Umeluleki - School Counseling

I. Problem and Solution

1. Problem

The problem we have to solve here is the academic orientation of new graduates.

2. Solution

The solution we propose here is an AI called Umeluleki (School Counseling in Zulu) which will take care, depending on the characteristics of each student, of providing him the most suitable field of study.

II. Motivation

The problem related to the orientation of new graduates is a constant obstacle to development. Indeed, in Terminale, most students do not yet know what to do with their lives, or in what area they want to practice. They therefore find themselves doing contests by follow-up and frequently fail. As a result, they find themselves in schools that are not very suitable or often completely opposed to what their qualities predispose them to do because of the lack of assistance. It follows from this poor choice of school or training, a significant drop in the level of the student, the use by the students of fraudulent methods such as cheating or corruption to pass, a complete disinterestedness of the work carried out, and even abandonment which leads to scourges like unemployment, theft, a drop in the level of workers. For all these reasons, it is obvious that the problem of academic orientation is a crucial problem in our African societies. It is therefore important or even vital to remedy this.

III. Similar works

Our work is certainly not the first in the field. There are already several high-performance digital guidance counselors. The main concern with them is that they are made by Westerners therefore adapted to their problems, their priorities and their customs. It is therefore necessary for us Africans to present a similar tool oriented by our own problems.

It is by wishing to respond to this problem of Cameroonian contextualization that we have proposed, in addition to the choice of possible profession, all of the schools where he can train in this profession.

IV. Methodology

1. Choice of characteristics to study

The orientation of an individual is mainly based on his personality. Indeed, the word personality opens quite wide universes. To choose which personality traits we were going to confine ourselves to, we started by listing them as many as possible. To help us in this choice, we started by listing the fields of study available in Cameroon. Then we looked for the qualities required to

excel in each of these areas. Finally, we have cleaned up the features of our first list according to their relevance and similarities. From this cleaning, we emerged with the 42 most important characteristics to take into account in the academic orientation.

3. Investigation

We obtained our data via a survey which aimed to determine each person's personality by asking questions related to the personality traits that were selected in the previous phase. To do this, we have passed the following form trying to take a significant and unbiased part of society. Indeed, being engineering students, it seems obvious that by taking our own friends and close family, there would be an overabundance of certain specific areas. To reduce this bias as much as possible, we have voluntarily shared the link on the form with people who are not very close and sometimes even unknown. The purpose of this form was to gather the maximum number of different profiles by asking questions and then using a guidance counselor to assign them one or more preferred areas. We were able to obtain 151 responses from this survey.

It is important to note that these fields concern the various possible trades followed by the contextualization of the schools in which one can be formed with these trades (for example for Medecine: FMSB, UDM...)

Survey form link:

https://docs.google.com/forms/d/e/1FAIpQLSdDvBNr7eFSQYwdSUnmhbOueShLn_12Zho8BeVmUF_MRbEk8w/viewform?usp=sf_link

4. Choice of model

For the choice of the model, we opted for random forests because it is particularly effective for classification problems with little data (which is our case). To properly understand what a random forest is, you must first know what a decision tree is.

5. Decision Tree

A decision tree is a flowchart similar to a tree structure, where each internal node designates a test on an attribute, each branch represents a result of the test and each leaf node (terminal node) has a class label.

Construction of the decision tree:

A tree can be "learned" by dividing the source set into subsets on the basis of an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. Recursion is complete when the subset at the node level all has the same value of the target variable, or when the splitting no longer adds any value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or any configuration, and is therefore suitable for the discovery of exploratory knowledge. Decision trees cannot handle large data. In general, the decision tree classifier has good accuracy. Induction of the decision tree is a typical inductive approach to gain knowledge about classification.

Representation of the decision tree:

Decision trees classify instances by sorting them in the tree from root to leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then descending the tree branch corresponding to the value of the attribute as shown in the figure above. This process is then repeated for the sub tree rooted at the new node.

Principle of construction:

At the start, the points of the learning base are all placed in the root node. One of the variables describing the points is the class of the point (the "ground truth"); this variable is called "target variable". The target variable can be categorical (classification problem) or real value (regression problem). Each node is cut (split operation) giving rise to several descending nodes. An element of the learning base located in a node will be found in only one of its descendants.

• The tree is built by recursive partition of each node according to the value of the attribute tested at each iteration (top-down induction). The optimized criterion is the homogeneity of the descendants compared to the target variable. The variable which is tested in a node will be the one which maximizes this homogeneity.

• The process stops when the elements of a node have the same value for the target variable (homogeneity).

Homogeneity of data is measured by calculating Gini impurity or entropy,

6. Random forests

Random forests are an improvement of the bagging for CART decision trees in order to make the trees used more independent (less correlated). Bagging is a technique used to classify models known as "weak classifiers", that is to say hardly more effective than a random classification.

Characteristics :

• They give good results especially in large dimensions,

• They are very simple to implement,

• They have few parameters.

Random forest algorithm:

• We randomly draw from the learning base B samples with discount z_i, i = 1, …, B (each sample having n points).

• For each sample i, we build a CART tree Gi (x) according to a slightly modified algorithm: each time a node must be cut (“split” step) we randomly draw part of

the attributes (q among the p attributes) and we choose the best division in this subset.

- Regression: aggregation by the mean
- Classification: aggregation by vote G (x) = Majority vote ($G_1(x)$,..., $G_B(x)$)

Trees are less correlated because:

- They are learned on a different set of attributes.
- They are built on different samples.

Comments:

- We generally limit ourselves to not very deep trees (for Bagging we need deep trees to reduce their correlation, but very deep trees suffer from over-learning).
- Each tree is small therefore less efficient, but the aggregation compensates for this failure (each attribute is typically found in several trees).
- As for Bagging we use the OOB error to prevent over-learning (we choose B where the error stabilizes and no longer descends).

The OOB (Out Of Bag) error is given by the average of the errors of the Gi classifiers such as $x_i \notin z_i$

Parameters (default values):

- Classification: $q=\sqrt{p}$,minimum knot size 1
- Regression: q = p / 3, minimum node size 5.

In practice, the "ideal" values depend very much on the base (and they must be found by cross-validation).

Regarding our own model, we no longer rewrote the random forest algorithm, we just used the one found in the scikit-learn (sklearn) library of the python language. In this library, the classifier named ‘‘ RandomforestClassifier ’’ is a class. We therefore used its attributes and methods to create our model and measure its accuracy. It turns out that our model (program) predicts the area best suited to an individual with an accuracy of 78%.

Conclusion

In short, our work focused on solving the problem of the academic orientation of new graduates who is the source of many scourges in our society and is a constant obstacle to the development of Africa in general and Cameroon in particular. To do this, we used a Machine Learning program, more precisely a

random forest program. Said program was trained with data collected via a form survey and obtained an accuracy of 78%.

For an optimization of our program, the first problem to solve is that of the lack of data. We have judged after having seen and analyzed several projects similar to ours that it would require an unbiased (even balanced i.e. where all the labels are in same quantity) bagatelle of 1500 data and if possible a team of experts of guidance (advisers of 'guidelines) to label them.