How to find a good paying job in the AI/ML and Big Data space in 2023

Data Mining Final Project

GROUP 3 (ADA LOVELACE)

Bibiloni Fontirroig, Josep Antoni   Comellas Fluxá, Cristian
Crespí Valero, Maribel   Fortes Domínguez, Odilo   Meneses Magon, Jhonier Duvan

2023-01-31

Contents

1 Introduction **2**

2 Data analysis **2**
2.1 First glance 2
2.2 Data and features selection 2
2.3 Cleansing and pre-processing 2
2.4 Feature addition 3
2.5 Summary 3

3 Questions **4**
3.1 Q1: How does location influence salary? 4
3.2 Q2: Can we cluster data to discover any patterns? 5
3.3 Q3: Which factors help the most to get high salaries? 6
3.4 Q4: What are the most influential characteristics? 7
3.5 Q5: Based on our characteristics, what salaries would we have? 8
3.5.1 Model selection 8
3.5.2 Validate model and our predictions 9

4 Conclusion **9**

1 Introduction

This project aims to discover the factors or patterns that can help find a well-paying job in the field of Artificial Intelligence, Machine Learning and Big Data. It is focused on extracting information from the past year, which is considered the best approximation for the next year, in which you want to find a job. Specifically, it seeks to discover how to find a full-time job.

To do so, different Data Mining techniques will be used, including both supervised and unsupervised learning. These cover classification techniques such as Decision Trees and Naive Bayes. Clustering techniques, such as k-modes, and Association rules with Apriori method, are also employed. All these methods will help to answer a series of questions, the answers to which will provide us with the necessary information to obtain the knowledge we are looking for.

2 Data analysis

In this section, we will analyze the data and make the necessary cleaning and changes to prepare it for the methods we will use later.

2.1 First glance

In first place, we read the data converting strings to factors and then we check if it is properly loaded. We also inspect now the structure of the dataframe.

```
## 'data.frame':    1332 obs. of  11 variables:
##  $ work_year         : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ..
##  $ experience_level  : Factor w/ 4 levels "EN","EX","MI",..: 3 3 3 3 3 3 3 4 4 4..
##  $ employment_type   : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3..
##  $ job_title         : Factor w/ 64 levels "3D Computer Vision Researcher",..:..
##  $ salary            : int  130000 90000 120000 100000 85000 78000 161000 1100..
##  $ salary_currency   : Factor w/ 18 levels "AUD","BRL","CAD",..: 18 18 18 18 1..
##  $ salary_in_usd     : int  130000 90000 120000 100000 85000 78000 161000 1100..
##  $ employee_residence: Factor w/ 64 levels "AE","AR","AT",..: 63 63 63 63 63 6..
##  $ remote_ratio      : int  0 0 100 100 100 100 100 100 100 100 ...
##  $ company_location  : Factor w/ 59 levels "AE","AL","AR",..: 58 58 58 58 58 5..
##  $ company_size      : Factor w/ 3 levels "L","M","S": 2 2 2 2 2 2 2 2 2 2 ...
```

There are a total of 1332 observations and 11 variables in the initial dataset. We can see that most of the variables are factors. In fact, the only numerical variables are the salary ones.

2.2 Data and features selection

We are interested in extracting knowledge from the most recent year, since we assume that it is the one that will most resemble 2023. Therefore, we remove data from the remaining years. We can do this because most of the observations are from 2022. In addition, we want to find full-time jobs, so we can also remove samples with other types of employment. Again, we can do this since most of the jobs in the dataframe are full-time. Because of these deletions, the variables `word_year` and `employment_type` only have one value, so we can remove them. Furthermore, as far as wages are concerned, we are only interested in the variable `salary_in_usd` since we need salary to be comparable regardless of the currency in which they are paid. For this reason, we get rid of `salary`.

2.3 Cleansing and pre-processing

First, we have decided to convert the remote ratio variable to a factor with order, since it behaves logically as such, but has a discrete numerical type. It will have the values "in-place", "partial" and "remote".

Then, we realized that the `job_title` variable had different values that meant the same thing, for example, "ML Engineer" and "Machine Learning Developer". Because of this, we had to do a manual filtering, merging this kind of groups into a single value. In the case of the example, both values end up being "Machine Learning Engineer".

We also discovered that there are a total of 9 outliers in the `salary_in_usd` variable, although we will not eliminate them because we are interested in discovering what factors allow us to obtain such high wages.

2.4 Feature addition

The first variable we add is the salary converted to an ordered factor. This will group the salary into three, the salaries below the first quartile ("LOW" is less than 90,000), those between the first and third ("MEDIUM") and those above the third quartile ("HIGH" is more than 172,500).

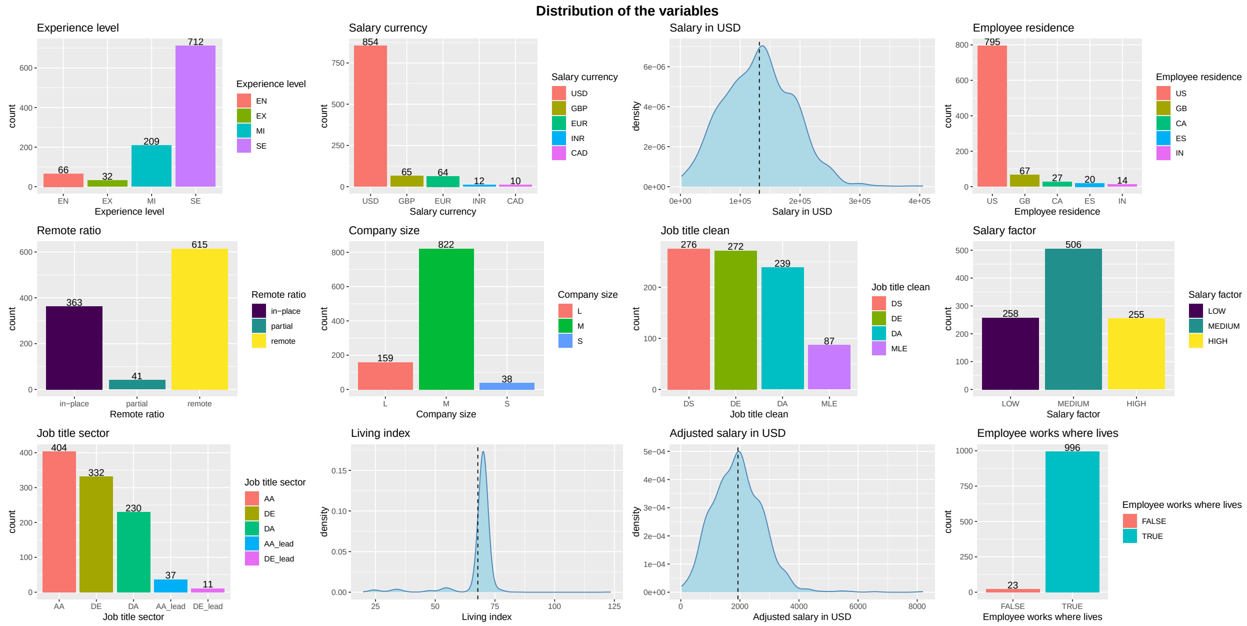
Second, we will group the different job titles according to their purpose. The new variable will have a total of six values: Advanced Analytics (AA), Data Engineering (DE), Data Analytics (DA) and their leaders, AA_lead, DE_lead and DA_lead.

Then, we added four variables containing the latitude and longitude of the employees' residence and the location of the companies. These data are extracted from another Kaggle dataset containing latitude and longitude by country code. We found that we have a sample with a code "AX" that is not found in the dataset, so we enter the data concerning this country manually. Due to the fact that we are replacing each factor by a concrete value, this variable can be considered as a factor logically, with the advantage that distances can be easily calculated between them. Related with that, we will add a binary variable that will indicate whether a worker works where he/she resides or not.

Finally, we will add a variable containing the salary adjusted for the cost of living in the employee's country of residence. This cost of living is from 2022 and it has been extracted from Numbeo.

2.5 Summary

Once the pre-processing is finished, we end up with 1019 samples and 17 variables. We can now inspect it for strange distributions or unbalanced factors. We note that most of the factors are unbalanced, which may pose a problem for subsequent analysis. In the case of the factors with more than five levels, only the five with the highest frequency are shown, ordered from highest to lowest. On the other hand, we see that the salary follows an approximately normal distribution, although there is a tail at the end, which represents outliers with salaries around 300,000 and 400,000 USD. Moreover, since most of the data is from USA, we don't see any change on the shape of the adjusted salary distribution.



We will review in more detail the variables related to latitude and longitude as we answer the questions.

3 Questions

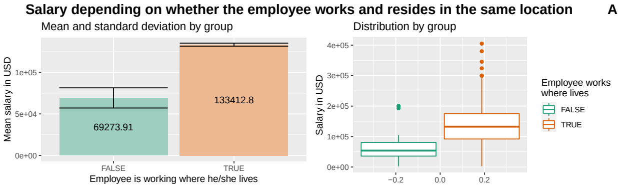
In this section, we will answer a series of questions to dig into the data and extract information that will be useful to gain knowledge and develop the best advice for finding a well-paying job next year (2023).

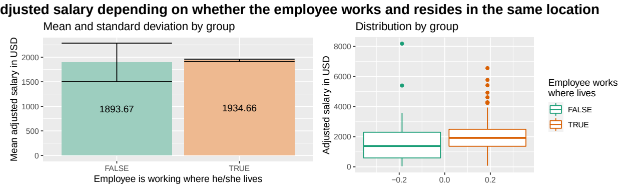
First, we are interested in determining whether the worker's location can affect his or her salary. We also want to find if there is any way to group workers more naturally from the data. In addition, we want to find associations between different factors that tell us which characteristics are more present in those workers with the highest salaries. Finally, we want to know what salary we could obtain next year with our current characteristics.

We will use different methods to answer these questions. Among these, we will employ more traditional methods such as the inspection of graphs and the study of different statistics. On the other hand, we will employ another more automated perspective, using techniques such as clustering, associative rule mining, and some classification methods, such as decision trees, k-nearest neighbors, and Naive Bayes.

3.1 Q1: How does location influence salary?

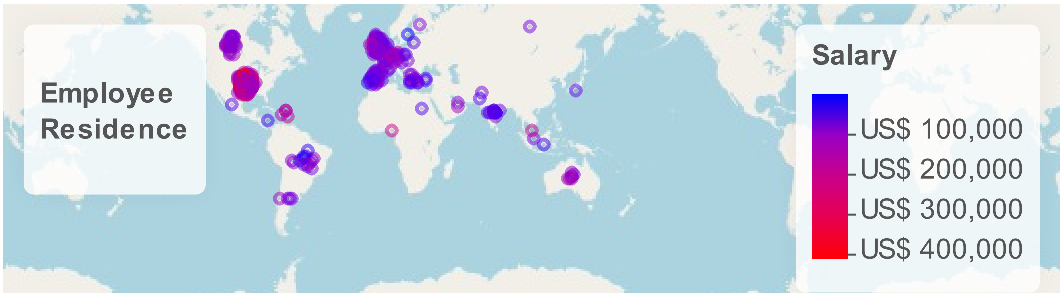
In this section we will use basic statistics and plots interpretation to answer the question. To begin with, we will check if there is any difference between the group of those who work in the same place they reside and those who do not. We do this because we want to know if either of the two is paid more than the other, because there are few workers who reside elsewhere than where they work, and if they were paid more it could be a good indicator for high salaries.





We observe that there is a clear difference between the salary in the group of those who reside in the same place where they work and those who do not, with the latter set of workers earning the least. We see this because the mean of the different groups does not fall within the range established by the mean plus minus the standard deviation of the other group. Furthermore, looking at the boxplot, we see that practically 100% of the samples are below 75% of those working where they reside, except for two outliers.

This fact could be due to the fact that employees working abroad are living in places with a lower cost of living, so the salary is regulated according to this factor. To check this, the same graphs have been shown with the adjusted salary. In this case, we observe that there are no longer significant differences between the two groups, indicating that salaries are adjusted according to the cost of living. Because of this, we see that whether you work where you live or not has no real impact on purchasing power.



We will now look at salaries in different countries using a Leaflet map. To better see the salaries, some normal noise has been added to the coordinates of each sample so that all the points do not overlap in one place. In addition, no differences were observed in the case of adjusting wages to the cost of living, so only the wage in USD is shown. The map shows that the highest salaries are in the United States. Good salaries are also found in Canada and Great Britain. In the rest of the world, some high salaries can be found in localized areas, such as Nigeria, but we have too few samples from certain regions to draw conclusions.

According to the results obtained in this section, in general, it is advisable to work for a company in the united states if you want to obtain a salary well above the average. It is possible to work from the same country or from abroad, but it must be remembered that the salary will be adjusted to the cost of living in the country of residence.

3.2 Q2: Can we cluster data to discover any patterns?

Clustering algorithms are specifically used to discover patterns and relationships within the data that may not be obvious from a simple visual inspection of the raw data. The application of this algorithm may result in a better understanding of the data. The results will then allow us to add a new (categorical) feature to the dataset, where for each row it will indicate the obtained cluster it belongs to.

First of all, we will have to carefully select the features to work with, as with too many features the clustering algorithm may have difficulties finding meaningful patterns.

We will delete all coordinates related variables except the company ones (as most workers live in the same country they work at), all salary related except salary_in_usd because it is the one that gives us the most information of all, and job_title_clean because it is too diverse.

The variables type is a key factor that affects which algorithms can be applied. In this case, we have left a similar quantity of categorical and numerical features (4 and 3 respectively) so we have decided to use k-prototype as it is specifically designed to handle mixed data types. It will allow us to cluster by taking into account the unique characteristics of both feature types.

The next step we need to do is select the optimal number of clusters (k), and to do so we can use the elbow method. The idea behind the elbow method is to plot the explained variation as a function of the number of clusters, and pick the elbow of the curve as the number of clusters to use. The plot for optimum k will be shown at the end of the section. If we take a look at the elbow it can be seen that we could either pick 4 or 5 as our number of clusters, because it represents the point where the increase in number of clusters does not yield much improvement in terms of WCSS or variance explained and avoids overfitting. In this case we will pick 4, but 5 would have been a good election as well. We will then execute the algorithm setting k=4.

```
optimum <- 4; set.seed(42); kpres <- kproto(data2[,c(-1, -2)], optimum, verbose = FALSE); kpres;
```

Now we will review the results. We know that the Euclidean distance is used to calculate the similarity between observations. The results show that the the number of observations of the clusters are 143, 203, 647 and 26. In addition, the within cluster error is 415109, 328329, 1208270 and 122392, respectively. This indicates how similar the observations are within each cluster. Lower values of this metric mean that the observations within a cluster are more similar.

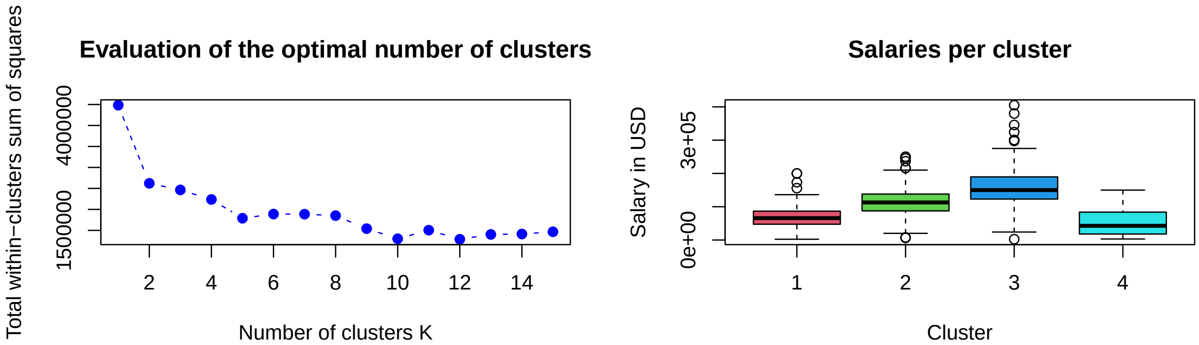
Based on the within cluster error values provided, it can be concluded that the clustering algorithm has done a relatively good job of grouping similar data points together. We can observe that cluster 2 has a low error for its size so its members must be pretty similar, cluster 3 has a very high error because it has a lot of members, and cluster 4 has a high error despite of it having very few members.

Now let's add the cluster column to the dataset with the new feature, then visualize and inspect the data. We will also plot the distribution of the salary in USD for every cluster to analyze them.

Clusters	Freq	Experience level	Remote ratio	Company size	Job title sector	Company location latitude	Company location longitude	Living index	Salary in USD
1	143	MI	remote	M	AA	49.449505	2.094679	64.09888	69956.99
2	203	SE	remote	M	DA	35.194639	-94.432672	67.58463	116174.46
3	647	SE	remote	M	AA	37.252614	-95.720383	69.55080	153844.38
4	26	EN	remote	L	AA	5.083435	96.904480	47.88385	51841.35

Overall we can see that cluster 3 earns significantly more than the rest (they earn US$ 150k aprox). Cluster 2 earns close to US$ 40k less than cluster 3 despite the only differences being that cluster 2 sector is data analysis while cluster 3 is advanced analytics. We can also observe that not living in the US makes a big impact to the income (clusters 1,4 vs clusters 2,3), taking into account that samples from cluster 1 are closer to Europe and samples from cluster 4 are closer to Asia.

This adds to the fact that to get a good paying job in this sector we should try to aim for a job in the US, specifically in the advanced analytics sector.



3.3 Q3: Which factors help the most to get high salaries?

In the initial stage of our Apriori analysis, it is necessary to transform the original dataset into a transactional format. This involves converting any non-factor values, such as salary, into categorical variables. In this case, we will categorize salaries into three distinct groups: HIGH, MEDIUM, and LOW. The specific salary values themselves are not of interest in this analysis.

Our objective is to secure top-paying positions in the field of Machine Learning and AI. To this end, we have conducted an analysis to uncover the factors that play a role in determining high salaries in this industry. The results of this analysis provide valuable insights for those seeking to secure high-paying positions in the field of Machine Learning and AI.

```
rules_high_salary = apriori(data = dataset,
          parameter = list(support = 0.001, #Support small value since you think the data set is relatively large.
          confidence = 0.7,target = "rules"), #Confidence value of 0.7 for strong rules.
          appearance = list(rhs = "HIGH")) #In rhs we are interested in the "HIGH" value of salary.
 inspect(sort(rules_high_salary[size(rules_high_salary) > 6], by = 'lift')[1:15])
 # We use is.maximal since we want only the most general rules.
 reglas_maximales <- rules_high_salary[is.maximal(rules_high_salary)][1:7]
```

In general, it appears that remote jobs and companies located in the US that pay in USD are the most common factors associated with high salaries. It is also important to note that face-to-face jobs can also offer high salaries. When it comes to the most well-paying jobs, the positions of Applied Data Scientist within the AA context, and large companies, as well as having senior experience, are key factors.

An analysis of the factors influencing high salaries in small, medium and large companies respectively shows differences in job titles such as Applied Data Scientist, Machine Learning Scientist, Machine Learning Engineer, DE_lead and Data Science Manager.

Regardless of company size, experience and seniority, location in the US and payment in USD appear to be important factors. The highest paying sector in small and large companies is AA, while DE_lead is the most prominent sector in medium-sized companies. No significant difference is observed between remote and face-to-face work.

For smaller companies, the highest paying jobs are Machine Learning Scientist and Machine Learning Engineer, for medium-sized companies it is Data Science Manager and for large companies it is Applied Data Scientist.

Finally, an analysis of factors affecting high salaries in both remote and face-to-face jobs shows that senior experience, location in the US and payment in USD are important factors. The highest paying remote jobs are Applied Data Scientist, AI Scientist and Machine Learning Engineer/Scientist, while the highest paying face-to-face job is Data Science Manager in the AA_lead sector.

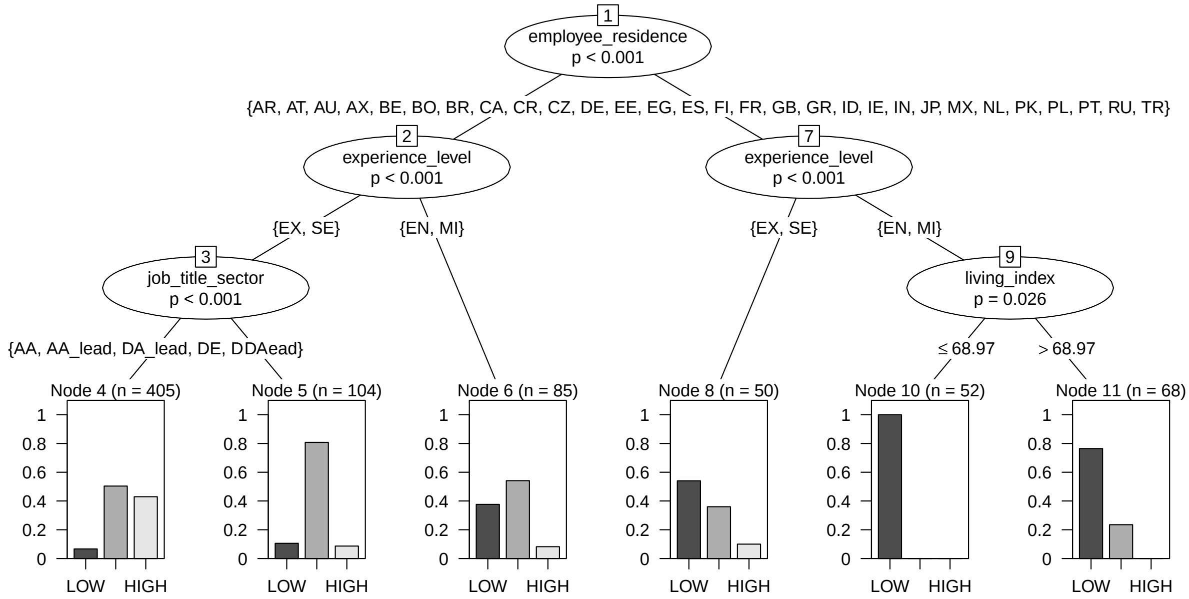
3.4 Q4: What are the most influential characteristics?

The process uses the salary dataset, which has information on job type, work experience, geographic location, etc., to train a decision tree model. The model finds patterns and relationships in the data to identify key factors in predicting AI/ML and Big Data salaries.

We will use the ctree algorithm from the party package in R, which builds a robust, reliable conditional inference tree based on non-parametric tests. It is effective for datasets with many predictors, and can handle continuous and categorical variables, making it suitable for analyzing AI/ML and Big Data salaries.

```
##
##   Conditional inference tree with 6 terminal nodes
##
## Response:  salary_factor
## Inputs:  experience_level, living_index, job_title_clean, employee_residence, remote_ratio, company_location, company_size, job_title_sector
## Number of observations:  764
##
## 1) employee_residence == {AE, MY, NG, PR, US}; criterion = 1, statistic = 247.049
##   2) experience_level == {EX, SE}; criterion = 1, statistic = 77.696
##     3) job_title_sector == {AA, AA_lead, DA_lead, DE, DE_lead}; criterion = 1, statistic = 61.722
##       4)*  weights = 405
##     3) job_title_sector == {DA}
##       5)*  weights = 104
##   2) experience_level == {EN, MI}
##     6)*  weights = 85
## 1) employee_residence == {AR, AT, AU, AX, BE, BO, BR, CA, CR, CZ, DE, EE, EG, ES, FI, FR, GB, GR, ID, IE, IN, JP, MX, NL, PK, PL, PT, RU, TR}
##   7) experience_level == {EX, SE}; criterion = 1, statistic = 63.16
##     8)*  weights = 50
##   7) experience_level == {EN, MI}
##     9) living_index <= 68.97; criterion = 0.974, statistic = 24.211
##       10)*  weights = 52
##     9) living_index > 68.97
##       11)*  weights = 68
```

After training a decision tree, the most important factors are residence (company location and employee location are practically the same), experience, job sector, and living index. Location is considered the top factor, followed by experience and job sector. The importance of each feature may vary with the dataset and problem at hand. It is essential to remember that the feature importance may change depending on the dataset used to train the decision tree. If we plot the decision tree, we can see that to increase salary, the company must be located in the US, UAE or Malaysia, Puerto Rico and Nigeria (which are outliers), and the candidate must have strong experience in AA or DA. Lower paying jobs are in other countries with limited experience and a living index below 68.97.



The accuracy of the decision tree, which is 63.53%, representing the proportion of correct predictions made by the model. Accuracy can be helpful, but it may not fully reflect a model's performance, especially for imbalanced datasets.

It has been noted that the location of the company or employee has the most impact on salary, followed by work experience, job type and living index. In particular, the country where the company is located greatly affects salary.

3.5 Q5: Based on our characteristics, what salaries would we have?

The objective of this question is to train a model that allows us to know in which salary range a worker is. We can use this model later to predict ourselves and get an idea of the salary we could have based on our characteristics.

First, we will select the characteristics of the model. We will remove from the previous dataset `salary_in_usd` and `adjusted_salary_in_usd`, since it is a classification model and therefore we will use "salary_factor" as target.

We will also remove `company_location`, `company_location_longitude`, `company_location_latitude`, `employee_works_where_lives` and `employee_residence` since the `employee_residence_latitude` and `employee_residence_longitude` variables provide the same information. Finally we will scale the numerical data of the dataset.

3.5.1 Model selection

We are going to select the classification model. To do so, we are going to split the data into 80% train and 20% test. We will train different models and choose the one that gives us the best results. It should be noted that the models reviewed in class such as knn and Naive Bayes do not give us good results, so we will use other unseen models apart from these.

Let's see the models that we trained. We have used **Naive Bayes** with the default hyperparameters and **k-Nearest Neighbors** (with kmax=5). We have also used other unseen models that we will briefly explain below.

- **Support Vector Machines** (kernel=linear, cost=1): it is a supervised machine learning algorithm used for classification and regression problems. It aims to find the best boundary between classes by maximizing the margin between classes, which is the distance between the boundary and closest data points of each class. The closest data points are known as support vectors and have the greatest impact on the boundary. SVM can handle non-linearly separable data by transforming them into a higher dimensional space mediante un kernel where a linear boundary can be found.
- **Random Forest** (default): this is 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 prediction (regression) of individual trees. The trees are constructed using bootstrapped samples of the data and a random subset of the features for each split, adding to the randomness of the model. The combination of many trees in a random forest leads to a decrease in overfitting and an increase in accuracy compared to a single decision tree.
- **XGBoost** (objective="multi:softmax, booster="gbtree"): Xgboost is a library that implements gradient boosting algorithm on decision trees and has been used in various real-world applications. XGBoost is known for its speed and performance and often used for large-scale data science projects.

Let's see the results of each model:

Metric	Naive Bayes	K-NN	Linear SVM	Random Forest	XGBoost
precision	0.539	0.598	0.667	0.672	0.545
error	0.461	0.402	0.333	0.328	0.455

We can see that KNN and Naive Bayes do not give us good results. Xgboost is more complex and its results are regular for having a small dataset. SVM and random forest give us better results. We will select random forest as it is the best.

3.5.2 Validate model and our predictions

Once the model is selected, to give more robustness to the metrics, we will use cross validation with 10 folds. It gives us a similar result (64.71% precision), therefore it is not a coincidence. The final model used to predict gives us a precision of 67.65%.

Finally, we are going to use the model to predict ourselves:

Name	Experience level	Salary currency	Remote ratio	Job title clean	Company size	Salary predicted
Cristian	MI	USD	partial	Machine Learning Engineer	M	**MEDIUM**
Pep Toni	MI	EUR	partial	Machine Learning Engineer	L	**LOW**
Odilo	MI	USD	in-place	Machine Learning Engineer	M	**MEDIUM**
Maribel	MI	GBP	remote	Data Scientist	L	**MEDIUM**
Jhonier	MI	EUR	partial	Data Engineer	L	**LOW**

4 Conclusion

Having answered the above questions, we have gained the necessary knowledge to build a guide on how to get a good paying job in 2023 in AI/ML and Data Science space.

Generally, salary is largely influenced by a company's or employee location, with factors like work experience, job type, and cost of living also playing a role. The country where the company is based has a big impact on salary, with higher salaries typically seen in companies based in the US or UAE. To boost pay, a candidate should target jobs in these countries and have extensive experience in Advanced Analytics (AA) or Data Analytics (DA).

In particular, we know that it is beneficial to learn English, since the best way to get a higher salary is to work for a company in the United States or other English-speaking areas such as Canada and Great Britain. In general, both remote and in-place jobs are well paid, but there are higher salaries on remote, so it could be convenient to find a remote job. Even so, it is important to take into account the cost of living in the country where you reside, as it will affect your remuneration. Regarding the type of work, the AA sector has the highest salaries, with Applied Data Scientist and Machine Learning Engineer being common roles, where you would be interested in working to earn more money. A curious fact that we found out is that US seniors working remote and in medium companies in the AA sector earn close to 50% more than their Data Analyst counterparts. Finally, although it is obvious, it is also important to stress the importance of gaining experience in order to get a better salary, since the salary increase between mid-level and senior level is enormous. Because of this, the salaries predicted for us by the model are low to medium, as we lack experience.

It is important to keep in mind that, in order to improve the results, it would be useful to look for another data set. This is because the data used is very unbalanced, with the vast majority of the samples being from U.S. workers, which may bias the results. Also, having a dataset with more numerical variables could help us when using other algorithms. In the future, it would also be interesting to use more advanced methods, such as neural networks.

From a personal point of view as students, we consider that this practice is a good way to finish the course, since it forces you to review all the material and to explore the data in every way you can think of to extract as much information as possible from it. The most difficult points for us have been dealing with the data, since they present several difficulties such as how unbalanced the classes are or having practically only factors, and the fact of presenting the whole study in only eight pages.