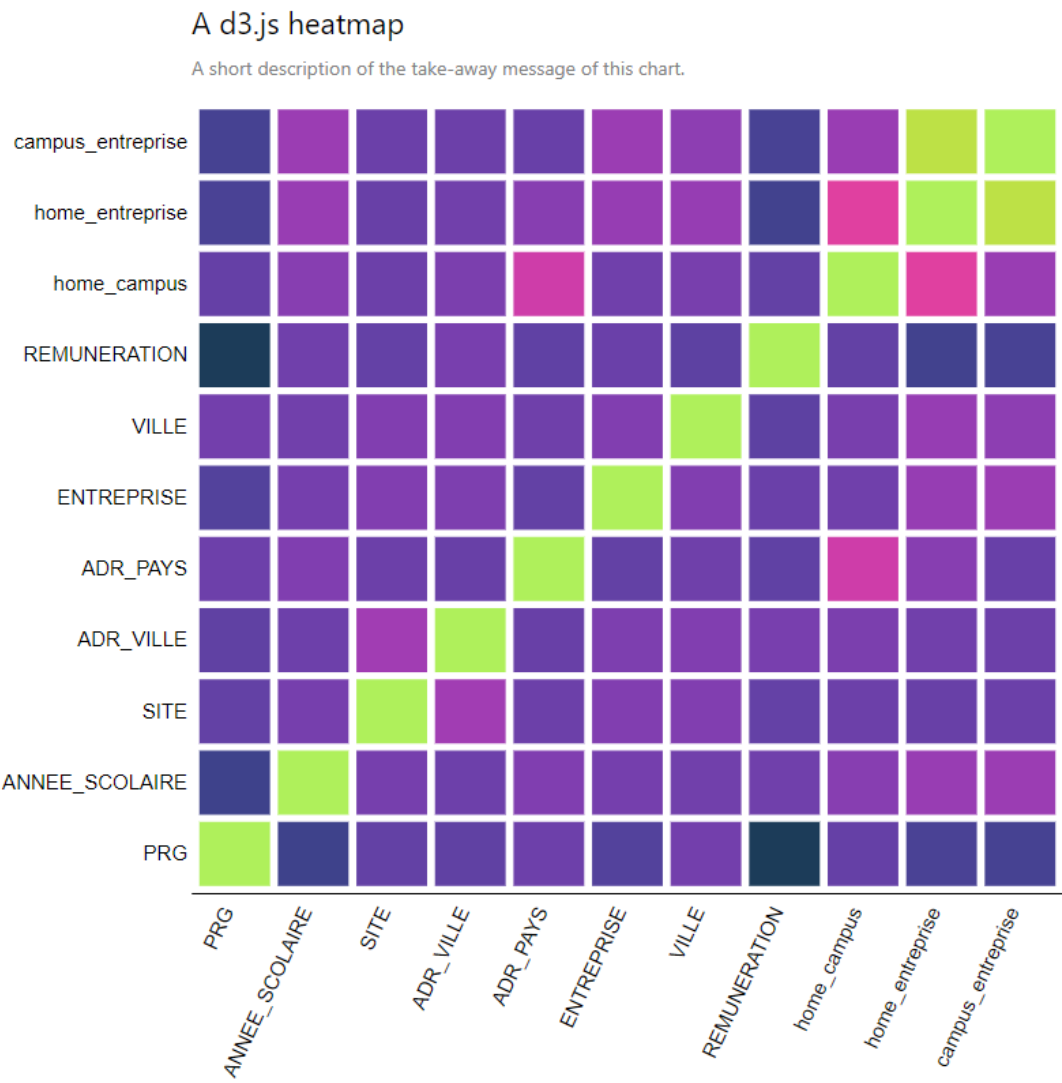


Figure 19: Descriptive statistics -Heatmap

V.4 Geographical statistic of engineering internships

Maps

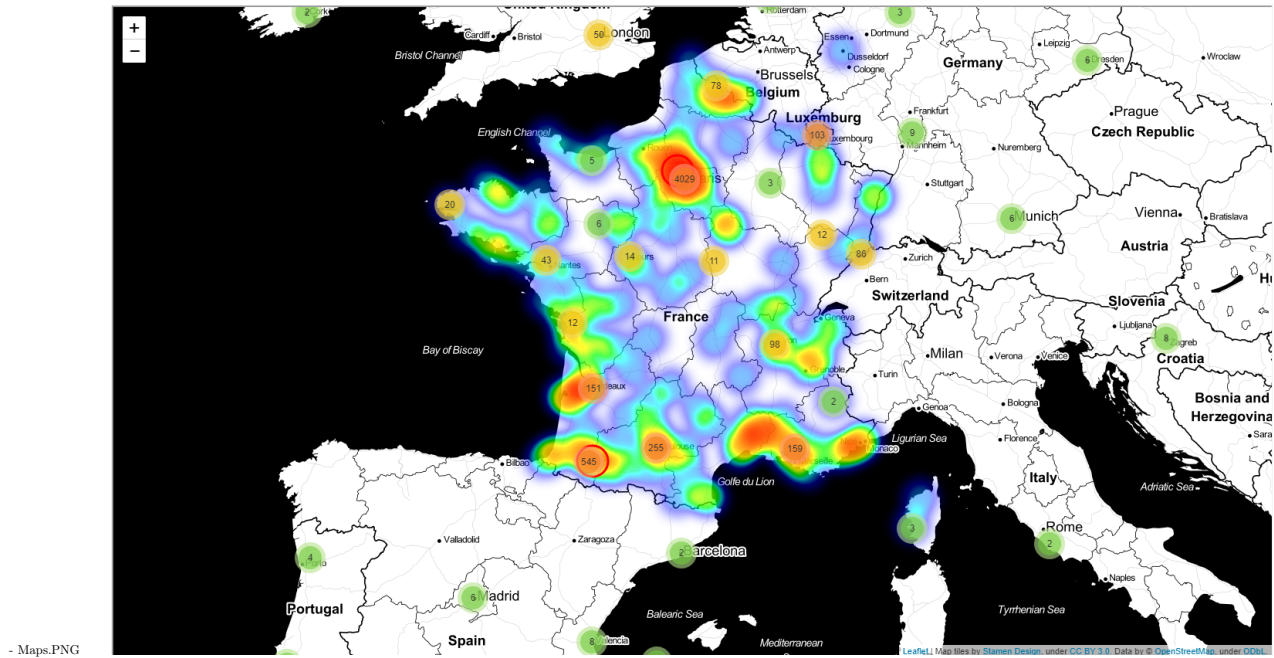


Figure 20: Distance between the students homes and campus

VI GENERAL CONCLUSION AND PERSPECTIVES

The general objective of the project was to build a tool for EISTI that can be used to visualize descriptive statistics of students internships, as what factor governs their choices to apply and go for that specific internship, allowing the business team to improve their skills in student orientation as well as decision making.

During this project we have been able to meet the needs of the customer, however, this system developed as a prototype needs to be refined. In perspective, we think about putting the system online, it would be very useful to the actors of the system, in addition we have listed some additional features and improvements that we can add in future versions.

- More friendly user interface with helps and Pop-ups.
- Suggestion and Guiding error message
- Uniform graph interfaces
- Additional graphs
- Reporting (Ex: PDF) /Printing Options
- More admin options, such as user privileges
- Problems with enterprise forecasting (name error's)

Appendices

A USER INTERFACE

A.1 Home

First Interface is the home screen, the user can register if he does not have account already, login or send a message through contact us.

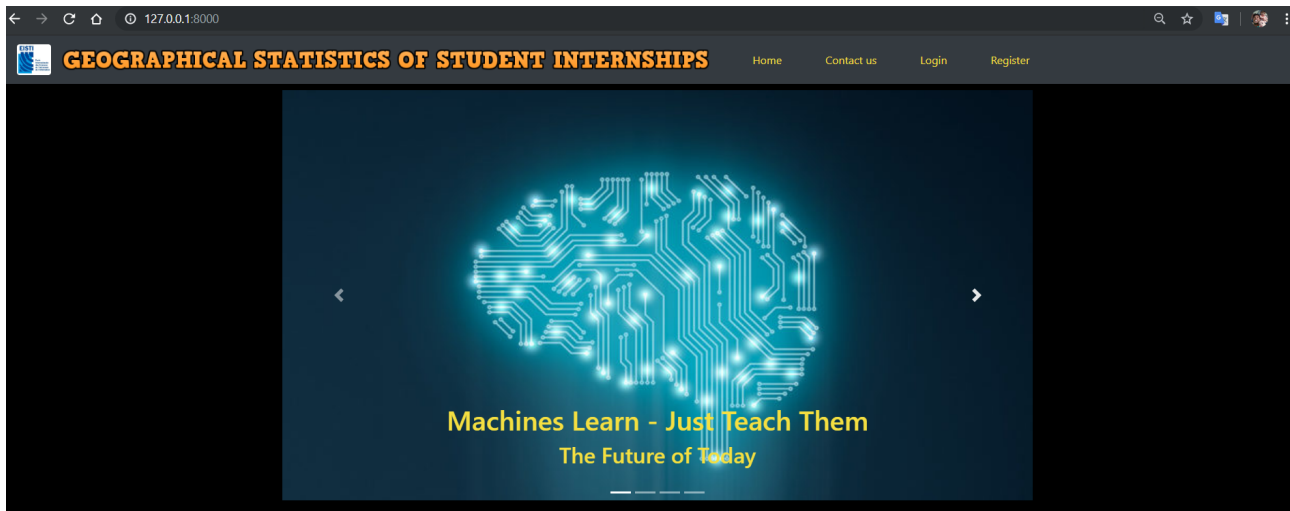


Figure 21: Home Screen Interface

A.2 Login

The user login to his account using a User name and a Password, if he forgets his, password, he can click on Forget Password.

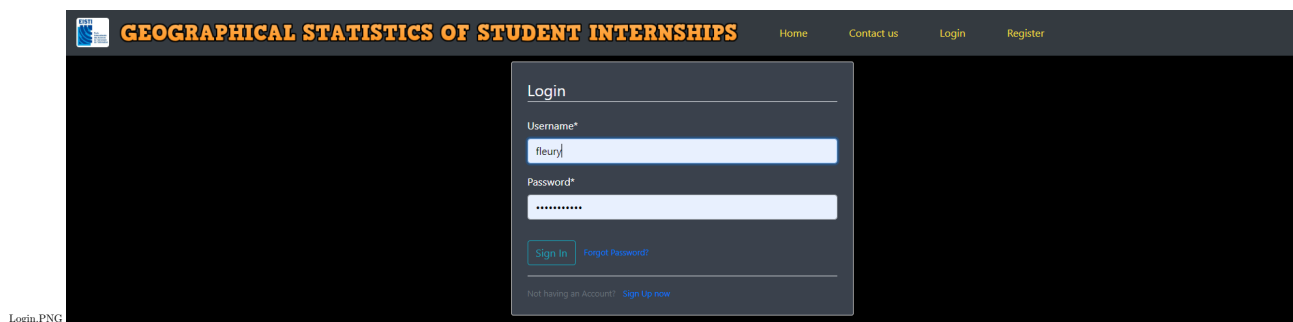


Figure 22: User Login Interface

A.3 Reset Password

To reset his password, the user needs to enter the e-mail address linked to his user account, after clicking Request Password Reset, an e-mail will be sent to allow the user to set a new password.



Figure 23: Interface for Password Reset

A.4 Contact Us

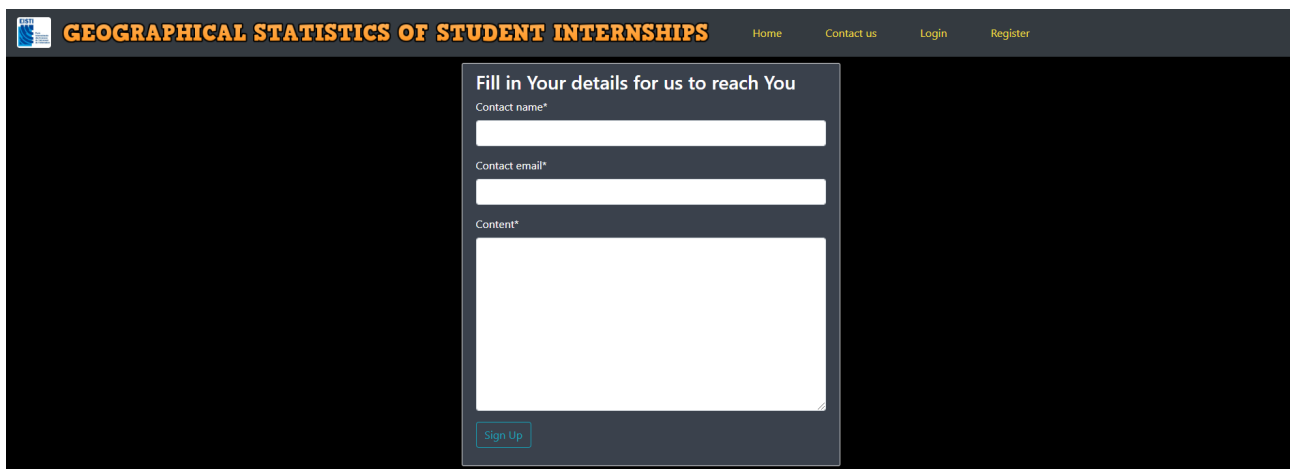


Figure 24: Contact Us Interface

A.5 Register

The users are allowed to register through this form, by entering a user name, e-mail and password, the requirements for the password are shown to the User.

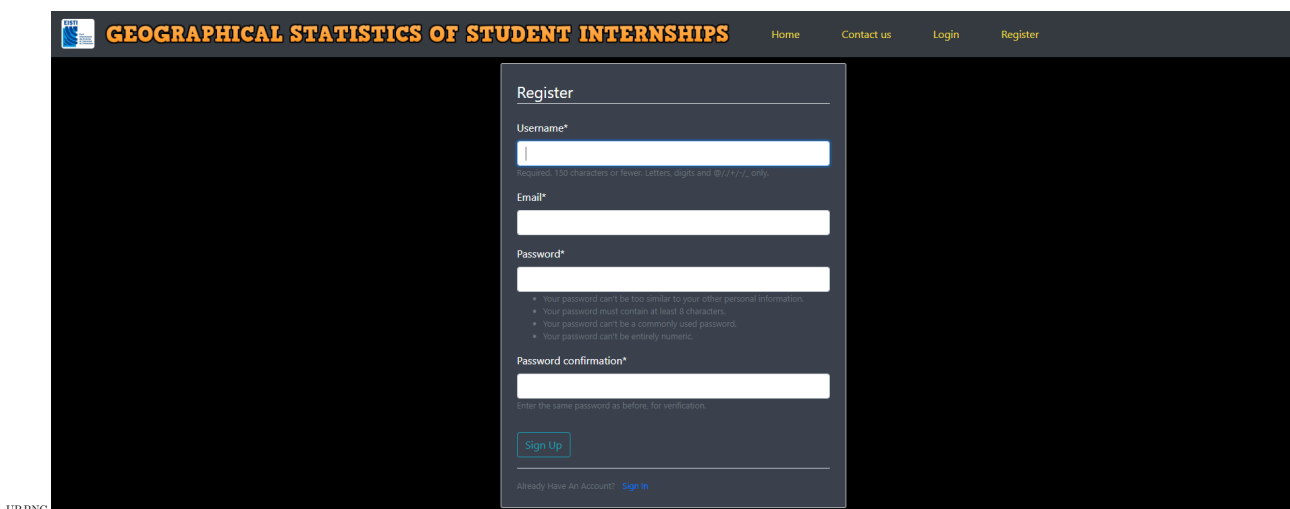


Figure 25: Register Interface

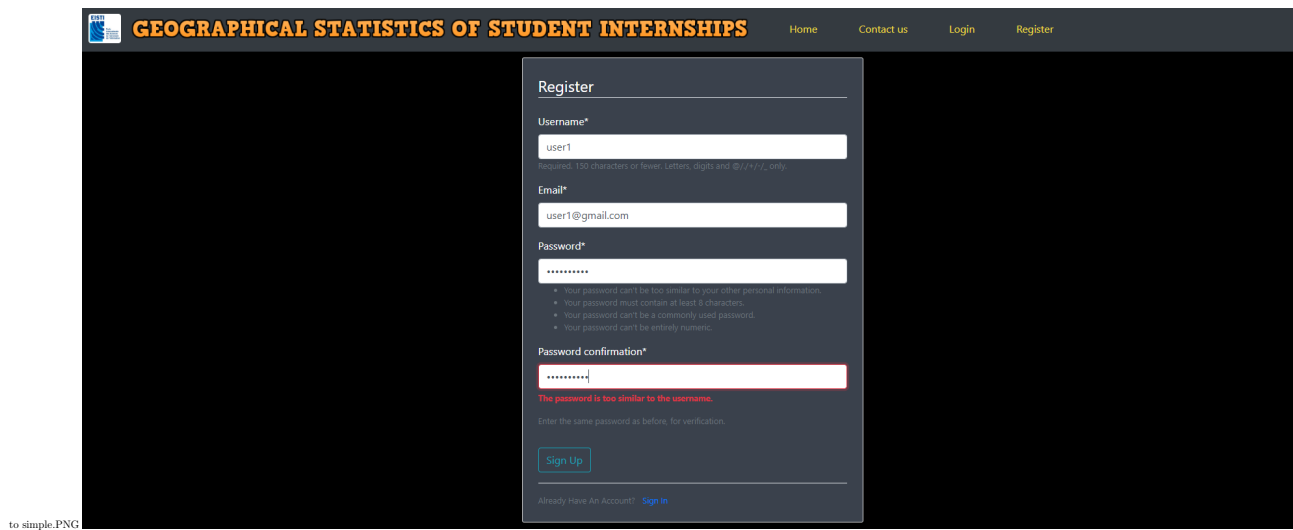


Figure 26: Register Interface password restrictions

A.6 Main Page

Basically this main page will give the user an overview of all the mains features of the application :

- Load Data
- Descriptive statistics
- Distance Calculation
- Access to the main Menu
- Filters applications
- Roles (Admin/User)

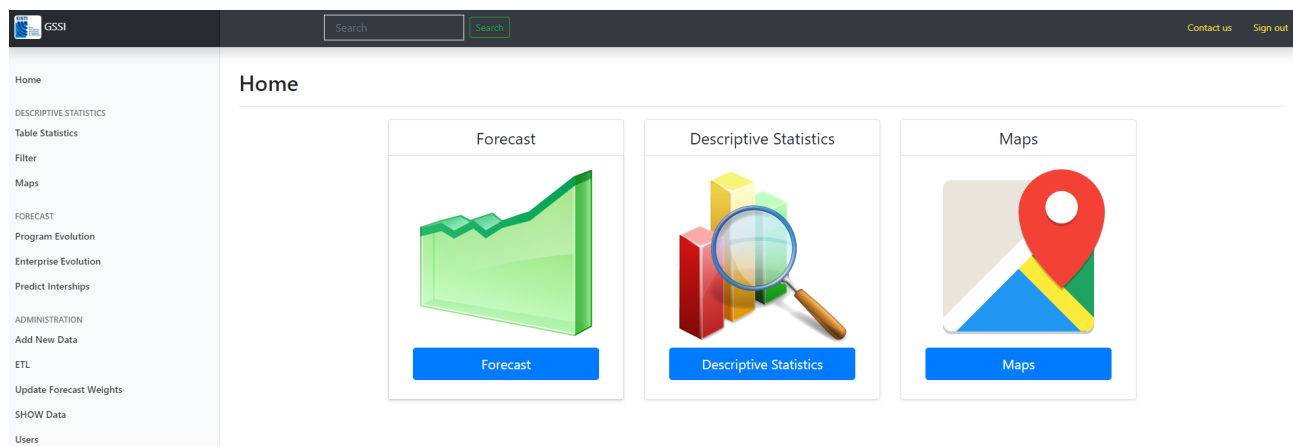


Figure 27: Main Page for Admin

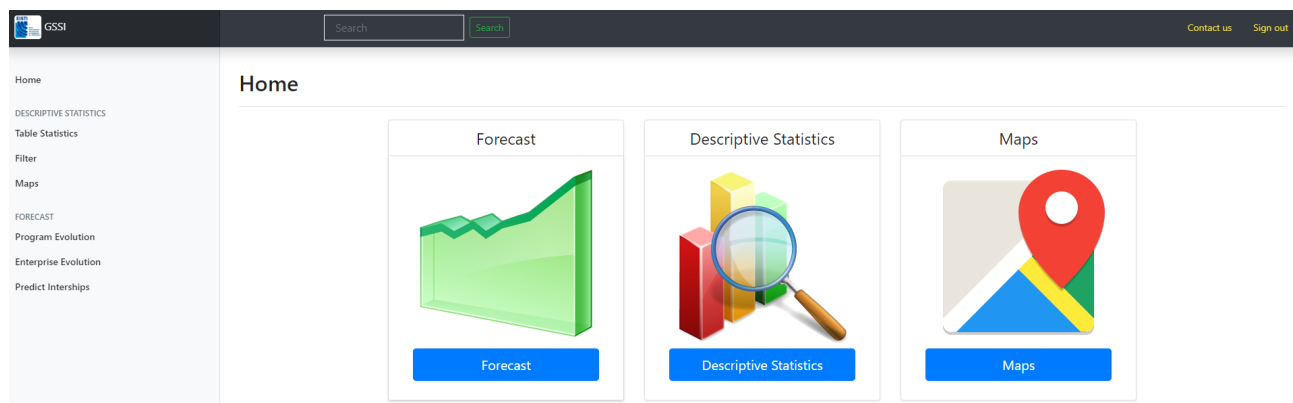


Figure 28: Main Page for simple user

A.7 CSV Files Load

The admin is supposed to upload the 3 sets of csv file. The model doesn't accept file types other than csv. The csv file that admin is allowed to upload also has restrictions like the number of attribute each csv file has and the order of the attributes within each of the csv file.

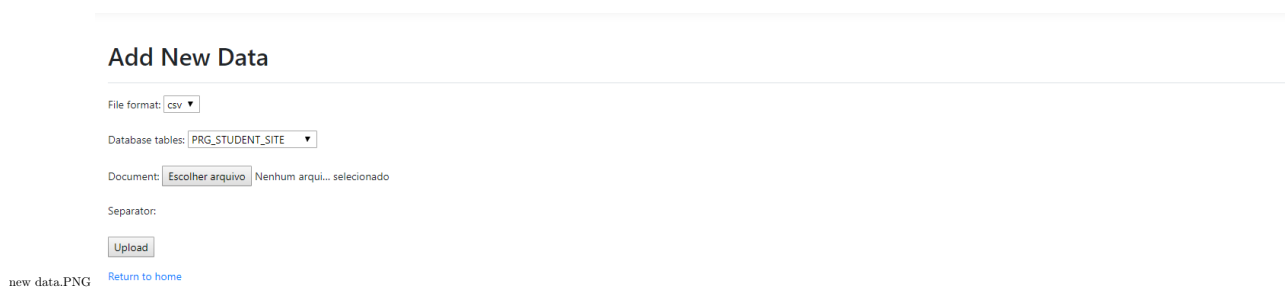


Figure 29: CSV Files Load

A.8 ETL

Once he has uploaded the csv file of right format, the backend code lists down the number of null and other data's that are random anomalies. The back end code is programmed either to delete these data's with anomalies or to use the machine learning algorithm namely random forest to intelligently fill the empty data's and datas with anomalies. The end user is given the choice as whether to apply the algorithms or just delete.

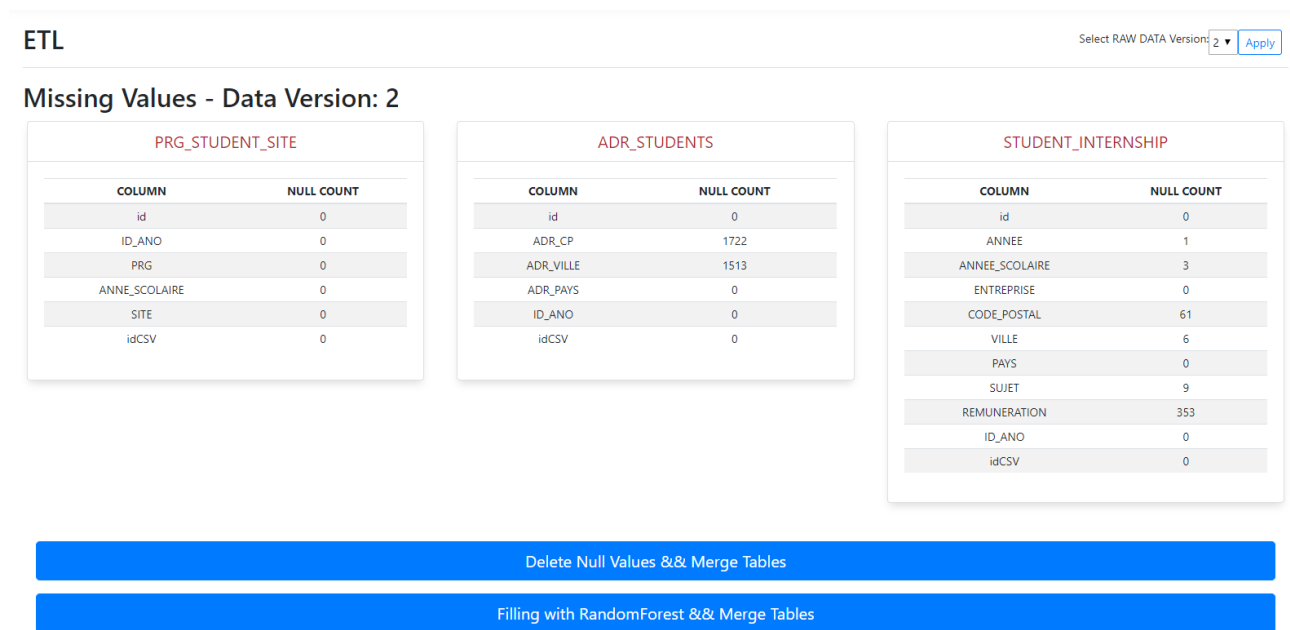


Figure 30: ETL Merged Tables

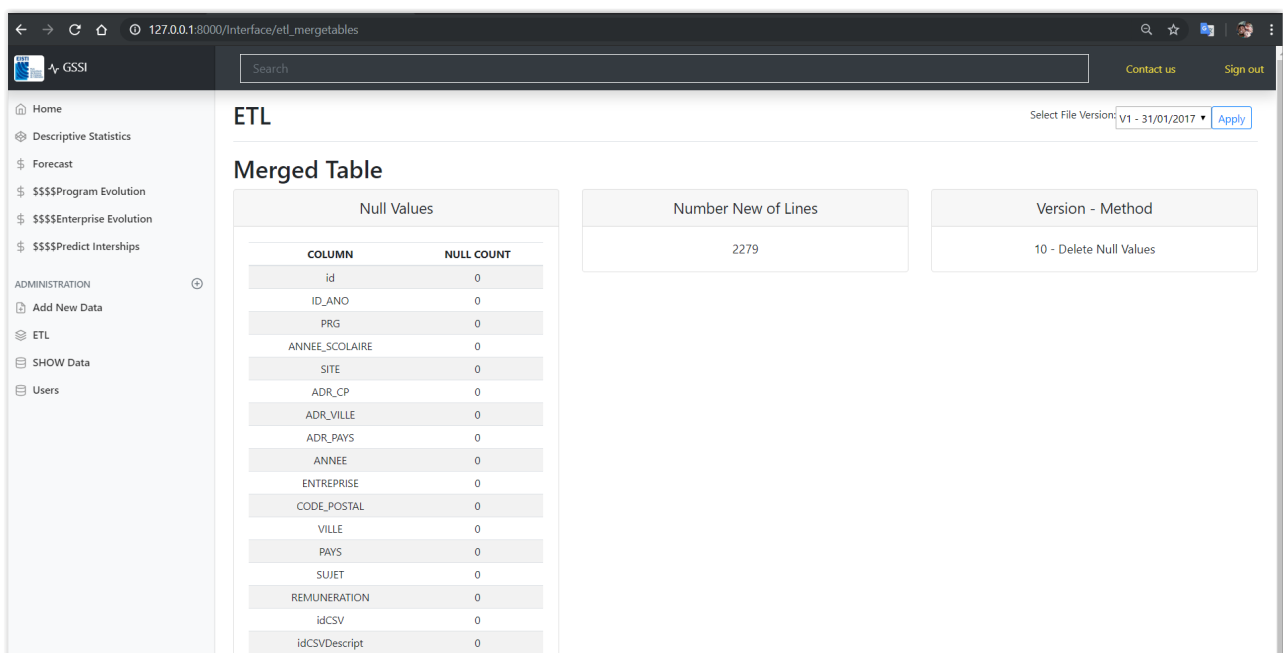


Figure 31: ETL

- Show Data.PNG

Program Student Site		Student Address	Student Internship
ID_ANO	PRG	ANNE_SCOLAIRE	SITE
56255130	MAIN	2013/2014	Cergy
56255130	ING1	2013/2014	Cergy
58229224	INEM	2015/2016	Cergy
58229224	ING3	2015/2016	Cergy
58229224	GSI	2014/2015	Cergy
58229224	ING2	2014/2015	Cergy
58229224	GI	2014/2015	Cergy

Figure 32: Data base Tables