P4 Advisory Report

Advisory Group 28

Hari Joshithaa Aghilah Senthilprathiban (2725112)
Małgorzata Zdych (2730740)
Mara-Iuliana Dragomir (2727283)
Yozlem Hashim Ramadan (2732940)

Overview

- Introduction
- Data Understanding
- Data Preparation
- The Models
- Models Comparison
- Final Models Exploration
- Conclusion

Introduction

Introduction

Nowadays, it has become common for firms to utilize machine learning algorithms to aid with the hiring process. This is because there has been an increase in competition and rise in the number of well-qualified candidates. Machine learning algorithms can help simplify the recruitment process making it easier to choose qualified candidates.

Thus, in this presentation/report, we will focus on the implementation of three models: Decision Tree Classifier, Support Vector Machine Classifier and Logistic Regression in the hiring process for a sports recruitment company. The name of the company is A and the sports involving the company are swimming, football and golf.

Data Understanding

Data Understanding

Collecting and describing the data

Firstly, we read the data into pandas dataframe and displayed it. Our data is from the company A and includes Golf, Football and Swimming as sports. We started to explore the data. We applied .head(), shape(), describe(), value_counts(). The data was read correctly. We checked for the missing values, which were not present.

Data Understanding

Gender	Min	Max	Mean	Median
female	21	32	26	26
male	21	32	27	27
other	23	29	27	27

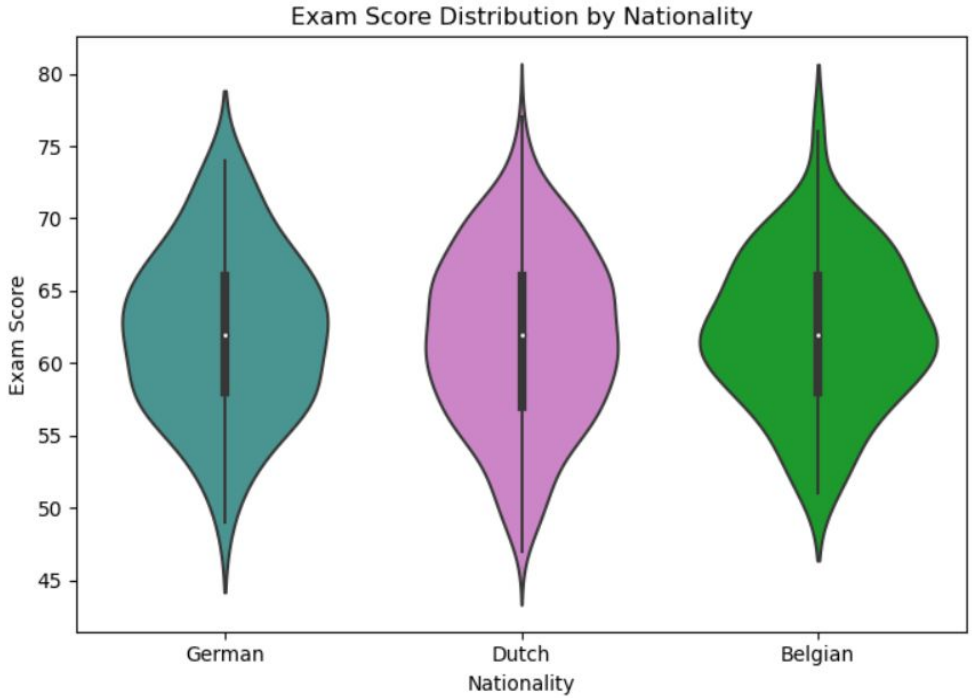
The violin plot illustrates the age by gender in the dataset.

It shows that the age range for both males and females spans from 21 to 32 years. For individuals in subgroup *other*, the range is slightly narrower from 23 to 29.

Moreover, the demonstrated mean and median for all gender groups indicates a well-balanced age distribution. For subgroup *female* there is slightly lower mean and median compared to the other two groups. Overall, all three gender categories exhibit overlapping age distribution.

Data Understanding

Nationality	Min	Max	Mean	Median
Belgian	51	76	62	62
Dutch	47	77	62	62
German	49	74	62	62



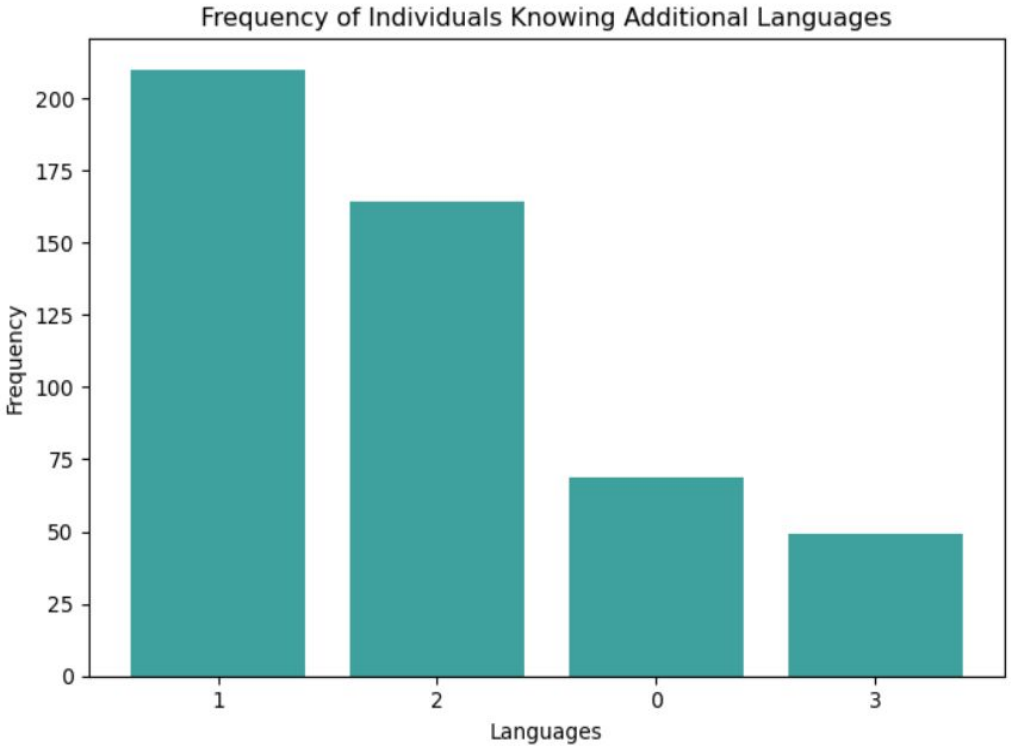
The violin plot shows the distribution of university scores among individuals based on their nationality.

It reveals that individuals from all nationality groups have scores spanning from 47 to 77, with consistent mean and median of 62.

Belgian group have the highest minimum score and the second-maximum after the *Dutch* individuals.

Overall, this suggests a notable similar performance on exams among these nationalities.

Data Understanding



Languages	0	1	2	3
Frequency	69	210	164	49

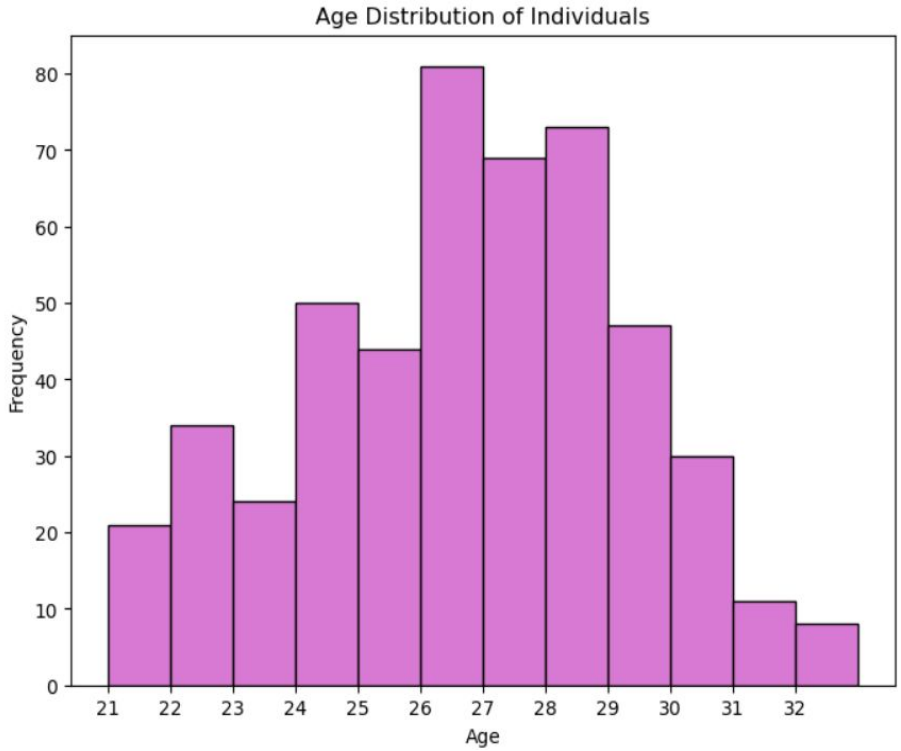
The histogram reveals a diverse range of language proficiency within the dataset.

Significantly, there are 210 bilingual individuals, while 69 individuals are monolingual, indicating a subset with proficiency limited to their primary language only. Moreover, the multilingual group is the highest in the dataset with 213 individuals knowing 2 or 3 additional languages, where 3 languages is less common with only 49 individuals.

This distribution provides insights into the linguistic diversity of the dataset.

Data Understanding

Age	21	22	23	24	25	26	27	28	29	30	31	32
Frequency	21	34	24	50	44	81	69	73	47	30	11	8

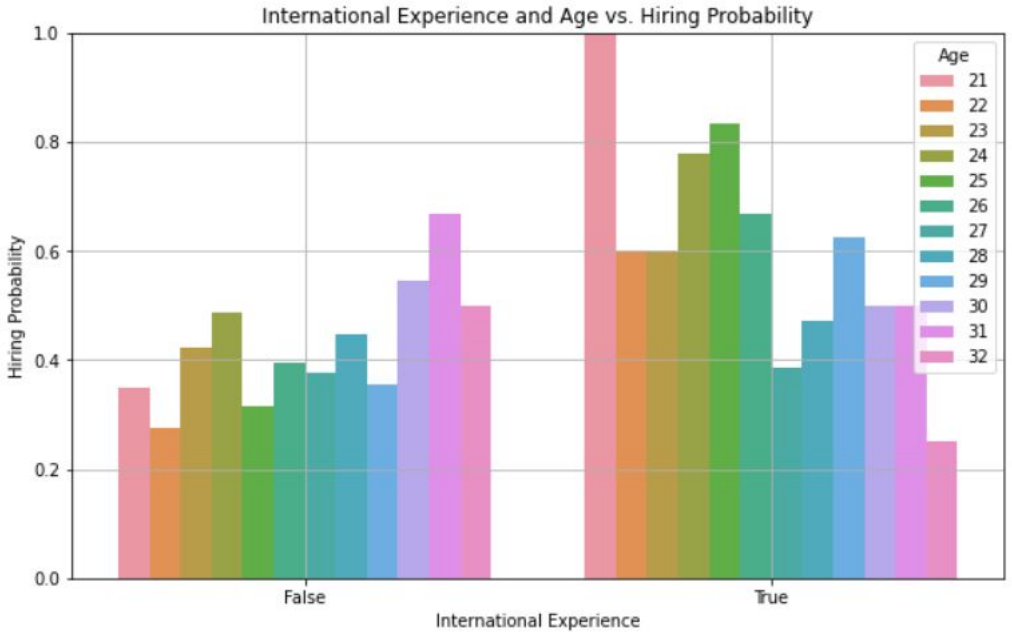


The histogram represents the distribution of individuals by age in the dataset.

The range is from 21 to 32 years.The most frequent ages in the dataset are 26 and 28 years, with 81 and 73 individuals, respectively, suggesting a concentration of individuals around these ages.Individuals from 25 to 29 years have higher frequency within this age range. Furthermore, individuals in their early twenties and early thirties have the lowest frequencies.

Overall, the histogram highlights that people in their late twenties (25 to 29 years) are the most common age group in the dataset, providing a comprehensive understanding of how ages are distributed in the dataset.

Data Understanding

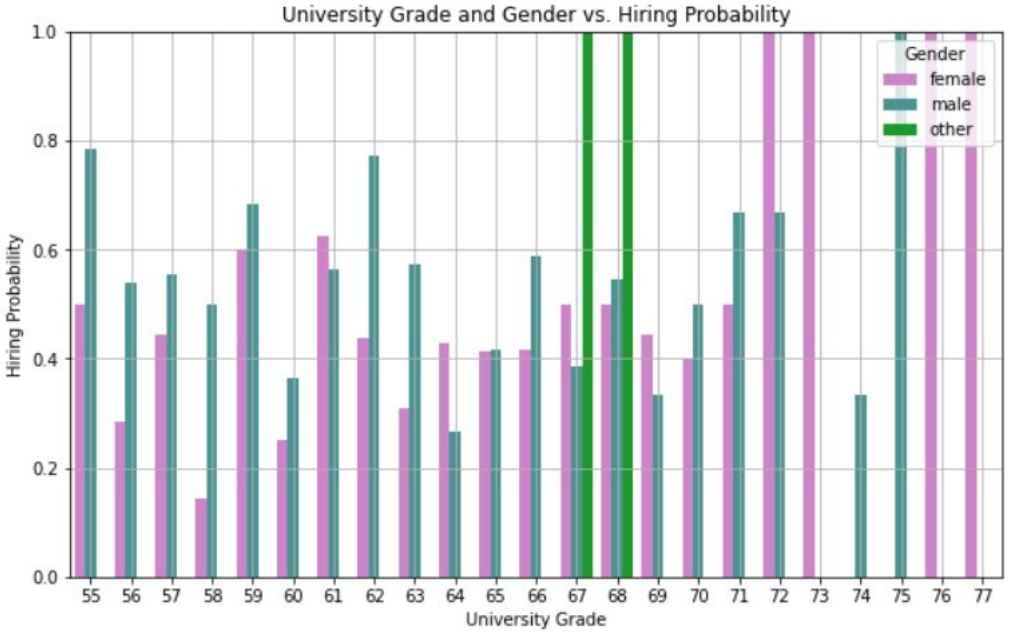


Multiple bar charts were made to showcase the probability of being hired for various subgroups. In this graph the subgroups are people of each age and their international experience - whether they have it or not.

The graph illustrates an interesting relationship between age and international experience in the context of hiring. In the case of no international experience what matters more is the age. In general, older candidates tend to be hired more often.

Whereas, for the subgroup with international experience, the tendency seems to be the opposite, people that are the youngest are hired the most and the oldest ones the least. However, the values in between may suggest that when individuals possess international experience, their age becomes a less decisive factor in the hiring process. In conclusion, the international experience significantly helps with getting hired. However, in the absence of such experience, age becomes a more significant factor in determining hiring outcomes.

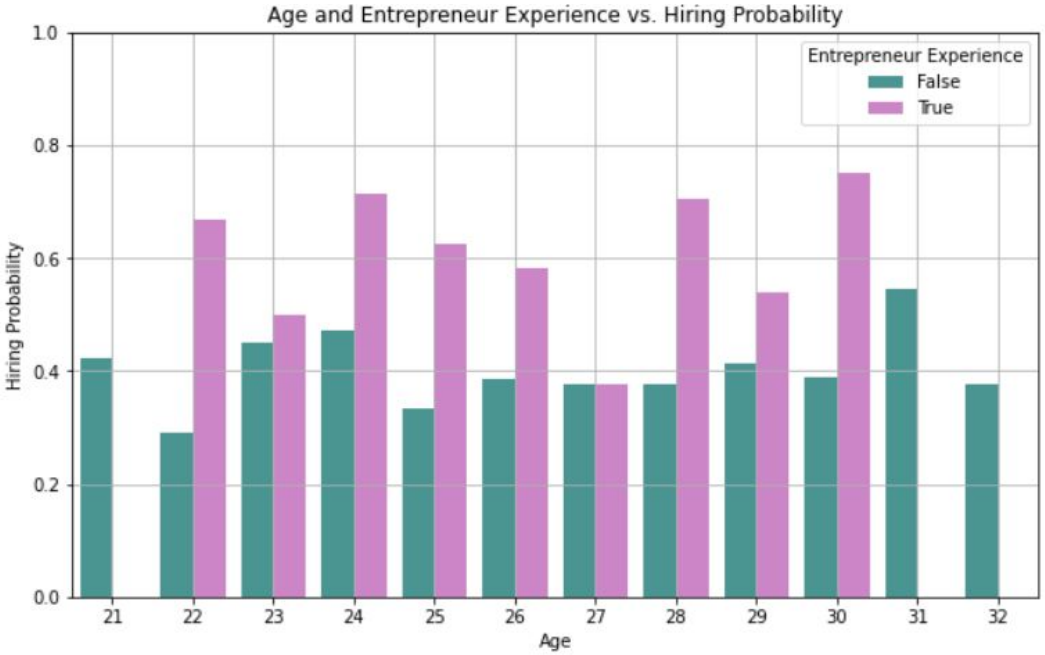
Data Understanding



The graph illustrates a relationship between university grade and gender in the context of hiring. It is noticeable that for the lowest grades from 55 to 60, men are hired significantly more often. The hiring patterns get subsequently more balanced with both man and women being hired equally. For the grades 67 and 68, the other genders are hired the most, but it is probably, because those were singular instances and thus, their hiring probability is 100%. For the grades 72 and higher every woman that applies for a position gets it, but it's possible that these are also singular instances.

To sum up, women tend to be hired more, when their grades are higher. Conversely, for men when they have one of the lowest grades they are hired more frequently in comparison to women and what is more, the proportions between hiring probabilities seem to be the highest there.

Data Understanding



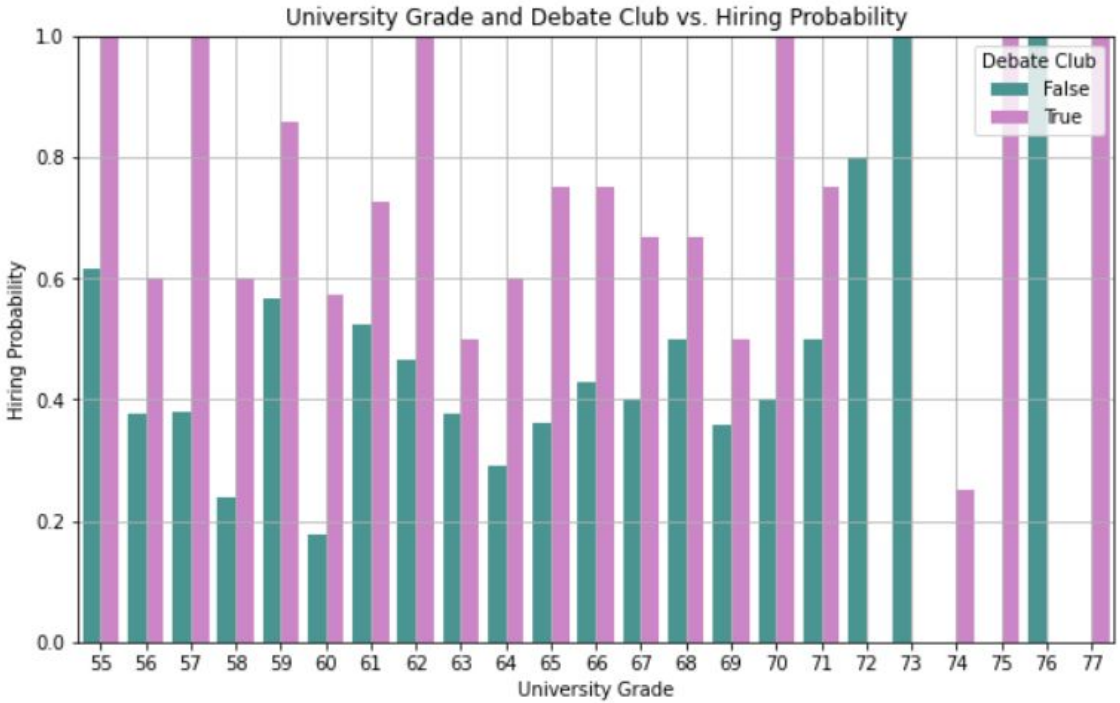
The graph provides insights into the relationship between age and entrepreneur experience in the context of hiring probability. It is prominent that candidates with entrepreneur experience in the vast majority were hired in contrary to the ones without such experience. Thus, this suggests that employers highly value entrepreneurial skills and take into account in their hiring decisions.

What is more, it turns out that people that are on the opposite sides of the age spectrum, so the youngest (21) and oldest (31-32) either do not have any experience at all or that people with experience int those age ranges were not hired at all. At the age 21, it's possible that candidates could be still in their education stage and they did not have much opportunity to gain entrepreneurial experience. Whereas, older candidates could have pursue different career path before they decided to transition, which could explain their lack of experience.

Moreover, candidates in the age ranges 22-26 and 28-30 people with entrepreneur experience tend to be hired more frequently.

In summary, the positive correlation between entrepreneur experience and hiring probability is determined. However, the influence of age seems to have a varying significance.

Data Understanding



The graph illustrates a relationship between university grade, participating in the debate club and the likelihood of being hired.

It is noticeable that candidates with debate club experience not only get hired more often, but it can be stated that their employability is certain. Participation in such a club can bring about a lot of benefits, for example strong communication and critical thinking skills. For candidates who did not join the debate club, the university grade had a more determining impact on their probability of being hired, especially in the case of grades exceeding 71. However, the exceptions are evident, such as candidates with lower grades 55 and 59 whose employability rate is still relatively high - around sixty percent.

To sum up, the data showcases remarkable influence of debate club participation on hiring probability. However, the importance of university grade is also highlighted, although it is not a sole factor in candidate's employability.

Data Preparation

Data Preparation

In order to prepare our data, we first check for outliers, missing values and duplicates. We did not find any of these and did not remove an data points from the data frame. We then split the data into train and test sets with a 80 to 20 percent ratio.

Then, we decided to One-Hot Encode the Categorical Variables and use a Standard Scaler for the Numerical values. We used a Standard Scaler so that the data is normally distributed and easier for the model to interpret.

The Models

Decision Tree

We decided to implement a Decision Tree classifier to predict whether a certain candidate will be recruited by company A or not. We use the features ['sport', 'ind-university_grade', 'ind-programming_exp', 'ind-languages', 'ind-degree’]to train our model. We used these features to make a prediction for ‘decision’.

We split the dataset into a training and test set with a 80 to 20 percent ratio.

For categorical features, we decided to perform one-hot encoding and used a Standard scaler for numeric features by subtracting each instance by the mean and dividing it by the standard deviation.

Decision Tree - Validation

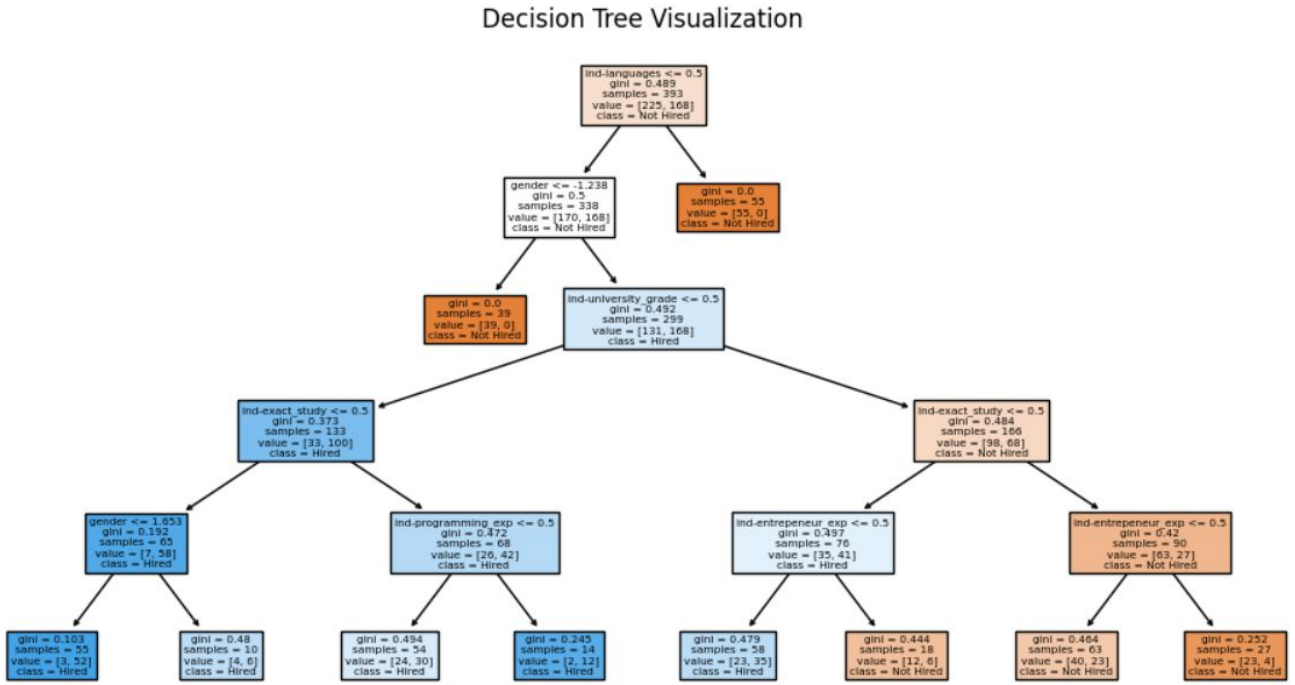
We first implemented GridSearch and 5 fold cross validation to test the performance of our model with different parameters using the f1 score. We decided to use GridSearch to search for a range of parameters for the max_depth, min_samples_split and min_samples_leaf. After performing Grid Search, we tested the model on both the training and validation dataset. We decided to make a confusion matrix with which we can compute the precision, recall and f1 score to determine the performance of our model.

We decided to use three metrics to compare the performance of our model because we wanted to see if our model makes more false positives or false negatives. We also decided to use the f1 score as this takes into account both precision and recall. Thus, our main metric is the f1 score but we also look at the precision and recall to see if our model makes false positive or false negatives.

Decision Tree Model

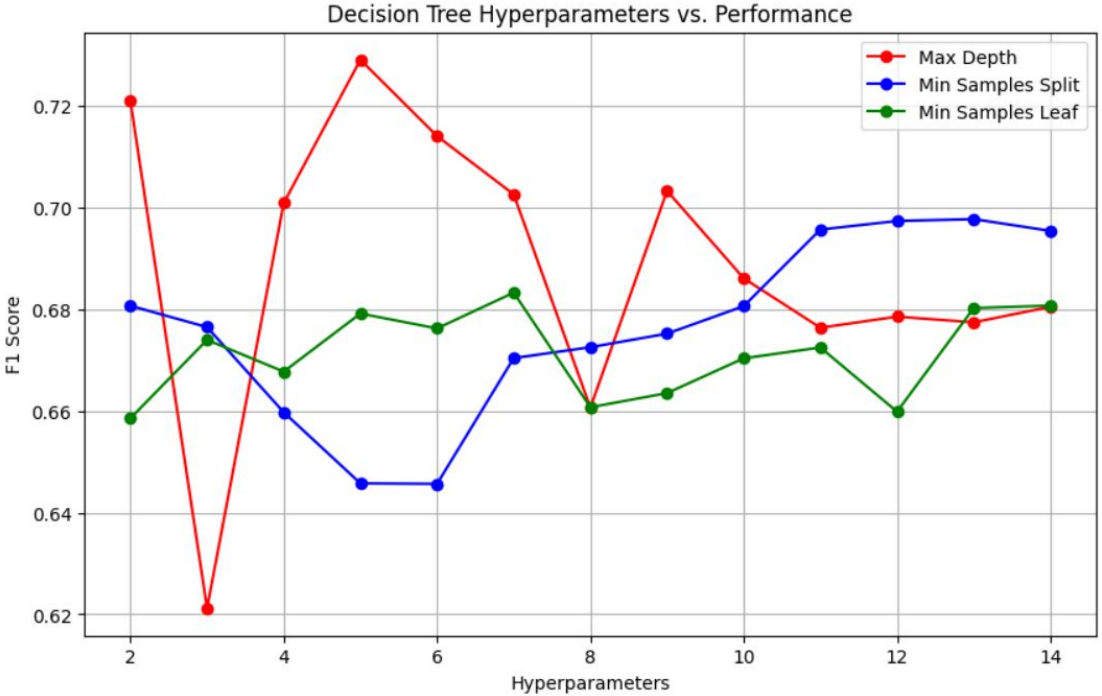
The final model after performing grid search has the following hyperparameters:

```
DecisionTreeClassifier(max_depth=5, min_samples_split=2, min_samples_leaf=10)
```

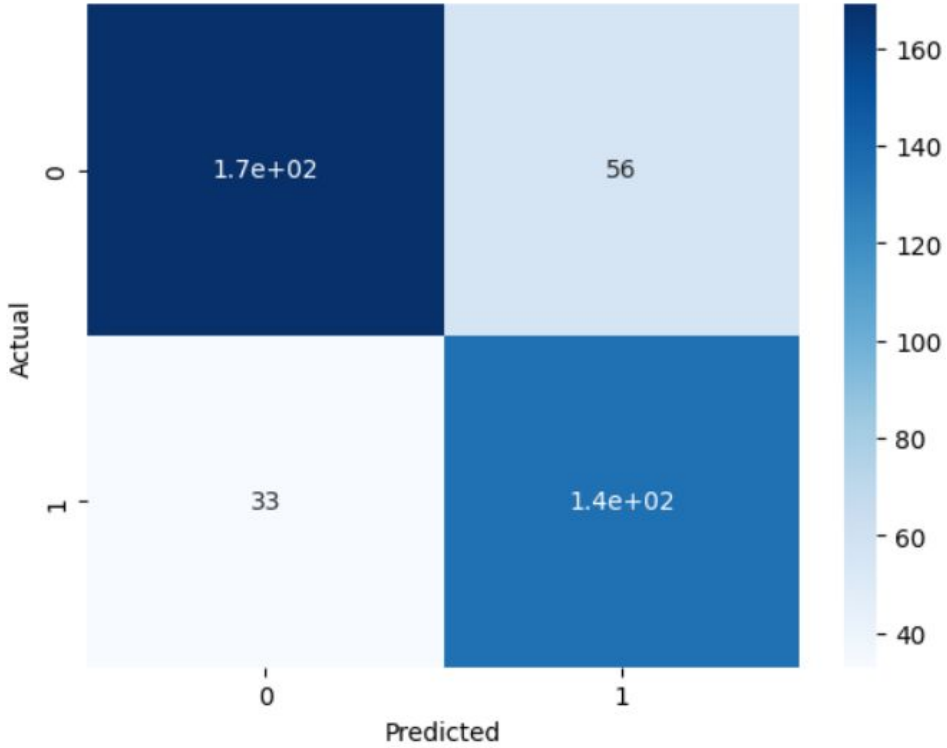


Hyperparameters

We tested values for 3 main hyperparameters. We used the f1 scores to determine the performance of the Decision Tree. As seen, a good value for the Max Depth is 5, for Min Samples Split is around 2 and Min Samples Leaf is around 10.



Confusion Matrix Training Set: Decision Tree



Training set

On the training set, we have the following metrics:

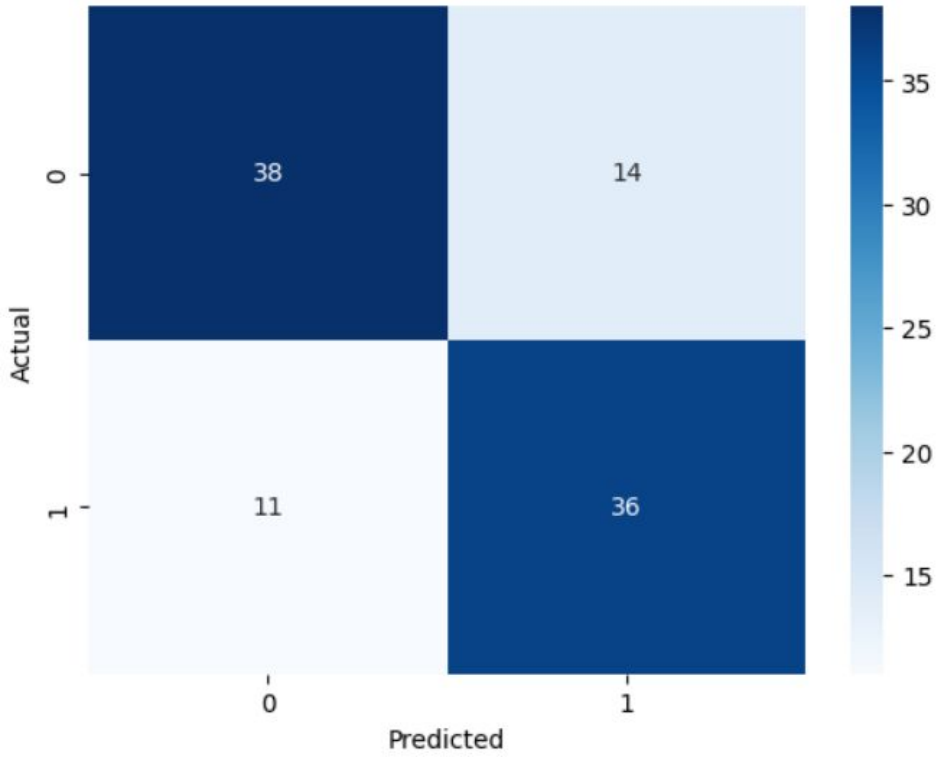
Precision: 0.71

Recall: 0.80

F1-score: 0.75

The model has a somewhat good performance on the training dataset.

Performance on Test Dataset: Decision Tree



Precision: 0.72

Recall: 0.77

F1-score: 0.74

These values indicate that the model is not too good but it is moderately accurate in making predictions.

SVM - model selection and validation

We chose to implement Support Vector Machine classifier due to its ability to handle non-linear and complex classification tasks. We use the features ['sport', 'ind-university_grade', 'ind-debateclub', 'ind-programming_exp', 'ind-degree', 'ind-international_exp', 'ind-languages', 'ind-exact_study', 'company'] to train our model, as we did for Decision Tree (DT) classifier. We employed these features to make a prediction for ‘decision’.

Moreover, we took the budget constraint into account and only four indicators were implemented: 'ind-degree,' 'ind-programming_exp,' 'ind-languages,' and 'ind-university_grade.'

SVM - model selection and validation

Similarly to DT, for categorical features we decided to perform one-hot encoding and used a Standard scaler for numeric features by subtracting each instance by the mean and dividing it by the standard deviation.

For all the models we split the data into a training and test set, adhering to 80-20 ratio. For the assessment of the model's performance, we used precision, recall, and F1-score as the main metrics. We aim to achieve results exceeding 80% by optimisation of the model's hyperparameters and utilizing kernel.

SVM - input and hyperparameters selection

We attempted to maximise predictive accuracy by fine-tuning process. It included exploring various C-values that ensure minimisation of classification errors, gamma values that influence the shape of decision boundary and also kernel type that in general influence model's performance. In such a way we were able to find the most suitable hyperparameters and kernel for the model.

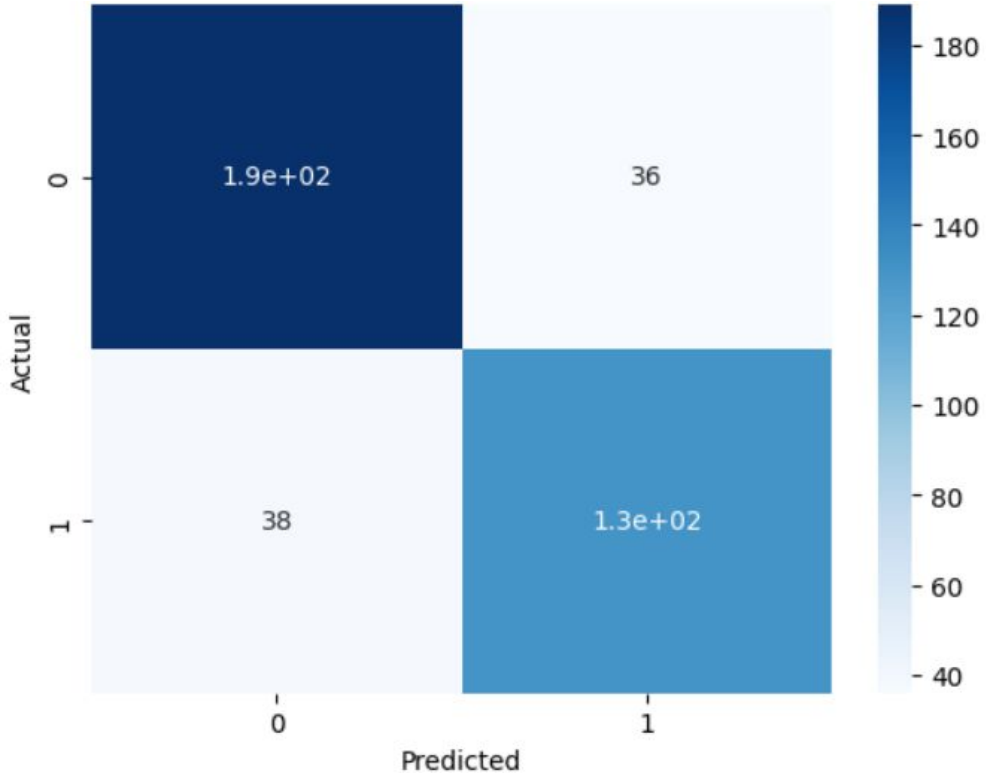
Initial values of parameters:

```
param_grid = {
    'classifier__C': [0.1, 1, 10],
    'classifier__kernel': ['linear', 'rbf'],
    'classifier__gamma': [0.1, 1],
}
```

After fine-tuning process:

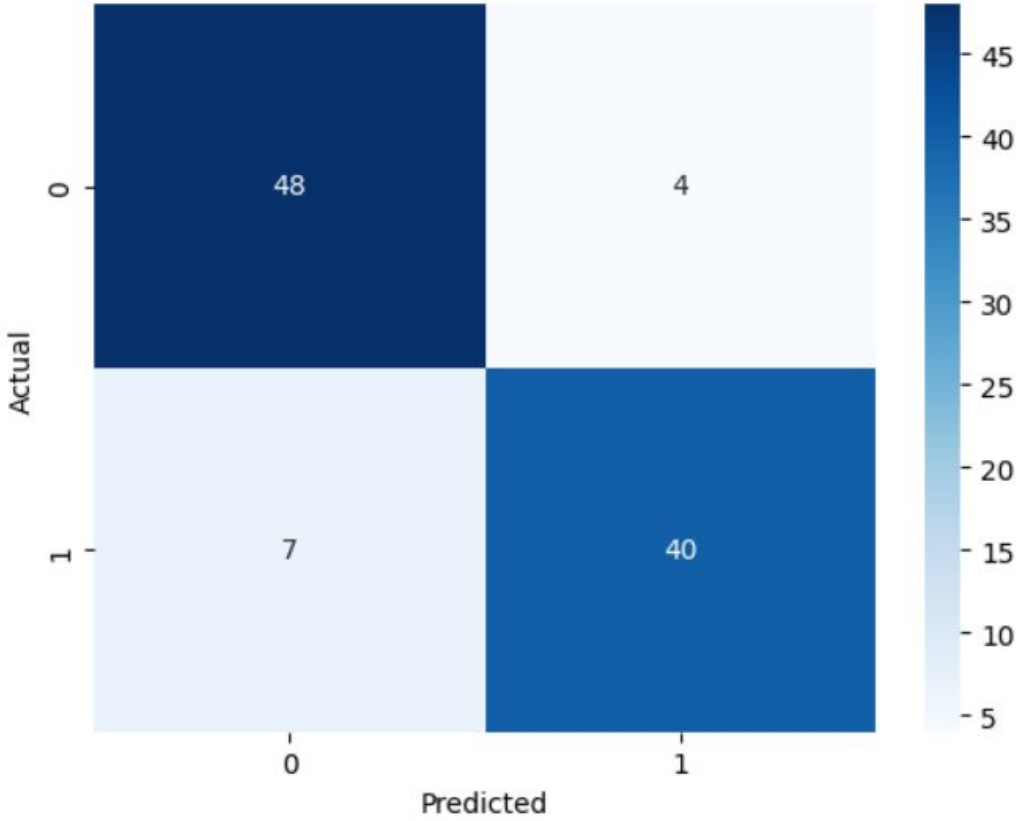
```
param_grid = {
    'classifier__C': [1, 10],  # A
    'classifier__kernel': ['rbf'],
    'classifier__gamma': [1],  # A
}
```

Performance on Training Set



On the training set, the precision was around 0.83, recall around 0.84 and f1 score around 0.84.

Performance on Test Set



On the test set, the precision is 0.87, the recall is 0.92 and f1 score is 0.9.

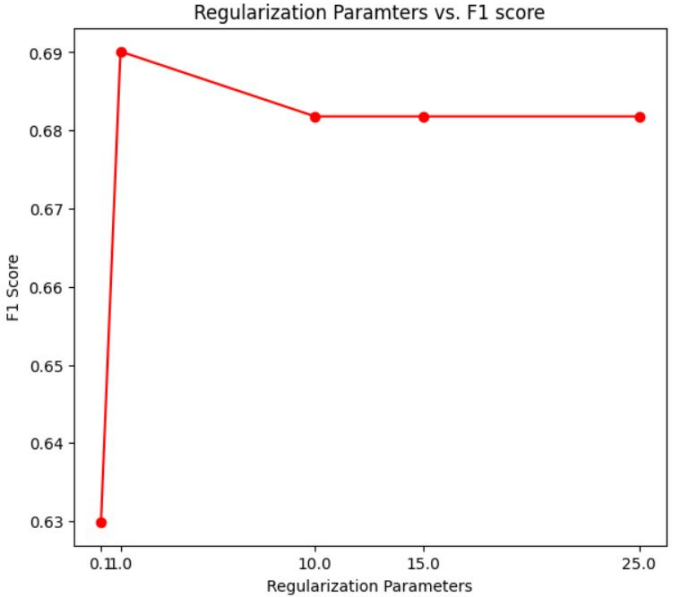
Logistic Regression - Model

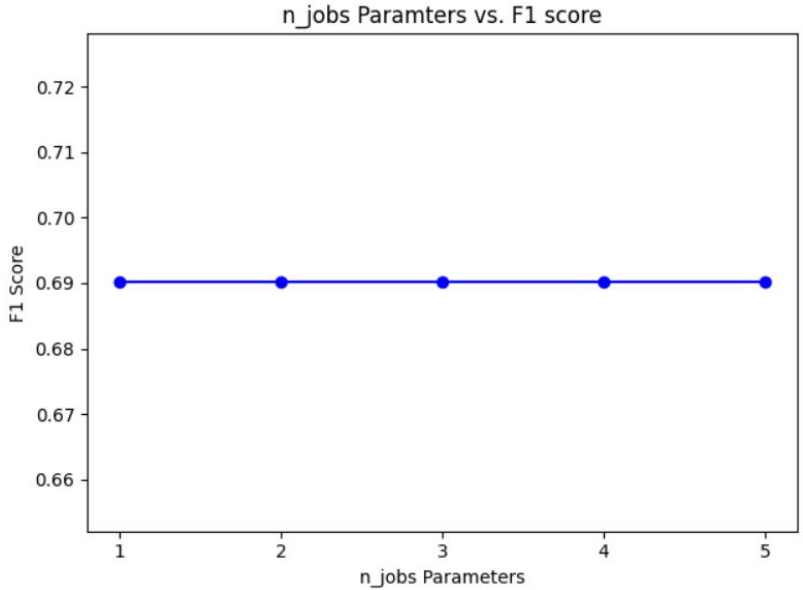
Again, we used the features ‘ind-degree’, ‘ind-programming_exp’, ‘ind-languages’ and ‘ind-university_grade’ to make a prediction for ‘decision’.

We decided to tune the model for the hyperparameters C (regularization) and n_jobs. The values we tested for C are [0.1, 1, 10, 15, 25] and n_jobs are [1,2,3,4,5]. Our final logistic regression model is as follows:

```
LogisticRegression(C=1, n_jobs=1)
```

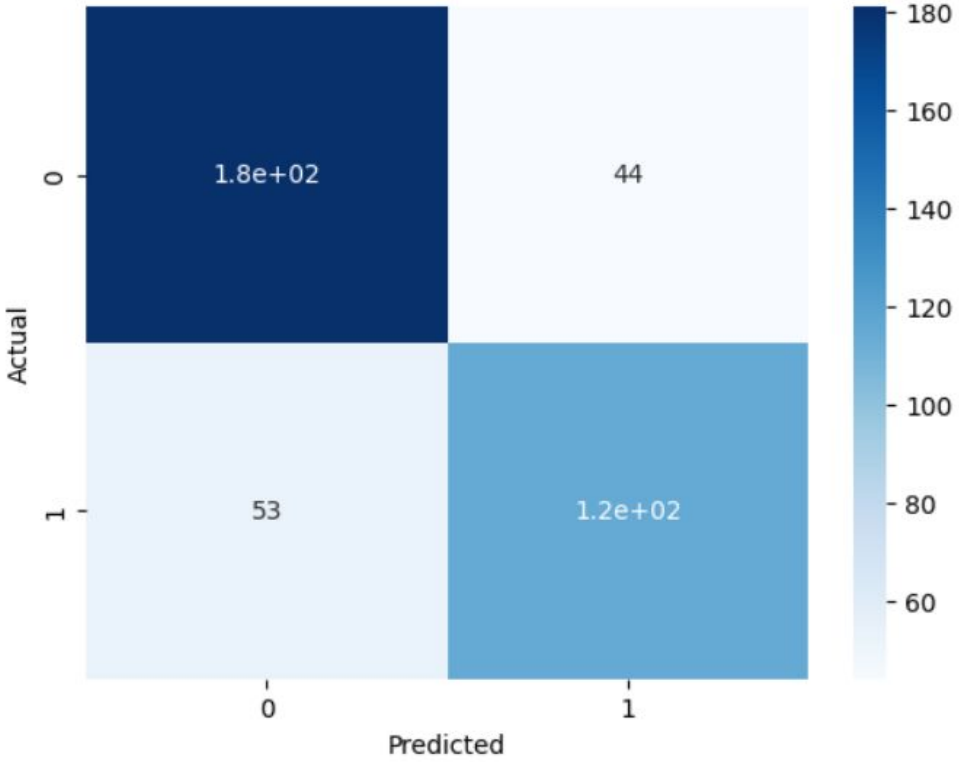
Validation of Hyperparameters





We performed cross validation for 2 hyperparameters in the Logistic Regression model. As seen the regularization parameter which gives a good f1 score is 1. Similarly, the parameters for n_jobs do not influence the f1 score, thus a good value for n_jobs would be 1.

Performance in Training Set: Logistic Regression



Training set

On the training set, we have the following metrics:

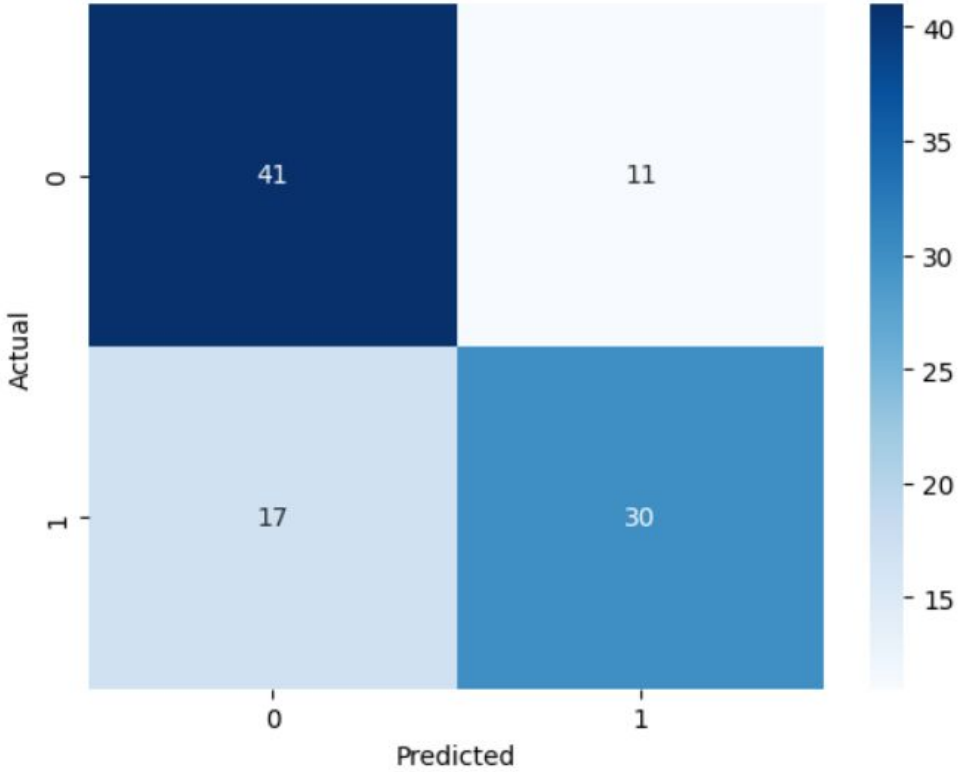
Precision: 0.72

Recall: 0.68

F1-score: 0.70

The model has a somewhat good performance on the training dataset.

Performance on Test Set: Logistic Regression



On the training set, we have the following metrics:

Precision: 0.73

Recall: 0.64

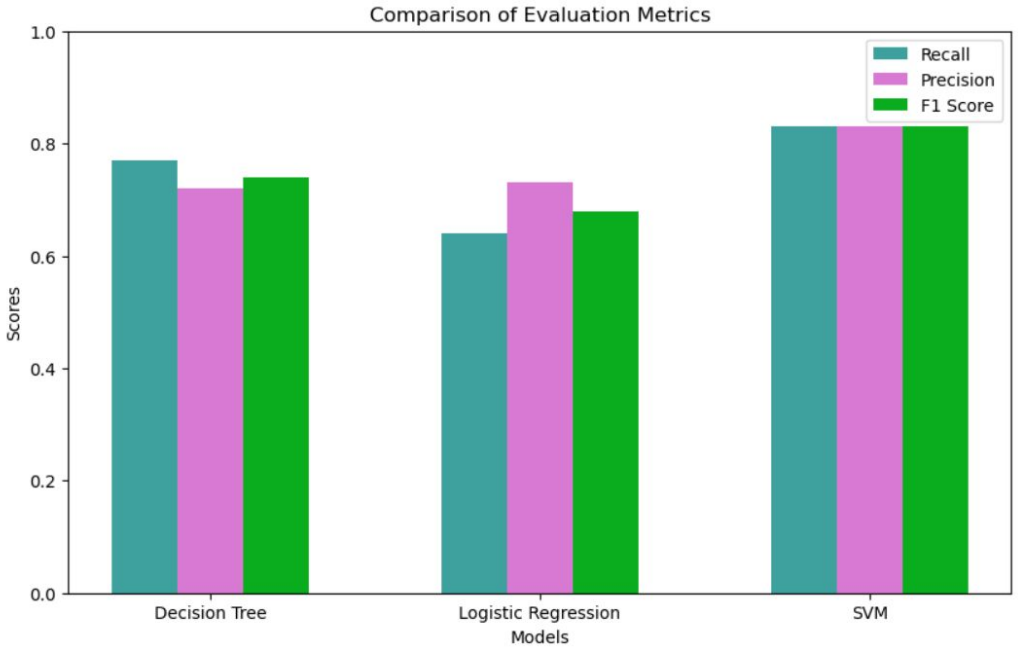
F1-score: 0.68

The model has a somewhat good performance on the training dataset.

Models Comparison

Models Comparison

	Precision	Recall	F1
Decision tree	0.72	0.77	0.74
Logistic regression	0.73	0.64	0.68
SVM	0.83	0.83	0.83



- In this classification task the team focused on SVM and Decision Tree, because those methods are well suited for both classification and regression tasks. Moreover, logistic regression focuses on binary classification tasks, which meets the objective of the task.

- The best model turned out to be SVM model and it had the highest performance on all precision, recall and F1 all 0.83.

- SVM captures complexity and non-linearity of the patterns in the data most proficiently

Final Models Exploration

Are our models biased?

We decided to evaluate the fairness of our best models by observing how they behave for different subgroups of the population, split by sensitive features: specifically by gender & nationality, and by age & sport. For each of the models, the process was split in 2 phases, based on the type of metrics used and representation:

- Phase 1: Model performance (accuracy, precision, recall & F1 score) for each subgroup, in tables
- Phase 2: Predicted vs. Real Hiring Probability for each subgroup, represented in bar charts

Benchmark: we consider that the maximum acceptable difference between actual and predicted hiring probabilities is 20% (in either direction), as it seems a reasonable amount. Anything exceeding this implies that the model is biased.

Decision Trees Model Performance

Decision Trees Model performance for subgroups based on nationality and gender

Nationalities: Dutch, German, Belgian; Genders: Male, Female, Other

	DM	DF	DO	GM	GF	BM	BF	BO
Accuracy	70.97%	68.42%	0.00%	100.00%	80.00%	100.00%	71.43%	100.00%
Precision	73.68%	63.16%	0.00%	100.00%	100.00%	100.00%	50.00%	0.00%
Recall	77.78%	70.59%	0.00%	100.00%	75.00%	100.00%	50.00%	0.00%
F1 Score	75.68%	66.67%	0.00%	100.00%	85.71%	100.00%	50.00%	0.00%

Decision Trees Model performance for subgroups based on sport and age group

Sports: Football, Golf, Swimming; Age Groups: 21-23, 24-26, 27-29, 30-32

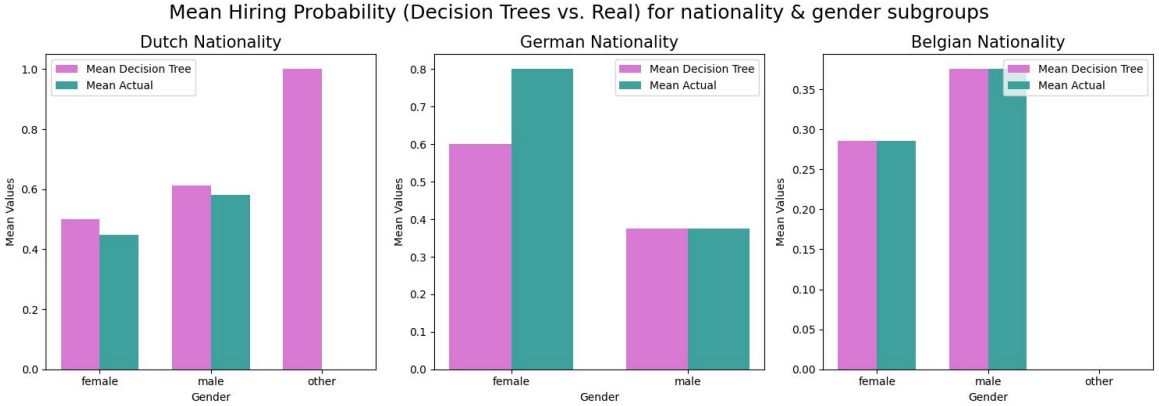
	F(21-23)	F(24-26)	F(27-29)	F(30-32)	G(21-23)	G(24-26)	G(27-29)	G(30-32)	S(21-23)	S(24-26)	S(27-29)	S(30-32)
Accuracy	50.00%	86.36%	71.43%	100.00%	100.00%	66.67%	62.50%	50.00%	66.67%	75.00%	76.92%	100.00%
Precision	25.00%	90.00%	55.56%	100.00%	100.00%	66.67%	60.00%	100.00%	75.00%	60.00%	100.00%	100.00%
Recall	100.00%	81.82%	100.00%	100.00%	100.00%	66.67%	75.00%	33.33%	60.00%	100.00%	50.00%	100.00%
F1 Score	40.00%	85.71%	71.43%	100.00%	100.00%	66.67%	66.67%	50.00%	66.67%	75.00%	66.67%	100.00%

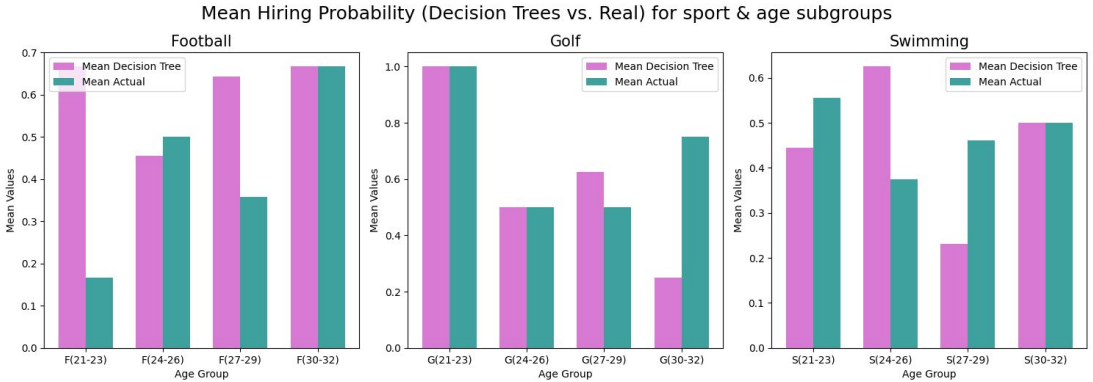
Decision Tree Model Performance

It seems that for most subgroups the decision tree model works decently well. Some positive exceptions are German / Belgian Men, Golf players between 21 and 23 years old and Footballers or Swimmers between 30 and 32 years old, which have 100% performance. This means that the model works perfectly for these subgroups. However, there are also subgroups that are suffering from our model, most importantly nonbinary people (regardless of nationality), but also Football players between 21 and 23 years old and Golf players between 30 and 32 years old. Now in order to detect actual bias, we need to analyze the selection rates as well.

Note: it’s possible that performance for nonbinary people is poor due to class imbalance, as there are only 2 nonbinary people in the test dataset.

Decision Tree Model Hiring Probabilities





Decision Tree Model Hiring Probabilities

There seem to be quite some subgroups for which the model is unacceptably biased (either negatively or positively): Dutch non-binary people, Footballers of ages 21-23 and 27-29 years old, Golf players of 30-23 years old and Swimmers of ages 24-26 and 27-29 years old.

Therefore, the Decision Tree Model is unfair, even though for the rest of the subgroups the model is either accurate or only slightly deviating (with a difference smaller than 20% from the real hiring probability).

Logistic Regression Model Performance

Logistic Regression Model performance for subgroups based on nationality and gender

Nationalities: Dutch, German, Belgian; Genders: Male, Female, Other

	DM	DF	DO	GM	GF	BM	BF	BO
Accuracy	64.52%	71.05%	100.00%	75.00%	60.00%	100.00%	71.43%	100.00%
Precision	76.92%	66.67%	0.00%	66.67%	100.00%	100.00%	50.00%	0.00%
Recall	55.56%	70.59%	0.00%	66.67%	50.00%	100.00%	50.00%	0.00%
F1 Score	64.52%	68.57%	0.00%	66.67%	66.67%	100.00%	50.00%	0.00%

Logistic Regression Model performance for subgroups based on sport and age group

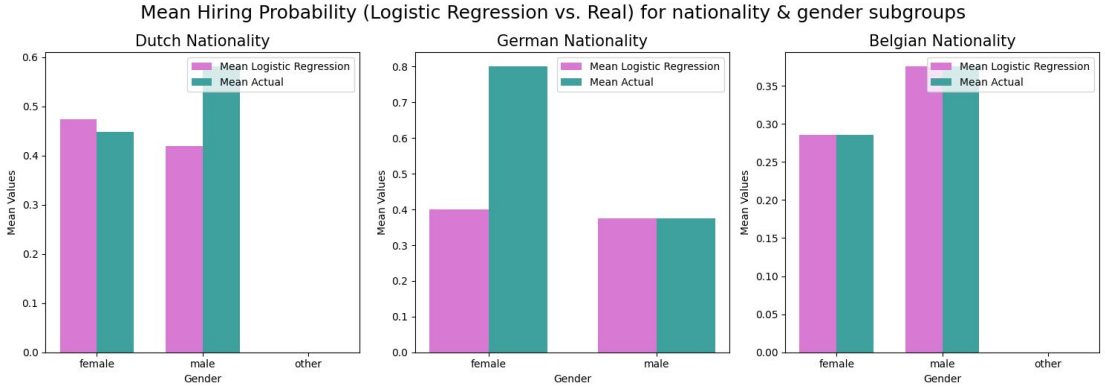
Sports: Football, Golf, Swimming; Age Groups: 21-23, 24-26, 27-29, 30-32

	F(21-23)	F(24-26)	F(27-29)	F(30-32)	G(21-23)	G(24-26)	G(27-29)	G(30-32)	S(21-23)	S(24-26)	S(27-29)	S(30-32)
Accuracy	66.67%	81.82%	78.57%	83.33%	100.00%	50.00%	62.50%	50.00%	88.89%	50.00%	61.54%	100.00%
Precision	33.33%	88.89%	75.00%	100.00%	100.00%	50.00%	66.67%	100.00%	83.33%	40.00%	66.67%	100.00%
Recall	100.00%	72.73%	60.00%	75.00%	100.00%	33.33%	50.00%	33.33%	100.00%	66.67%	33.33%	100.00%
F1 Score	50.00%	80.00%	66.67%	85.71%	100.00%	40.00%	57.14%	50.00%	90.91%	50.00%	44.44%	100.00%

Logistic Regression Model Performance

The performance of the logistic regression model is relatively similar to the decision tree model, though many subgroups have slightly lower values for the 4 metrics. Some notable differences are that Dutch non-binary people have an accuracy of 100% now, instead of 0%, and that the subgroups with a performance of full 100% on all metrics are the same except for German Men, for whom Logistic Regression performs poorer.

Logistic Regression Model Hiring Probabilities



Logistic Regression Model Hiring Probabilities

With logistic regression, differences between predicted and real probabilities seem to be larger generally, though most of them are in acceptable ranges, and there are some subgroups for which it performs perfectly.

There are still subgroups where bias is noticeable though, but fewer than for the previous model. These subgroups are: German women, Footballers between 21 and 23 y/o, Golf players between 30 and 32 y/o and Swimmers of ages 24-26 and 27-29 y/o. This means that this model is also unfair, but perhaps less biased than the Decision Tree model.

SVM Model Performance

SVM Model performance for subgroups based on nationality and gender

Nationalities: Dutch, German, Belgian; Genders: Male, Female, Other

	DM	DF	DO	GM	GF	BM	BF	BO
Accuracy	90.32%	92.11%	100.00%	87.50%	60.00%	87.50%	85.71%	100.00%
Precision	94.12%	93.75%	0.00%	100.00%	100.00%	75.00%	66.67%	0.00%
Recall	88.89%	88.24%	0.00%	66.67%	50.00%	100.00%	100.00%	0.00%
F1 Score	91.43%	90.91%	0.00%	80.00%	66.67%	85.71%	80.00%	0.00%

SVM Model performance for subgroups based on sport and age group

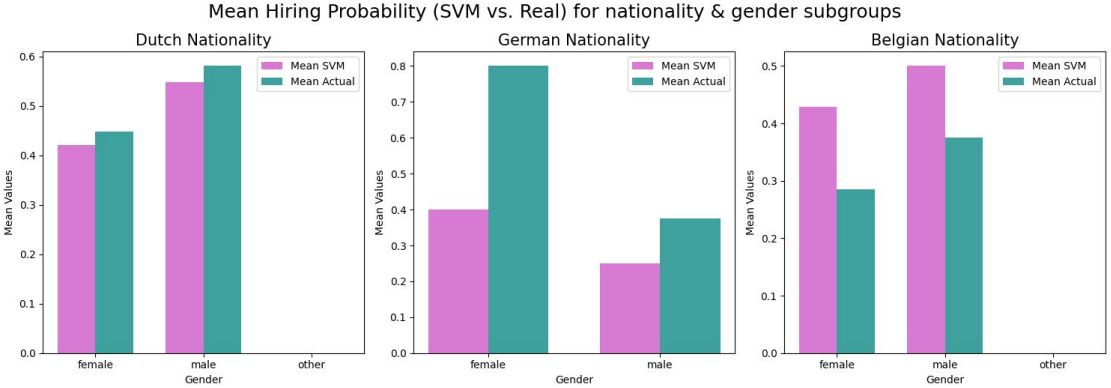
Sports: Football, Golf, Swimming; Age Groups: 21-23, 24-26, 27-29, 30-32

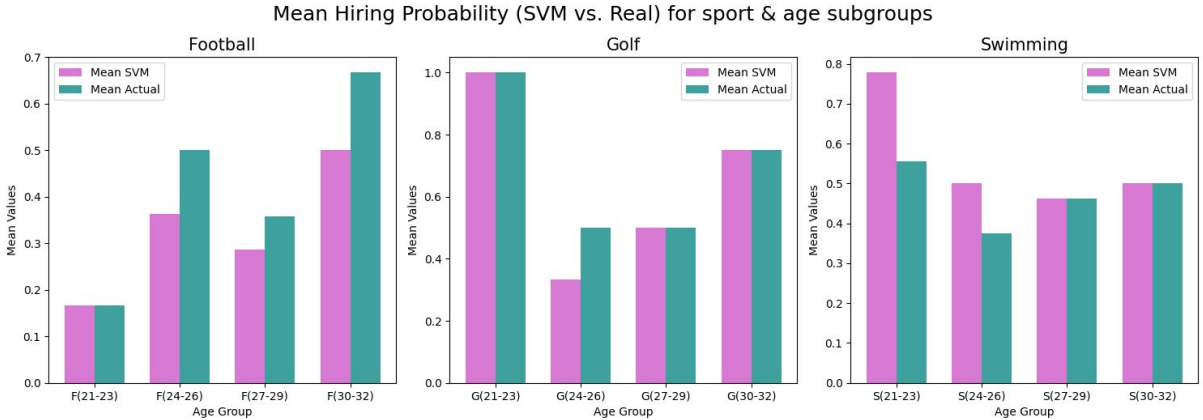
	F(21-23)	F(24-26)	F(27-29)	F(30-32)	G(21-23)	G(24-26)	G(27-29)	G(30-32)	S(21-23)	S(24-26)	S(27-29)	S(30-32)
Accuracy	100.00%	77.27%	92.86%	83.33%	100.00%	83.33%	100.00%	100.00%	77.78%	87.50%	100.00%	100.00%
Precision	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	71.43%	75.00%	100.00%	100.00%
Recall	100.00%	63.64%	80.00%	75.00%	100.00%	66.67%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
F1 Score	100.00%	73.68%	88.89%	85.71%	100.00%	80.00%	100.00%	100.00%	83.33%	85.71%	100.00%	100.00%

SVM Model Performance

The performance of the SVM model seems to be better overall than for the previous models. This time, there are no nationality & gender subgroups with a perfect performance, but there are much more sports & age groups with a full 100% performance instead: Football (21-23), all age groups for Golf except for 24-26, and Swimming (27-29 and 30-32).

SVM Hiring Probabilities





SVM Model Hiring Probabilities

The charts show that hiring probabilities predicted by Support Vector Machines are closer to reality than the previous models. Of course, there are still many subgroups for which the predicted selection rate differs from the real one, but most of the time by a small difference, within our acceptable bound (20%). There are only 2 groups that exceed this bound: German women and Swimmers between 21 and 23 years old. This is the fairest of the 3 models, but still not good enough, as the bias against German women is very large: the predicted hiring probability is half of the true value.

Conclusion

Conclusion

After trying out three distinct classification models for predicting the hiring decision, we have observed that the best model is the Support Vector Machines model, as it has the best overall performance on the test data. It is also the fairest of the models, however still not a truly fair method, for the following reason: there is still bias towards sensitive features such as gender, nationality and age, in a measure that is beyond our acceptable threshold.