

# **Student Specialty Counselling Using Machine Learning Techniques**

**Duhamel Parfait Tsopgni Tsague**

TSOPGNIDUHAMEL@GMAIL.COM

*Computer Science Department*

*National Advanced School of Engineering of Yaounde*

*Yaounde, Cameroon*

*P.O. Box 8390 Yaounde*

**Oréal Shaguile Chimi Youkap**

CHIMISHAGUILE@GMAIL.COM

*Computer Science Department*

*National Advanced School of Engineering of Yaounde*

*Yaounde, Cameroon*

*P.O. Box 8390 Yaounde*

## **1 Introduction**

Education is one of the most important field in the development of a country, through which people acquire knowledge to go through the other sectors of the economy. Thus, it is important to make sure people perform well in their studies. In several training schools around the world, and mainly in Cameroon, training extends over a period of five years including two years of preparation (general training) and three years of specialization. The choice of specialization departments is made by the students themselves and is motivated by their personal wishes, the opinions of their friends, the demands of parents and relatives. This choice is generally made without taking into account the intellectual aptitudes of the learner, the results obtained during the two years of generalized training or even the disciplines taught in the departments of specialization; and this leads to a big problem: the poor performance of the students in the specialization departments they have chosen.

In this paper, we will build a recommender system that, based on the results obtained by the student during the two years of generalized training and the subjects taught in the specialization departments, will suggest to that student the three (03) specialization departments where his performance might be the best. In fact, the system will suggest the three (03) specialization departments that are best suited to the student.

## **2 Related work**

The state of the art carried out on the system of orientation of students towards specialization departments rather redirected us to the Career Guidance System or Student Career Prediction. We therefore see that the majority of machine learning work is rather

oriented towards the choice of careers of the students and not on the choice of the specialty that he wants to do.

The main problem here is that once you have made the wrong choice of department, the choice of professional career is completely skewed. It is therefore important to see crucial to make a good choice of department of specialization so that the future choice of the professional career of the student matches the intellectual skills. Among all those works on career Guidance System, the most relevant that we picked and studied to understand how models can be built are:

- STUDENT CAREER PREDICTION Idyapriya .C :  
<https://ijert.org/papers/IJCRT195700.pdf>
- Career Guidance System using Machine Learning For Engineering Students (CS/IT): <https://www.irjet.net/archives/V7/i6/IRJET-V7I6640.pdf>
- Student Career Prediction Using Advanced Machine Learning Techniques Roy:  
<https://www.sciencepubco.com/index.php/ijet/article/view/11738/4565>

### 3 Methodology and implementation

To successfully complete this project, we followed a methodology based on four (04) mainstages:

- Definition and understanding of the problem
- Data collection
- Data preprocessing
- Label Encoding
- Machine learning algorithms
- Training and testing
- Result

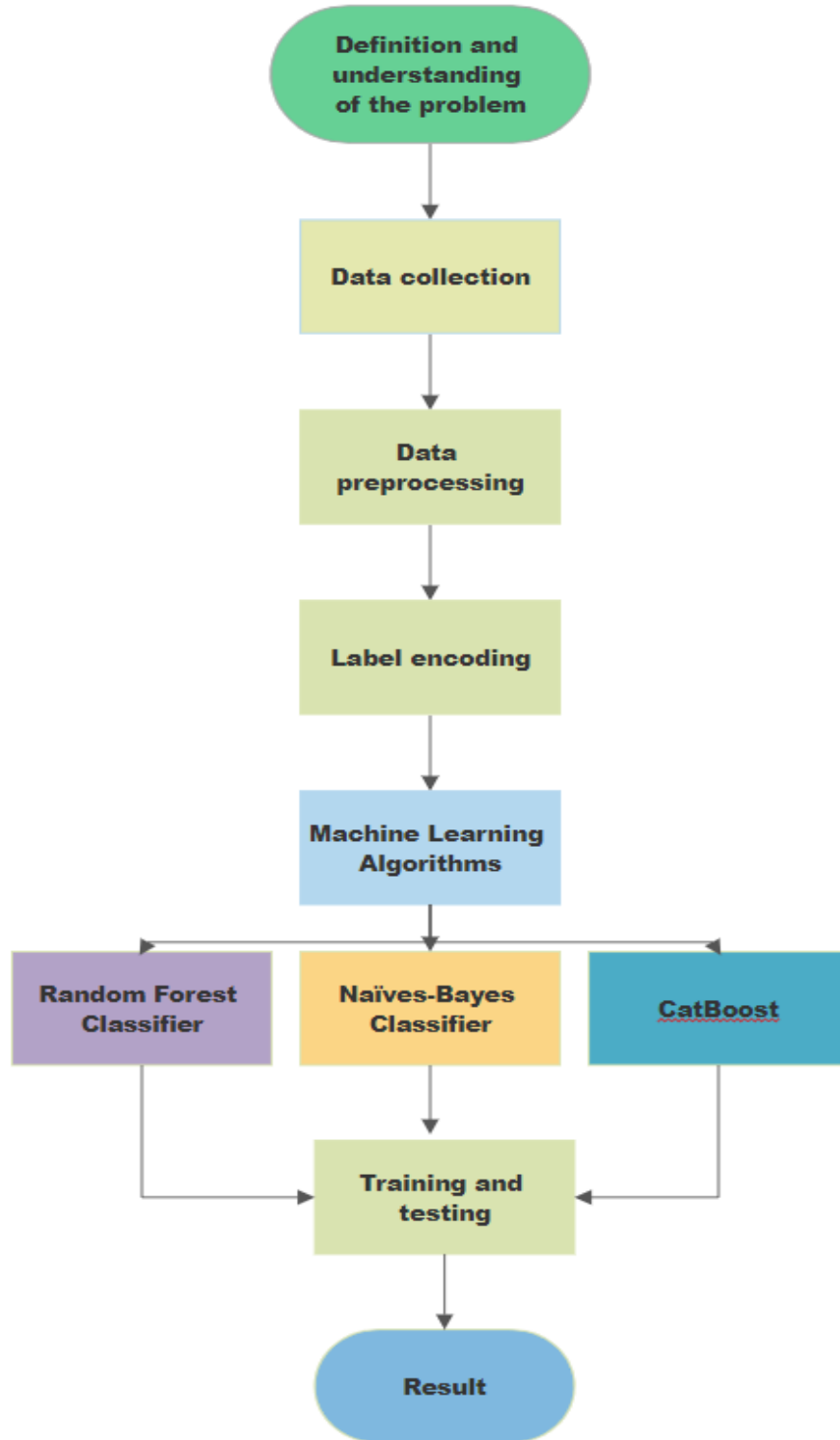


Figure 1: Methodology process

### 3.1 Definition and understanding the problem

In Cameroon , in training schools which extend over several two cycles (first cycle for general training and second cycle for specialized training), it was found that the performance of several students in specialization departments was very low compared to their performances during the first two years of training. This poor performance is due to the fact that the students were misguided in relation to the department of specialization which was best suited to their intellectual aptitudes. This problem impacts not only the training schools but also the students of the second cycle in high schools in the sense that after having made the classes of form 1 to form 5, the student is brought to make a choice of series A, C or D and these choices are generally badly made because they do not take into account the performances that the pupils have them in the preceding classes. The direct consequence visible to all is the high rate of failures in the Lower sixth (Probatoire) and Upper sixth (Baccalauréats) exams.

To solve this problem, the solution that we are going to construct is a system of recommendation, which is based on the marks obtained by the students in the previous classes and the subjects taught in the specialization departments, to offer the student the three best departments in which his academic performance would be the best.

### 3.2 Data collection

Collection of data is one of the major and most important tasks of any machine learning projects. Because the input we feed to the algorithms is data. So, the algorithms efficiency and accuracy depends upon the correctness and quality of data collected. So as the data same will be the output. For student specialization suggestion, many parameters are required like students' academic scores in various subjects and the subjects taught in specializations.

Therefore, for this project, we took the data of former students of the National Advanced School of Engineering of Yaounde. We took the data of the former students because with that, we can directly see if they have performed well in the specialization department that they had chosen.

### 3.3 Data preprocessing

Collecting the data is one task and making that data useful is an-an-another vital task. Data collected from various means will be in an unorganized format and there may be lot of null values, in-valid data values and unwanted data. Cleaning all these data, replacing them with appropriate or approximate data, removing null and missing data, and replacing them with some fixed alternate values are the basic steps in preprocessing of data.

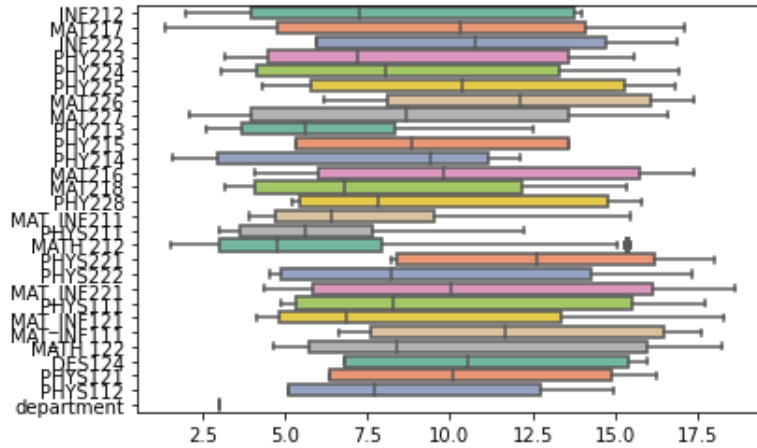


Figure 2: Distribution of the students' marks during the two first years

### 3.4 Label encoding

At the end of this stage, we have selected 1500 entries to use in our model. An entry being the marks that a student obtained in all the subjects of the two first years of training and the department which did after its first two years of training. It was therefore necessary to determine if a student performed well in the specialization department that he had chosen.

### 3.5 Machine learning algorithms

The problem that we had to solve being a problem of classification (with several classes), we tested several algorithms of classification in particular:

- **Naïves-Bayes**

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

Abstractly, naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector  $\mathbf{x} = (x_1, \dots, x_n)$  representing some  $n$  features (independent variables), it assigns to this instance probabilities  $p(C_k | \mathbf{x}_1, \dots, \mathbf{x}_n)$  for each of  $K$  possible outcomes or classes  $C_k$

Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k | \mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

Transposing this to our problem, the specialty department that will have the maximum probability given the old marks of the student will be the specialty department where his academic performance might be the best.

- **Random Forest Classifier**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  with responses  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ , bagging repeatedly ( $B$  times) selects a random sample with replacement of the training set and fits trees to these samples:

**For  $b = 1, \dots, B$ :**

1. Sample, with replacement,  $n$  training examples from  $\mathbf{X}, \mathbf{Y}$ ; call these  $\mathbf{X}_b, \mathbf{Y}_b$ .
2. Train a classification or regression tree  $f_b$  on  $\mathbf{X}_b, \mathbf{Y}_b$

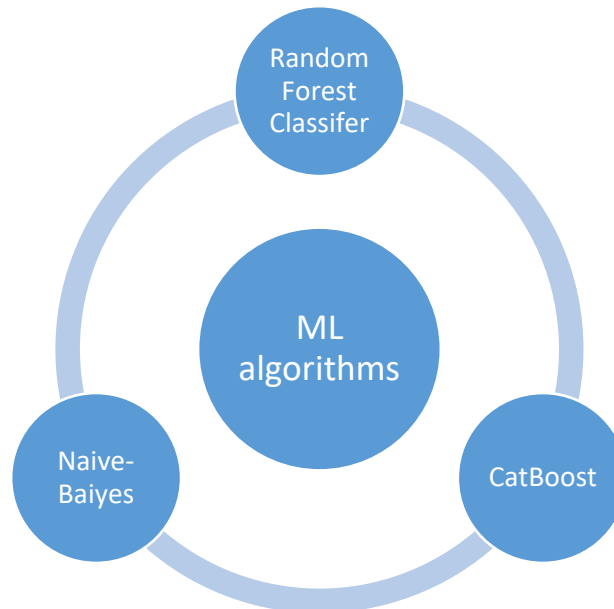
After training, predictions for unseen samples  $\mathbf{x}'$  can be made by averaging the predictions from all the individual regression trees on  $\mathbf{x}'$ :

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(\mathbf{x}')$$

Or by taking the majority vote in the case of classification trees.

- **CatBoost Classifier**

CatBoost is an open-source software library developed by Yandex. It provides a gradient boosting framework, which attempts to solve for Categorical features using a permutation driven alternative, compared to the classical algorithm.



**Figure 3: Overview of the various Machine Learning Algorithms**

### 3.6 Training and testing

Finally, after processing of data and training the next task is obviously testing. This is where performance of the algorithm, quality of data, and required output all appears out. From the huge dataset, collected **80 percent of the data is used for training and 20 percent of the data is reserved for testing**. Training as discussed before is the process of making the machine to learn and giving it the capability to make further predictions based on the training it took. Whereas testing means already having a predefined data set with output also previously labelled and the model is tested whether it is working properly or not and is giving the right prediction or not. If maximum number of predictions are right then model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model.

Also further new set of inputs and the predictions made by the model will be keep on adding to the dataset, which makes dataset more powerful and accurate.

### 3.7 Results

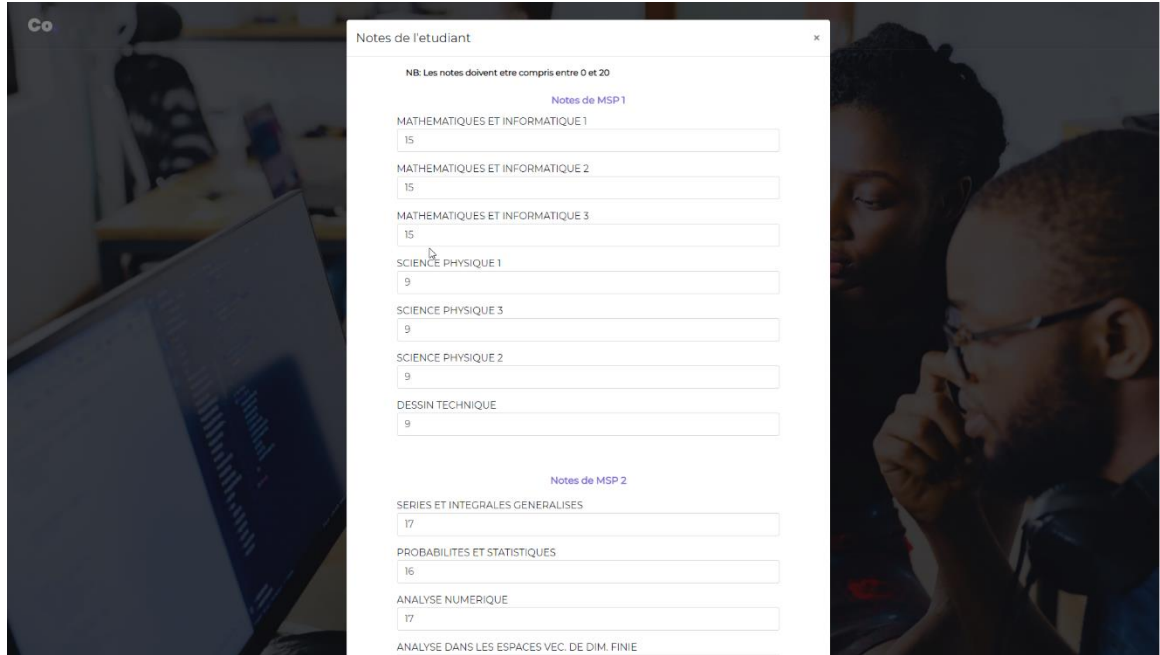
The data is trained and tested with many algorithms and out of all Random Forest Model gave more accuracy with **72 percent**. As Random Forest Model gave the highest accuracy, all further data predictions are chosen to be followed with that model. Finally, a web application is made to give the input parameters of the student and the final prediction is generated and displayed. The background algorithm being used is Random Forest Model and the new prediction are keep on adding to the dataset for further more accuracy.

Model	Random Forest	Naïve-Bayes	CatBoost
Accuracy (%)	<b>72</b>	<b>62</b>	<b>65</b>

#### 3.7.1 Prototype

We have integrated our final model into a web application that is our prototype. We have hosted it on Heroku and you can access it through the following link :

<https://mlpc-stud-dept-counc-prod.herokuapp.com/>



Notes de l'etudiant

NB: Les notes doivent etre compris entre 0 et 20

Notes de MSP 1

MATHEMATIQUES ET INFORMATIQUE 1  
15

MATHEMATIQUES ET INFORMATIQUE 2  
15

MATHEMATIQUES ET INFORMATIQUE 3  
15

SCIENCE PHYSIQUE 1  
9

SCIENCE PHYSIQUE 3  
9

SCIENCE PHYSIQUE 2  
9

DESSIN TECHNIQUE  
9

Notes de MSP 2

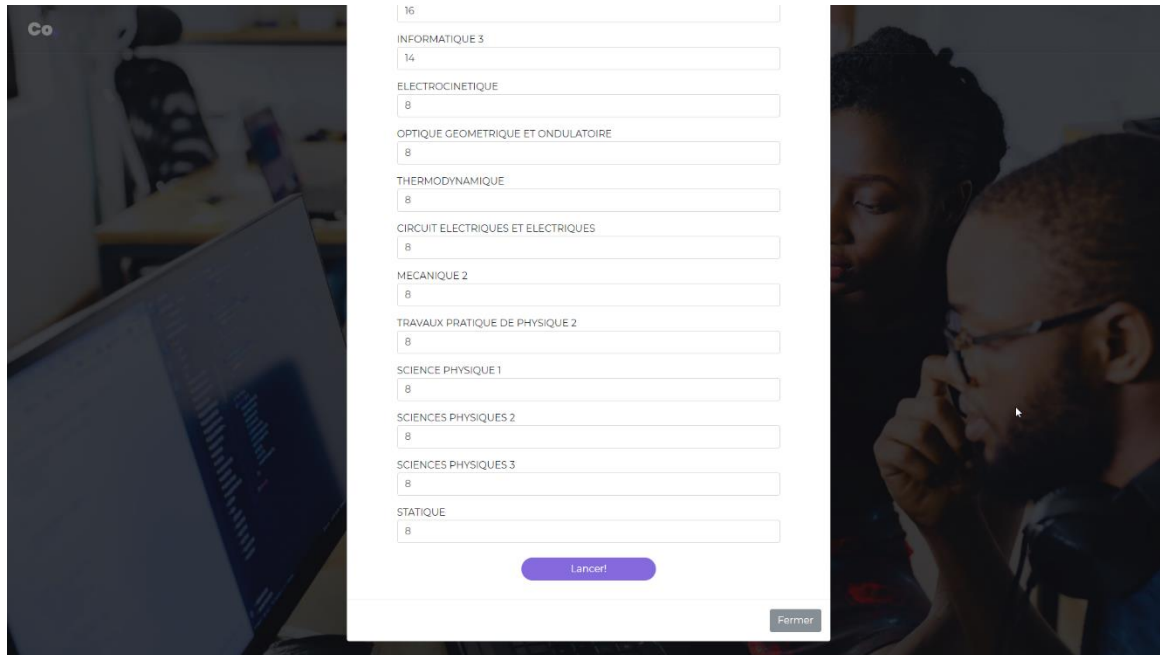
SERIES ET INTEGRALES GENERALISEES  
17

PROBABILITES ET STATISTIQUES  
16

ANALYSE NUMERIQUE  
17

ANALYSE DANS LES ESPACES VEC. DE DIM. FINIE

**Figure 4: Student enters the marks he/she had on papers of the two first years**



16

INFORMATIQUE 3  
14

ELECTRODYNAMIQUE  
8

OPTIQUE GEOMETRIQUE ET ONDULATOIRE  
8

THERMODYNAMIQUE  
8

CIRCUIT ELECTRIQUES ET ELECTRIQUES  
8

MECANIQUE 2  
8

TRAVAUX PRATIQUES DE PHYSIQUE 2  
8

SCIENCE PHYSIQUE 1  
8

SCIENCES PHYSIQUES 2  
8

SCIENCES PHYSIQUES 3  
8

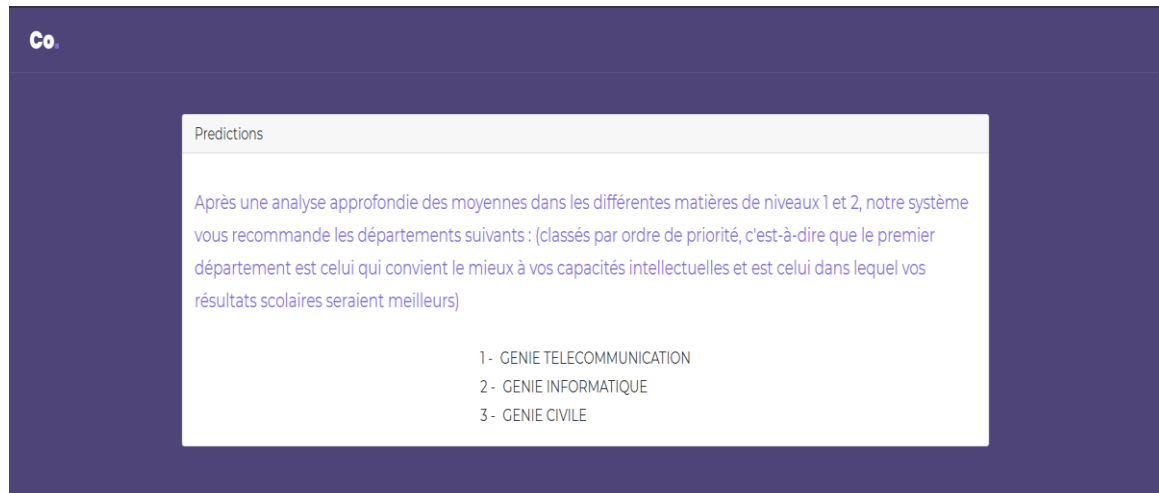
STATIQUE  
8

Lancer

Fermer

**Figure 5: He/she launch the application to see which specialty departments the system suggests him/her**





**Figure 6: The final recommendations of the system listed in order of priority (Genie telecommunication is the specialty department where this student performance might be the best; His performance might also be good in Genie informatique but not as in the first department, same for the third)**

#### 4 Conclusion and future work

In conclusion, this paper presents a system of orientation of the students in department of specialization and proposes to them the three best departments in which they are more likely to maximize their academic results. To achieve this result, we walked through a clear methodology that includes many stages (definition of the problem, data collection, data preprocessing, label encoding, machine learning techniques, training and test and results).

We plan to improve the accuracy of our model by getting more data choosing the right function that gives the label to data input data. We also plan to improve the system by making it more adaptive. In other words, we are going to make the features easily customizable so that the suggestion application can manage both students and high school students. In addition, we intend to integrate a feedback system to give the possibility to the students to give us feedback at the end of their training. This will allow us to see if our system is doing its job well and it will allow us to improve it based on this feedback.

## References

- [1] STUDENT CAREER PREDICTION Idyapriya .C:  
<https://ijcrt.org/papers/IJCRT195700.pdf>.
- [2] Career Guidance System using Machine Learning For Engineering Students (CS/IT):  
<https://www.irjet.net/archives/V7/i6/IRJET-V7I6640.pdf>
- [3] Student Career Prediction Using Advanced Machine Learning Techniques Roy:  
<https://www.sciencepubco.com/index.php/ijet/article/view/11738/4565>
- [4] Student Career Prediction System : [https://github.com/KLGLUG/student-career-area-prediction-using-machine-learning/blob/master/THESIS-STUDENT\\_CAREER\\_AREA\\_PREDICTION/PROJECT%20THESIS.pdf](https://github.com/KLGLUG/student-career-area-prediction-using-machine-learning/blob/master/THESIS-STUDENT_CAREER_AREA_PREDICTION/PROJECT%20THESIS.pdf)
- [5] A Machine Learning Approach for Future Career Planning:  
<http://cs229.stanford.edu/proj2010/LouRenZhang-AMachineLearningApproachForFutureCareerPlanning.pdf>
- [6] Random forest : [https://en.wikipedia.org/wiki/Random\\_forest](https://en.wikipedia.org/wiki/Random_forest)
- [7] Naïve Bayes classifier : [https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)
- [8] CatBoost Classifier : <https://en.wikipedia.org/wiki/Catboost>