UIB

Universitat
de les Illes Balears

Escola Politècnica Superior

Degree in Computer Engineering

Data Mining - 21746

Final Project - AI/ML Salaries

Professor:

Mónica Jennifer Ruiz Miró

Students:

Dawid Michal Roch Móll

Jonathan Salisbury Vega

Joan Sansó Pericàs

Joan Vilella Candia

Julián Wallis Medina

Introduction

Understanding salary trends in the field of Artificial Intelligence and Machine Learning is important for both individuals and organizations. For individuals, understanding potential salaries can help in making informed career decisions and negotiating fair compensation. For organizations, understanding the market value of different roles can improve the hiring process and compensation decisions.

Our project aims to provide valuable insights and recommendations to those seeking to navigate the AI/ML job market by analyzing salary data depending on different roles such as engineers, scientists, analysts, etc and examining the factors that influence salary such as location, company size, and years of experience.

In addition to this, we'll be applying different techniques like classification, prediction, and clustering using various models like multi-linear regression, k-modes, decision trees, etc.

As seniors, we are soon to be entering the job market and this project is an opportunity for us to gain a deeper understanding of the field and the job market around it. As a result, our goal is to give valuable recommendations for individuals (and organizations) seeking to navigate the AI/ML job market.

Data Analysis

Raw data (1332 data entries with 11 different variables)

Variable name	Values	Comments
work_year	2020, 2021 and 2022	2022: 77.2%, 2021: 17.2% 2020: 5.6%
experience_level	EN, MI, SE and EX	SE: 60.3%, MI: 25.3%, EN: 11%, EX: 3.4%
employment_type	PT, FT, CT and FL	FT: 98%, PT: 1%, CT: 0.6%, FL: 0.4%
job_title	64 unique job titles	Some are duplicates for example “Machine Learning” and “ML”
salary, salary_in_usd	Continuous	
salary_currency	18 different currencies	
employee_residence	64 different countries	Too many unique countries. It’s also unbalanced since US represents 68.31% of the data
company_location	59 different countries	Same as above
remote_ratio	0%, 50% and 100%	100%: 58.6%, 50%: 20.4%, 0%: 31%
company_size	S, M and L	M: 67.4%, L: 24.2%, S: 8.4%

Data pre-processing

Data cleansing:

We have ensured that our dataset does not contain any missing values (NA) and have thoroughly checked for outliers. We have decided to not delete any data entries that might be considered as outliers, as they may represent genuine variations in the population's distribution.

Data transformation:

We have deleted, modified and added different columns to improve our analysis. For example, we deleted the column "salary" since we already have "salary_in_usd" that allows us to do a better comparison. We also modified the column "remote_ratio" from 0, 50 and 100 to "Office", "Hybrid" and "Remote" for improved interpretability.

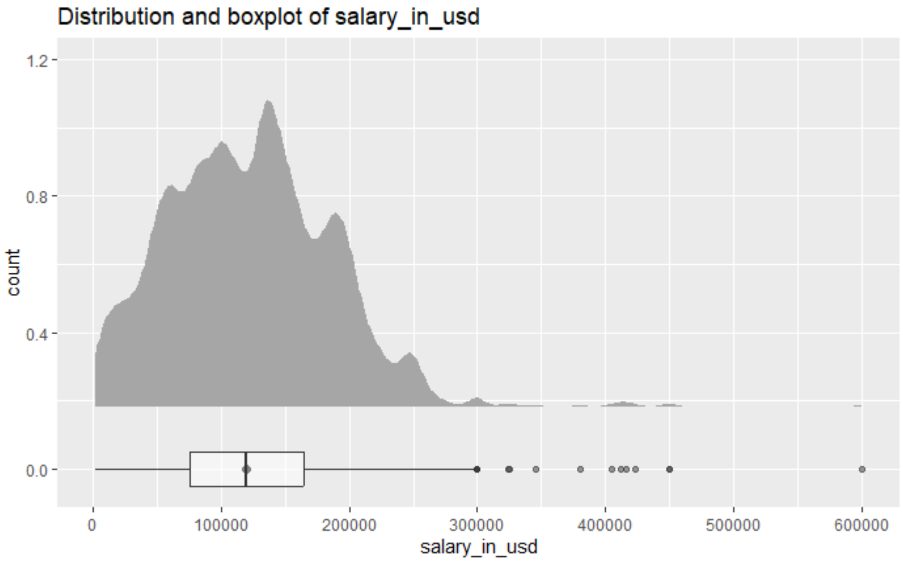
We created a new column "salary_group" based on the "salary_in_usd" variable, where we grouped the salary into 9 different categories (0-25k, 25k-50k, etc.). Additionally, we grouped the "employee_residence" and "company_location" columns into 7 regions (the 7 continents) due to the large number of unique values.

As we can see in the previous table the "job_title" column has 64 unique values. To better analyze the data, we created 4 new variables: "industry" (AI/ML, Data, and Other), "role" (Engineer, Analyst, etc.), "boss" (yes/no), and "research" (yes/no).

In order to enhance the flexibility of our analysis, we have chosen not to select any subset of the data at this stage. Instead, we will have the option to select specific subsets of the data while addressing each individual research question.

Graphical representation:

Given that the majority of our research questions revolve around salary, it is essential to investigate the distribution of salary in our dataset. The following histogram depicts the distribution of salary in our dataset. As we can see, the average salary is around 123k, with some data entries reaching as high as 600k. The first quartile (25th percentile) is at 75k, while the third quartile (75th percentile) is at 165k. This suggests that the majority of the data falls within the range of 75k to 165k, with a few outliers above 300k. It is also clear that the salary distribution is slightly skewed to the right, indicating the presence of some high-income earners



Research Questions

Profile-based prediction and maximization of compensation

For this question, we have **two main objectives** to try to accomplish. The first one is: having the profile of our teacher, can we try to **predict her salary**? And the second one would be: while leaving some parameters fixed (like residence, or experience level), which **attributes** would our profiles need to choose to **maximize their salary**?

With that in mind, we thought that the model that would fit best for these two questions is the Decision Tree.

We create a Regression Decision Tree using the *rpart* library with the “anova” parameter (for regression). We also tried to do another model for classification, where the prediction variable would be a salary group (class) instead of a number, but the model would not converge due to the high number of levels in certain factor variables, so we had to stick with the regression Tree.

Since we want the Tree to be deep (we care about the prediction being the closest to the real value, even if the tree is quite big), we set the **Complexity Parameter** to a really low value of 0.0001. With this, we get a tree that does not generalize (the average error is about the same but we have more levels for prediction). After experimenting with the model, we saw that even with a low **CP**, the model was not using some of the variables, so we decided to *remove* them from the train and test sets so that the Maximizing salary part would be easier to read (less variables = smaller grid = more readable results).

With this, we get a model that has a mean error of ~$3,500 and a standard deviation of ~$57,000. The mean being low means that the model is not biased towards predicting higher or lower values than the real value (since the mean could be considered close to 0 taking into account that the average salary is 123,000), but still is not very accurate since the standard deviation is quite high.

With this out of the way, we then try to predict the salary of our professor Monica. With the available attributes given by her, the salary that this model predicts is **$40,856.43 (37,577.70€)**.

When looking for the best parameters to maximize the salary, the following strategy has been followed. Create a grid with all possible combinations of the possible attributes (taking into account those that are fixed) and then use these parameters to make the prediction. Once all the salaries have been obtained, the highest salaries are selected and the parameters with which this result was obtained are searched. This may present a problem, since if very few parameters are set or if it is a very large dataset, the calculation would be impractical. That is why it is mandatory to make a good study of the parameters and leave as few attributes free as possible. The remaining attributes that have not been used for prediction have been discarded as they were not relevant to the decision tree. Such as RESEARCH or BOSS, there was no rule associated with them.

Wallis' profile is that of a recent ML graduate who wants to live in the United States. He would also like to work in the sector he has studied. This leaves the attributes of: remote work (living in the US) and company size. The results show the following for Wallis: remote work does not influence salary maximization, the only important requirement is **not to work in a large company** (there is no difference between medium and small). The maximum predicted salary is: **$102330.8 - (94331.60€)**.

Dawid's profile is of a recent graduate student with some experience. His requirements are to work from Spain and only remotely. So the attributes that remain to be determined are: Company location and company size. In this case, a rule has not been found to be as decisive as in the previous profile. The size of the company does not seem to be as relevant, since there are practically the same number of each type. It's a little bit the same with the company's location, there are all kinds of options. So in this case the conclusion would be that if you work remotely, you can prioritize secondary aspects such as whether you like the schedules they have or the work structure. The maximized salary will be around **$109.338,5 - (100.791,51€)**.

How can data mining techniques be used to identify which employees are at risk of leaving their company?

Or otherwise, how can we identify compensation disparities that may lead to an employee leaving a company. Being fairly compensated is a crucial part in the well being of an employee since it directly affects their quality of life and financial stability. Disparities in compensation can lead to dissatisfaction and mistrust in the workplace.

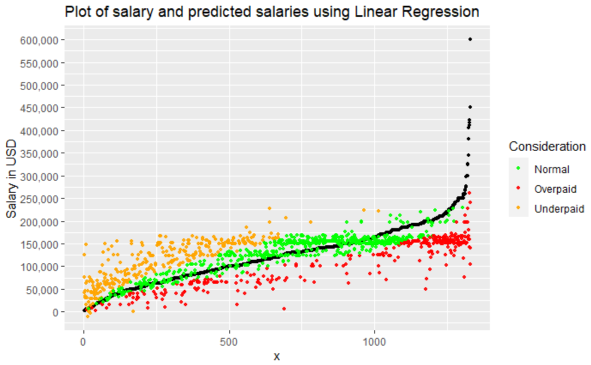
By utilizing data mining techniques, **patterns and trends in employee compensation can be identified and analyzed**. This information can then be used to address any discrepancies and implement fair and equitable compensation practices. Additionally, data mining can also reveal factors that may contribute to these disparities, such as location, industry, or experience level, allowing for targeted solutions to be developed. This information can be used by companies to retain top talent, or to hire outsiders for less, as well as by employees to make informed decisions when considering switching companies.

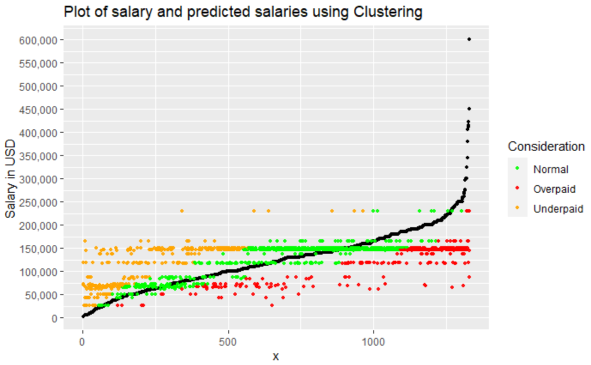
In order to address this research problem, we will use two techniques: Multi-Linear Regression and Clustering.

For **clustering**, we have chosen to use the **K-Modes** method as it is better suited for dealing with categorical attributes, as is the case in this analysis. Our process consists of the following steps: First, we group our data into 20 clusters (determined using the elbow technique), then we calculate the average salary for each cluster and calculate the percentage error. For example, if someone earns $100,000 per year and is in a cluster with an average salary of $105,000, the error percentage would be -4.76%. Once we have this percentage error, we calculate the quartiles of the errors. We have decided that employees in the 1st Quartile will be considered Underpaid, those in the 3rd Quartile will be considered Overpaid, and all others will be considered normally paid.

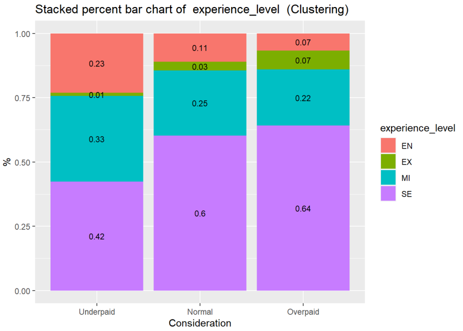
For **multilinear regression**, we follow a similar process. We use the same variables as in the clustering, fit the model, predict the salary for all employees, calculate the percentage error, and label the entries based on the quartiles.

We can then plot the actual salary and the predicted salary on a graph, where the black dots represent the actual salary, and the colored dots represent the predicted salary. The data entries are on the x-axis, and the salary is on the y-axis, and we have also colored them based on their "fairness" status. This will allow us to visually compare the actual and predicted salaries and identify any disparities in compensation.





In this graph we can see an example of the **different distribution of attributes amongst the three “fairness” considerations.**



Entry level employees represent the 23% of the underpaid while only being an 11% and 7% in the normal and overpaid. On the other hand, Executives and Senior employees are respectively 1% and 42% in the underpaid category, 3% and 60% in the normal category and 7% and 64% in the overpaid category. This can help us see that entry level employees might be underpaid compared to their peers (members of its own cluster) while senior and executive might be overpaid. By this we don’t mean that the salaries aren’t fair but we show that **amongst people with similar attributes entry level employees tend to earn less while seniors and executives tend to earn more**.

Employee Seniority: Upward mobility and career growth

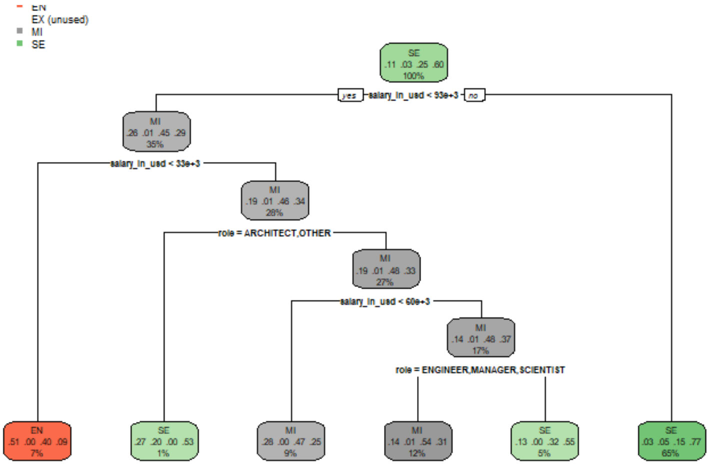
Assessing an employee's experience level can play a vital role in advancing their career, especially when they are at a crossroads in their professional growth. In certain circumstances, it can be challenging to determine their level of expertise based solely on the job title. However, by analyzing various attributes it is possible to make a more informed decision about their experience level and guide them towards the next steps in their professional journey. Additionally, this question can add value as a tool for employee management and development by identifying which employees have the potential for upward mobility within the company.

To accomplish this we’ve taken two approaches, one using Decision Trees and the other one using K-NN.

Decision Trees

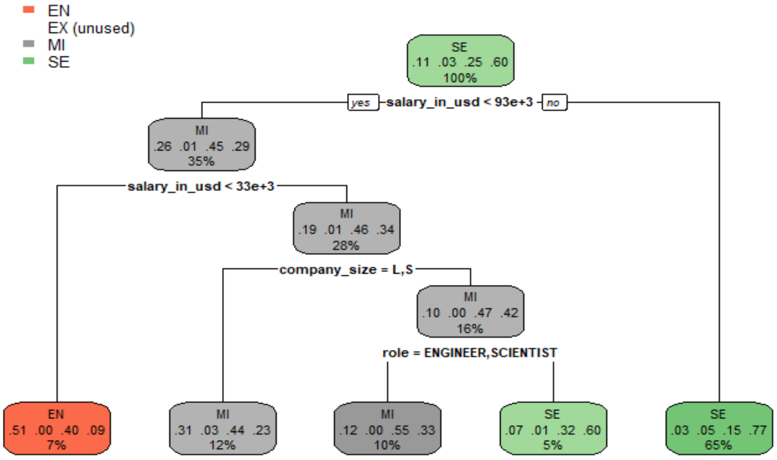
The tree structure of the model allows for **easy visualization** of the important factors that contribute to the prediction, making it **simple to understand the logic** behind the model's decisions.

As a start for our **base model** we only picked up the **industry, role and salary** of the employee as factors for predicting its experience level:



As we can see, the first factor that splits up the decision tree is if the salary is lower than $93.000, if it is higher then the predicted experience level is SE (senior), if the salary is lower than $33.000 the experience level is EN (entry level) and if the salary range is between $33.000 and $93.000 other factors come into play such as the role. Interesting enough the industry doesn't seem to affect the experience level of an employee so from this we can conclude that there is no special industry where a certain expertise is needed.

Next, we crafted a more **advanced model with all of the variables available** (apart from those which factor levels exceeded 32):



Here we can see that the salary ranges in the first two nodes of the tree are the same as in the previous model (less than $93.000 and less than $33.000) but the company size attribute has made a difference when predicting mid experienced employees, with this model the company size is used instead of the role and salary threshold of $60.000.

This model is more accurate since it can choose between all of the available features that we have in the dataset so with this we can say that the **main factors that determine employees experience level are the salary, company size and role of the employee**.

Moreover, from this last decision tree plot we can derive that being in a **small or large company and having the engineer or scientist role would increase the chances of having a higher salary**. This makes sense since these are some of the roles that have higher wages and, on the other hand, startups and big tech companies are the ones that usually offer higher salaries too.

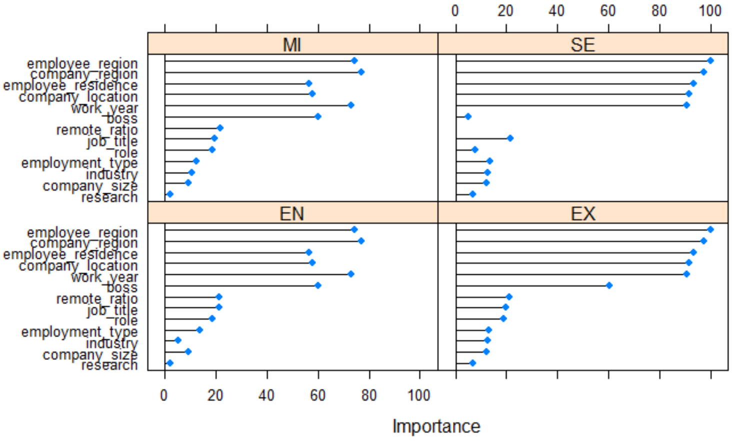
K-NN

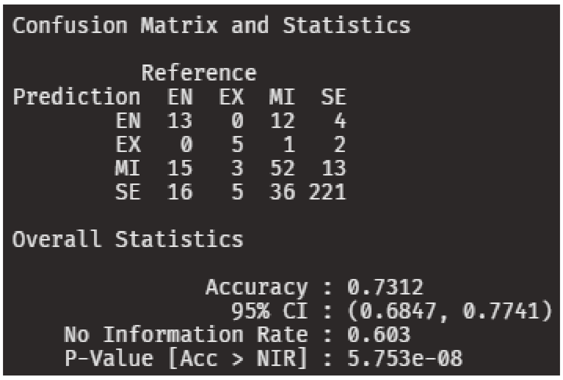
When doing the prediction with K-NN we will be excluding the salary as it is often influenced by factors such as location, industry, and negotiation skills, rather than actual experience and performance. By focusing on the other attributes the evaluation can provide a more accurate reflection of an employee's level of experience.

To answer the question, we first remove all salary attributes from the dataset, and split the dataset into a training set (70%) and a test set (30%). We then create the KNN model with a grid of possible k from 1 to 20. The best result is obtained when k = 1.

With the trained model, we can analyze which attributes in the dataset have the most significant impact on determining an employee's experience level within the company.

Now we can evaluate the knn model’s performance on the test dataset by comparing the predicted experience_levels to the true values. One way to do this is by creating a confusion matrix, which shows the number of true positives, true negatives, false positives, and false negatives for each job level. This can give us a sense of how well the model is performing overall and for each individual job level, allowing us to identify any areas where the model may need improvement. Additionally, we can use this confusion matrix to calculate various performance metrics such as accuracy, precision, recall, and F1-score to further evaluate the model's performance.





```
Confusion Matrix and Statistics

          Reference
Prediction  EN  EX  MI  SE
        EN  13   0  12   4
        EX   0   5   1   2
        MI  15   3  52  13
        SE  16   5  36 221

Overall Statistics

               Accuracy : 0.7312
                 95% CI : (0.6847, 0.7741)
    No Information Rate : 0.603
    P-Value [Acc > NIR] : 5.753e-08
```

A 72% accuracy suggests that the model is performing well. However, it is important to note that this accuracy is based on a specific dataset and may not generalize to other datasets or real-world scenarios. Additionally, it is important to consider other evaluation metrics such as precision, recall, and F1 score to get a more complete understanding of the model's performance. Overall, the model's performance is good, but it may be worth exploring other models or techniques to improve its accuracy and robustness.

Conclusions

While redacting this report we came across with one problem, we didn’t have enough space to explain all the work that we’ve done. In this conclusions we’ll give a brief summary of the discard questions or those who didn’t have good results.

At the beginning, one of the topics that we wanted to cover around this dataset was to see if we can figure out patterns that would lead employees to have a higher purchasing power by working remotely for companies located elsewhere their residency location, for example, work in Indonesia for an American company. But we soon realized that we didn’t have enough data for exploring this casuistry since only 64 out of 1329 employees work for other countries that they reside and only 13 out of this 64 live in countries with a lower cost of living index than the country that they work for (according to 2022 data).

Another research question we tried to answer was "What is worth more, to work in your region or remotely in another region?" Our analysis of the provided dataset revealed that the data is not representative and unbalanced in terms of the number of entries for each region. This leads us to the conclusion that it is not possible to provide a valid answer to the question using the current data. Further research and expansion of the dataset is necessary to accurately balance the representation of different regions and remote types, before we can make meaningful conclusions.

By working on this project we realized the importance of determining the correct scope of the predictions that we can do since, depending on the data balance and structure, these predictions might be biased. For example, we might observe a biased salary prediction for other countries than the US due to the majority of data being from the US where salaries tend to be higher than in the rest of the countries. This results in predictions that may overestimate the salary of a person that is not based in the US or even underestimate when predicting US based employee salaries. Probably reducing our scope (only take into account US data) and having a more clear objective from the beginning would have resulted in better outcomes.

This project provided valuable experience in applying the concepts and techniques learned in class, including data analysis and various models. It also presented the challenge of developing well-defined research questions and refining them throughout the project. Overall, the project was a valuable learning opportunity that allowed us to further develop and strengthen our skills.