Deliverable 5
Phase 4

CSI4142 : Fundamentals of Data Science
winter 2023

uOttawa

Faculté de génie
Faculty of Engineering

School of Electrical Engineering and Computer Science
University of Ottawa

Professeur: Yazan Otoum, Ph.D

Group members :
Souleymane Wilfried Sankara, #300100940
Marie Fabienne Sawadogo; #300101795
Tata Saidatou Berte, #300100935

PART A

The goal of our project is to predict the job classification (data scientist, data engineer, data analyst, etc…) given some information about the hiring company : company name, company's industry, the rating,type of ownership, size range and the location and the average salary.

We used mostly **countplot data** visualization since most of our data are categorical data.

- Countplot of job classification according the size range:

We noticed that all the size range categories hire a lot of data scientists, but the distribution is different from one size range to another. So we decided to keep it for the classification task

- Countplot according the industry:

We noticed that the job classification can be somehow depending on the sector of activity, for instance health care services hire more researcher while aerospace companies hire more data scientist

- Countplot according to state, type of ownership and company name :

In all of these cases the job classification depends on these data. We decided then to use them for the classification the type of data scientist jobs

We used a **barplot** to analyze the impact of the company rating on the job classification and we noticed an almost uniform distribution, which implies that the rating doesn't have a significant effect on the job classification, so we decided then to not use it.

We also used a barplot to analyze the distribution of the average salary given the job classification, and we noticed a real huge difference between the different type of jobs, so we decided to keep it for the classification task.

After the data visualization, we decided to only use the : company name, industry, location, type of ownership, size range and the average salary to get a model that can classify the job type.

For the classification task almost all of our data are categorical. We decided then to convert them into numerical data using one-hot encoding since we're going to use random forest, decision tree and gradient boosting.

.

Part B

Based on the results obtained, it appears that the models are not able to accurately predict the 'Job Title' attribute. The precision is very low for the Random Forest and Gradient Boosting algorithms, indicating that they tend to predict a large number of false positives. The precision of the Decision Tree Classifier algorithm is slightly better, but still relatively low.

The recall is also quite low for Gradient Boosting, meaning that this model tends to miss a lot of true positives. The recall for Decision Tree Classifier is a bit higher, meaning that this model tends to detect more true positives, but it may also include many false positives.

Overall, these results suggest that the 'Job Title' attribute is difficult to predict from the other attributes in the database. It is possible that the use of other algorithms or different data preprocessing techniques could improve the performance of the models. It is also possible that the addition of new variables to the database could improve the performance of the models.

We can conclude that our data is not sufficient to predict the job title using the given attributes.

Algorithm	Accuracy	Précision	Recall	Time
Random Rorest	0.315	0.132	0.315	3.857
Gradient Boosting	0.239	0.0063	0.014	88.415
Decision tree Classifier	0.103	0.099	0.130	0.359

Table1: Metrics of classification.