DT 2

MACHINE LEARNING AND NEURAL NETWORKS

Digital Technology 2

08 March 2024

Eindhoven

Prepared by **Sultan Basar**

Table Of Content

1. Introduction

In recent years, significant advances have been made in the fields of machine learning and deep learning, transforming various industries and fields. With the proliferation of data and computational resources, the capabilities of machine learning and deep learning models have expanded, providing solutions to increasingly complex problems. The solutions offered by artificial intelligence in the fields of economic, environmental and social sustainability have become a major focus of attention today. In particular, as stated in Lee' (2021) study, it has been observed that the masses tend to use artificial intelligence mostly in the field of economic and environmental sustainability, but the importance of the field of social sustainability is not sufficiently understood.

Considering the fact that artificial intelligence increases productivity, reduces costs and contributes to the environmental field, the potential of this technology is quite large. However, according to the United Nations Environment Program - SETAC Life Cycle Initiative, it is stated that there is not enough focus on social sustainability.(2021, Lee) This makes it difficult to fully evaluate the potential impact of artificial intelligence. Social sustainability is mostly associated with quality of life.Although findings support this perspective, various areas where artificial intelligence is utilized from the standpoint of social sustainability are also recognized. It is notable that algorithmic decision-making processes are becoming increasingly common, especially in the field of Human Resources Management (HRM). This area constitutes the intersection between economic and social sustainability. Köchling's(2020) study reveals that in recent years, companies such as Google, IBM, SAP and Microsoft, which have an important place in the market, have used artificial intelligence in areas such as recruitment and performance evaluation. These developments constitute an important example of how artificial intelligence can be used in the field of social sustainability. Studies conducted in this field can help us better understand the potential effects of artificial intelligence in terms of social sustainability.

Addressing this issue from a social perspective, discrimination in recruitment occurs based on candidates' gender, race, age, or other personal characteristics. Such discrimination is contrary to the principle of equal opportunities and may lead to an unfair recruitment process. The existence of discrimination is considered a significant obstacle to promoting diversity in the workplace and ensuring that everyone has equal chances. Therefore, conducting recruitment processes in a fair and transparent manner is vital to ensuring fairness and diversity in the workforce. In this sense, artificial intelligence has the potential to help make objective decisions and reduce discrimination. Conducting recruitment processes impartially through the use of algorithms and evaluating candidates based on their abilities can be an effective strategy to address this social problem.

This study focused on exploring how various machine learning and deep learning models can be used in fair recruitment processes. The aim of this study is to train machine learning models such as Support Vector Machine and Random Forest, as well as Multi-layer Perceptron, a deep learning model, to make hiring decisions and to evaluate how these models behave in terms of intergroup fairness.Additionally, demographic characteristics such as gender and ethnicity were considered for intergroup evaluation.

Requirements List

- *Achieving at least 75% accuracy in the project.*
- *Minimizing False Positive (FP) and False Negative (FN) values.*
- *To improve the fairness features of the model.*
- *The data set used must be clean, balanced and up-to-date.*
- *Ensuring a balanced distribution of demographic and other sensitive characteristics in the data set.*
- *Data pre-processing steps are carried out by observing the principles of fairness.*
- *Comparison of the performances of different machine learning algorithms (MLP, SVM, Random Forest).*
- *Following a meticulous process for tuning model parameters and hyperparameters.*
- *Integrating fairness metrics into the model analysis process to measure fair behavior.*

2. Methodology

In this study, various machine learning and deep learning models were trained in order to digitalize recruitment processes. First, potential forms of discrimination in the hiring process were identified by reviewing existing literature and reviewing relevant research. The methods used to prevent discrimination in the recruitment model used and the metrics that need to be examined have been identified. The recruitment process was then automated using various machine learning and deep learning models such as Support Vector Machine, Random Forest and Multi-layer Perceptron.These models were then trained and evaluated using a variety of metrics to assess gender and cross-national fairness across different groups. Finally, the results were analyzed and suggestions were made to reduce discrimination in recruitment processes. Finally, the results were analyzed and suggestions were made to reduce discrimination in recruitment processes.

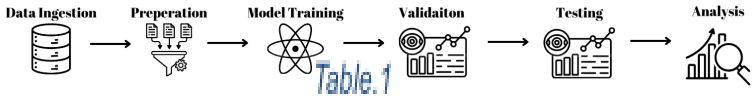


Table.1

The work is based on a pipeline and the table.1 below shows a rough path. Details are explained in the report.

2.1 Characteristics of the Dataset

```
[308]  data.shape

...    (4000, 15)
```

Fig.1

The dataset used in the study is a synthetic dataset containing the hiring decisions of more than 500 employees of four companies.The dataset consists of 4000 rows and 15 columns. The dataset comprises general descriptors and several indicators for each candidate, including the ultimate decision. The features of the raw dataset include gender, age, nationality, sports involvement, university grade, participation in debate clubs, programming experience, international experience, entrepreneurial experience, language skills, field of study, academic degree, and the associated company. These variables collectively contribute to the decision-making process. *See Fig.1 and Fig.2*

```
[307]  data = pd.read_csv("recruitmentdataset-2022-1.3.csv")
       data.head(5)
```

Python

	Id	gender	age	nationality	sport	ind-university_grade	ind-debateclub	ind-programming_exp	ind-international_exp	ind-entrepeneur_exp	ind-languages	ind-exact_study	ind-degree	company	decision
0	x8011e	female	24	German	Swimming	70	False	False	False	False	1	True	phd	A	True
1	x6077a	male	26	German	Golf	67	False	True	False	False	2	True	bachelor	A	False
2	x6006e	female	23	Dutch	Running	67	False	True	True	False	0	True	master	A	False
3	x2173b	male	24	Dutch	Cricket	70	False	True	False	False	1	True	master	A	True
4	x6241a	female	26	German	Golf	59	False	False	False	False	1	False	master	A	True

Fig.2

A brief overview of the dataset was made. With the '*.info()*' method, it was determined that the data set consisted of three different data types: boolean, int64 and object. By using the '*.isnull().sum()'* method, it was determined that there were no null values in the data set. Statistics of numerical data were examined using the '*.describe()*' method, and no significant results were obtained. For categorical data, the '*.describe(include='object')*' method was used to obtain an idea about the most frequently recurring values. The selected dataset is relatively small in size.

Working with a small data set has advantages and disadvantages:

Advantages:

1. Rapid Training and Trial Processes
2. Less Computing Power Required
3. Easy Discovery of Meaningful Patterns

Disadvantages:

1. Generalization Power May Decrease: The model may overfit the training data and its ability to generalize to new data may decrease.
2. Fewer Samples, Less Diversity: Small data sets can reduce the diversity of data represented. This may cause the model to be less adaptable to real-world data.

3. Danger of Overfitting: In cases where training data is less, the model may lose its ability to generalize by overfitting the data.
4. Less Statistical Confidence: Small data sets can make less reliable predictions about how the model will perform in the real world.

2.2 Exploratory Data Analysis

Self Reflection : *At the beginning of the study, an error was noticed and determined to be corrected. The entire data set was divided into five replicates each, namely Company A, Company B, Company C, Company D and Total data set. This approach aimed to analyze each company's recruitment process individually, as well as examine the general characteristics of the entire data set. However, before starting the data set operations, it was not divided into Train, Validation and Test sets. This showed that the transactions carried out could carry the risk of data snooping. In order to correct this error, Exploratory Data Analysis (EDA) operations performed on the total data set were deleted. Then, Company A, B, C and D parts of the copied data set were analyzed separately and the modeling process was performed on the Total data set. The EDA process for the total data set was carried out after separating the data set into Train, Validation and Test parts. This step was intended as a precaution to obtain more reliable and consistent results.*

In *Section 2.1*, it was stated that the data set has three different data types. Intergroup analyzes were carried out by taking these different data types into account in the data set review and visualization. Numerical data 'age', 'ind-university_grade' and 'ind-languages' were visualized through histograms. Categorical data were visualized using bar plots and pie charts for columns such as 'gender', 'nationality', 'sport' and 'ind-degree'. Additionally, bar plots and countplots were used to examine the ratios to the 'decision' column for categorical data. The risk of data snooping occurs at this point. Although efforts were made to prevent errors that were discovered later, they were not sufficient. The rates in the 'ind-debateclub', 'ind-programming_exp', 'ind-international_exp', 'ind-entrepeneur_exp', 'ind-exact_study' and 'decision' columns, which have Boolean data type, were examined through countplots. The same procedures were applied and evaluated separately for 4 companies.

Below are the findings from the examinations conducted for each company.

2.2.1 Company A

The following observations can be made regarding hiring decisions:

- **Gender:** Males have a higher decision ratio (45.66%) than females (38.60%) and individuals with other genders (37.50%).
- **Nationality:** Belgian, Dutch, and German nationals exhibit similar decision ratios (around 42%).
- **Sport:** Rugby players have the highest decision ratio (55.03%), followed by football players (46.09%).
- **Degree:** Ph.D. holders have the highest decision ratio (61.82%), followed by master's (51.97%) and bachelor's degree holders (32.30%).
- **Debate Club Participation:** Participants have a higher decision ratio (61.16%) than non-participants (36.42%).
- **Programming Experience:** Individuals without programming experience have a higher decision ratio (46.39%) than those with (33.44%).
- **International Experience:** Those with international experience have a slightly higher decision ratio (48.44%) compared to those without (40.65%).

❖ **Entrepreneurial Experience:** Having entrepreneurial experience significantly increases the decision ratio (58.22%) compared to those without (37.81%).
❖ **Exact Study Background:** Individuals without an exact study background have a higher decision ratio (49.17%) compared to those with (34.42%).

Overall, certain factors like gender, sport participation, and degree level influence hiring decisions, while others like nationality and programming experience have less impact. Participation in extracurricular activities and possessing entrepreneurial experience appear to be positively correlated with hiring decisions.

2.2.2 Company B

The analysis of hiring decisions based on various factors reveals interesting insights:

❖ **Gender**: Male candidates have a higher likelihood of being hired, with a decision ratio of 45.47%, compared to 13.97% for females and 26.32% for individuals with other genders.
❖ **Nationality**: The decision ratios for Dutch (31.42%) and Belgian (29.36%) nationals are slightly higher compared to German nationals (28.09%). However, the differences are relatively small, indicating that nationality may not be a major factor in the hiring decision.
❖ **Sport**: Participation in sports like rugby and tennis is associated with higher decision ratios (53.59% and 50.50%, respectively) compared to other sports. In contrast, individuals involved in chess have the lowest decision ratio at 2.20%.
❖ **Degree**: Candidates with a master's degree have the highest decision ratio at 36.59%, followed by those with a bachelor's degree (26.87%). Ph.D. holders have a lower decision ratio of 23.08%.
❖ **Extracurricular Activities:** Participation in activities like the debate club and entrepreneurial ventures significantly increases the likelihood of being hired, with decision ratios of 79.67% and 77.55%, respectively.
❖ **Programming and International Experience**: Lack of programming or international experience does not seem to hinder employment opportunities, as candidates without these experiences have higher decision ratios (37.23% and 32.59%, respectively) compared to those with such experiences.

These findings emphasize the importance of considering a diverse range of factors, including gender, sports involvement, educational background, and extracurricular activities, in the hiring process.

2.2.3 Company C

The following observations can be made regarding hiring decisions:

❖ **Gender**: Males (27.67%) have slightly higher decision ratios than females (23.40%) and individuals with other genders (30.43%), but the differences are small.
❖ **Nationality**: Decision ratios for Dutch (26.21%) and German (27.37%) nationals are slightly higher than for Belgians (21.95%).
❖ **Sport**: Cricket (38.03%) and rugby (39.88%) players have higher decision ratios than other sports, indicating a positive influence on hiring likelihood.
❖ **Degree**: Decision ratios increase with education level, with Ph.D. holders (62.96%) having the highest ratio.
❖ **Debate Club:** Participation (28.45%) slightly increases the decision ratio.
❖ Programming Experience: Lack of experience (30.59%) shows higher ratios than experienced individuals (15.63%).

❖ **International Experience:** Those with international experience (59.36%) have significantly higher decision ratios.

❖ **Entrepreneurial Experience:** Experience (28.02%) slightly affects decision ratios.

❖ **Exact Study Background:** Individuals without an exact background (33.52%) have higher decision ratios.

In summary, education level and international experience positively impact hiring decisions, while gender and nationality show minimal effects. Certain sports and debate club participation also influence decision ratios, while programming or entrepreneurial experience have less significance.

2.2.4 Company D

The analysis of hiring decisions based on various factors reveals interesting insights:

❖ **Gender**: Females (34.31%) have higher decision ratios compared to males (21.99%) and individuals with other genders (28.00%).

❖ **Nationality:** Decision ratios for Belgian (31.63%) and Dutch (27.74%) nationals are higher than for German nationals (21.98%).

❖ **Sport:** Chess players (41.30%) and runners (39.73%) have the highest decision ratios, while rugby players (14.12%) have the lowest.

❖ **Degree:** Decision ratios increase from bachelor's (13.47%) to master's (40.49%) and slightly decrease for Ph.D. holders (39.73%).

❖ **Debate Club Participation:** Non-participants (31.58%) have higher decision ratios compared to participants (13.57%).

❖ **Programming Experience:** Individuals with programming experience (53.71%) have higher decision ratios compared to those without (14.33%).

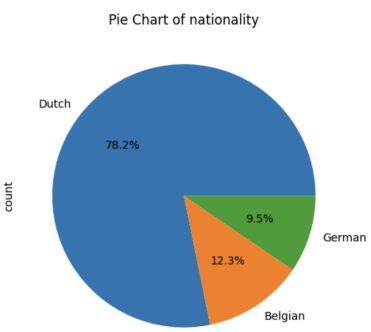
❖ **International Experience:** Decision ratios show little difference between individuals with (28.64%) and without (27.31%) international experience.

❖ **Entrepreneurial Experience:** Individuals without entrepreneurial experience (30.60%) have slightly higher decision ratios compared to those with (16.02%).

❖ **Exact Study Background:** Individuals with an exact study background (50.21%) have substantially higher decision ratios compared to those without (6.73%).

Overall, programming experience and having an exact study background appear to be significant factors positively affecting hiring decisions, while participation in the debate club, international experience, and entrepreneurial experience show minimal impact.

2.2.5 Result Of Exploratory Data Analysis



Upon examining the data, it is evident that each company has its own set of criteria and prioritizes different groups. This prioritization is generally considered acceptable when it includes issues such as school degree, activities performed, etc. However, discrimination based on personal characteristics such as gender and nationality indicates a sensitive and unethical situation. According to the analysis results, although there are differences between companies, it is observed that some companies openly discriminate against individuals of certain gender and nationality. For this reason, it was decided to proceed by taking into account discrimination based on gender and nationality when training and

examining models. This approach aims to ensure that the study is conducted within an ethical and fair framework.

There is a noteworthy point about data distribution, which is a very important issue in "fairness" research. There are clear differences in distribution between different classes. For example, in applications to Company C, you may see at *Fig.3* that there are significantly more applications from "Dutch" nationalities than others. This may cause bias in the established models. *However, this situation can be prevented by taking various precautions.* The details of these measures are explained in detail in the Model Implementation section4.

2.3 Preprocessing

While examining the data structure, no missing data was found. However, during review, an attribute called "ID" was found unnecessary and was removed from the dataset. This was deemed unnecessary because it was thought that values such as ID could reduce the generalization ability of the model, since such unique IDs generally do not contain a pattern. Additionally, the unbalanced distribution between data groups could have been balanced by measuring synthetic data during preprocessing, *but instead it was preferred to resolve it by weighted classification during modeling.* Finally, the necessary procedures were carried out to detect the outlier and no outlier was found.

2.3.1 Feature Engineering

Self reflection: *Label encoding may be more appropriate for Boolean values, since in this case true is usually encoded as 1 and false is usually encoded as 0. However, in some models it may be more effective to use boolean values directly, because in this case they are already interpreted as 1 and 0.Therefore, applying encoding to boolean data can make the model unnecessarily complex. In this study, this application error was noticed and notes were taken to prevent similar errors from being repeated in the future.*

An environment based on numerical data is more suitable for machine learning and deep learning models to work. Therefore, the text data was converted to numerical values as the dataset has a suitable format to convert text data into numerical data.

In the study, boolean data was transformed with the *label encoding* method. "True" values were assigned to "1" and "False" values were assigned to "0". Boolean properties are: ['ind-debateclub', 'ind-programming_exp', 'ind-international_exp', 'ind-entrepeneur_exp', 'ind-exact_study', 'decision']

Additionally, categorical data were transformed with the *one-hot encoding* method. The attributes used for categorical attributes are: ['gender', 'nationality', 'sport', 'ind-degree', 'company']

In this way, the feature engineering part is completed by transforming the data using two different methods. As a result of the preprocess, the new dataset consisted of 30 columns and 4000 rows.

After the sampling process, it applies a scaling process to transform the features in the data set into a standard distribution. Using the *StandardScaler* class, features are standardized to have a mean value of 0 and a standard deviation of 1. Standardizing features eliminates problems caused by data at different scales and helps the model perform better. In addition, the transformation is performed using the same scaling parameters for sensitive features. The purpose of this process is to ensure that models react more fairly to sensitive data.

2.3.2 Data Sampling

Before proceeding to implement the models, it was decided that each model would undergo the same process. As a result, the aim is twofold:

1. To train the models to carry out recruitment with the highest accuracy.
2. To determine which model makes fairer decisions according to the fairness metrics decided upon through research.

Before proceeding with the implementation of the models, the next step is to split the available data into three sets: training, validation, and testing. The reason for this is that it prevents the model from overfitting and ensures that the model still works with high performance when faced with real world data.
At this stage, the fundamental task is to separate *sensitive features*. In this way, the data to check whether decisions are made fairly will be divided into train, validation, and test groups. As previously mentioned, the selected sensitive features include nationality and gender information.
The data is splitted into 3 different subsets. This separation was made for the purpose of model training of the data set, evaluating the performance of the model and testing its generalization ability. Additionally, separate subsets were created for sensitive features. As a result of the operations performed, the test set was divided into 0.15 of the total data, the validation set into 0.15, and the train set into 0.7. The "random_state=42" parameter ensured that this division process was repeatable.

Then, sensitive features were separated for each set. These features are specified as 'gender_female', 'gender_male', 'gender_other', 'nationality_Belgian', 'nationality_Dutch' and 'nationality_German'.
The independent variable of the model is determined as 'decision' feature.
As a result, a training set, validation set and test set containing sensitive features were created along with training, validation and test sets. Separating sensitive features can be used to assess discrimination in subsequent analyzes and is important to determine whether the model is fair to these features.

2.3.3 EDA in TOTAL

The graphs and analyzes obtained as a result of the data analysis after separating the entire data set into train, validation and test sets are given in the Appendix.

3. Model Selection

Self reflection: *Considering the size of the data set, a model such as Naive Bayes could have been chosen instead of Random Forest, which works better on small data sets and handles complex problems more easily. Applying a Random Forest model consisting of a large number of trees to little data may increase the risk of overfitting. Therefore, it may seem that there is no wiser choice. Literature reviews have shown that methods such as SVM and Naive Bayes are frequently preferred for such recruitment processes.*

Three different models were chosen to look at model selection from various perspectives and gain insight into which model works better in this regard. Support Vector Machine (SVM) and Random Forest were preferred as machine learning models. Multi-layer Perceptron (MLP) model was chosen for deep learning. The reasons for choosing these models and their advantages and disadvantages are explained below:

***Support Vector Machine (SVM)*:**
Reasons for Choosing: SVM can work effectively in linear and non-linear classification problems. Additionally, it can perform well on high-dimensional datasets and especially low-dimensional datasets.
Advantages: SVM generally requires fewer hyperparameters and is not prone to overfitting. Additionally, it is resistant to outliers and can be easily extended to multiple classification problems.
Disadvantages: SVM can be time-consuming in terms of training and prediction times on large data sets. Additionally, the complexity of SVM may increase in high-dimensional datasets. However, this disadvantage can be ignored in this study because it works with small-sized data.

Random Forest:

Reasons for Choosing: Random Forest is an ensemble learning method created by combining multiple decision trees. Therefore, it may be more effective in evaluating complex data structures.
Advantages: Random Forest generally gives strong and stable results because it is created by combining multiple decision trees. Additionally, it is capable of learning complex structures in the data set and is resistant to overfitting tendencies.
Disadvantages: Random Forest's complexity and slow prediction times can cause performance issues on very large data sets.

Multi-layer Perceptron (MLP):
Reasons for Selection: MLP can be used to evaluate fairness in the recruitment process with the ability to learn more complex structures.
Advantages: MLP is capable of modeling complex non-linear relationships. Additionally, it has high scalability.
Disadvantages: MLP may take longer to train and require sufficient computing power for large data sets. Additionally, there may be a tendency to overfit and hyperparameter tuning is difficult.

4. Model Implementation

6.1 SVM

6.1.1 Model Building and Raw Performance

In the first stage of the SVM application of the study, the SVM model was directly created without any hyperparameter adjustment and adapted to the training set. Before hyperparameter tuning was performed, raw performance metrics were examined and the results were similar to the values presented in *Figure 4.*

```
Validation Accuracy: 0.705
Confusion Matrix:
[[423   0]
 [177   0]]
Precision: 0.49702499999999994
Recall: 0.705
F1 Score: 0.5830205278592375
Sensitivity (Recall): 0.0
Specificity: 1.0
```

Fig.4

Looking at the *confusion matrix* values, the presence of false positives (FP) indicates that the model incorrectly classified certain samples as positive, while the absence of false negatives (FN) reveals that the model failed to classify any samples as positive.
Precision: A high precision value indicates that the model has low false positives (FP) and most of the samples it classifies as true positives are correct. Here, precision is given as 0.497. This indicates that only about half of the samples that the model predicts as positive are actually positive. A low precision means that the model also includes false positives (FP), meaning that the model appears to incorrectly classify some negative outcomes as positive.
Recall: A high recall value means that the model correctly identified most of the positive examples. The recall value is given as 0.705, indicating that the model correctly identified more than 70% of the samples that were positive. A high recall indicates that the model correctly identified most of the positive samples.
F1 Score: A high F1 score indicates that both precision and recall are high. F1 score value is given as 0.583. This indicates a balanced performance between precision and recall.
Sensitivity: A low sensitivity value indicates that the model is poor at identifying the positive class. The sensitivity value is given as 0.0, which indicates that the model is poor at identifying the positive class.
Specificity: Specificity measures how much of the true negatives the model correctly identifies. The specificity value was given as 1.0, meaning that the model correctly identified the negative class.
These results provide a reference point for evaluating the performance of the model and for subsequent hyperparameter adjustments. Finally, validation accuracy measures overall model performance. However, considering the accuracy rate, it seems that the model only predicts the majority class. This shows that the accuracy

rate can be misleading in case of class imbalance. In particular, it may be important to increase the sensitivity of the model and reduce the number of false positives (FP).

6.1.2 Hyperparameter Tuning

In this study, the GridSearchCV (Grid Search Cross Validation) method was used for hyperparameter tuning and model evaluation for the SVM (Support Vector Machine) model. The 'param_grid' variable is a dictionary containing different combinations of the model's hyperparameters. Within this dictionary, the most appropriate combinations of hyperparameters are selected and applied to the model to maximize the performance of the model. The evaluated hyperparameters are as seen in *Fig.5*.

```
from sklearn.model_selection import GridSearchCV
param_grid = {
    'C': [0.1, 1, 10, 100],
    'gamma': [0.1, 0.01, 0.001],
    'kernel': ['linear', 'rbf'],
    'class_weight': ['balanced', None]
}
```

Fig.5

The reasons for selecting the parameters used are as follows:

Class_weight: As mentioned in *Section 2.3*, a significant imbalance is observed among the data classes. This imbalance in distribution can lead to unfair behavior of the model and result in bias. To address this imbalance, it was deemed appropriate to use hyperparameter tuning methods. When creating a fair model, it is important to use class weights where the rare classes are given more weight, thus ensuring balanced adaptation to the model.

C: The C parameter determines how flexible or rigid your SVM model will be. Larger C values cause the model to fit more closely to the training data, thereby reducing errors in the training data. However, this can lead to overfitting. Smaller C values allow the model to fit more generally, but accept some errors. When creating a fair model, it is important to choose an appropriate C value to avoid overfitting.

Gamma: Larger gamma values narrow the influence area of each training example, making the classification boundary more distinct. Smaller gamma values widen the influence area, resulting in a softer classification boundary. When creating a fair model, it is important to choose an appropriate gamma value to create a balanced classification boundary.

Kernel: When creating a fair model, it is important to choose an appropriate kernel type that will properly separate the data and make a fair distinction between classes.

```
Best Parameters: {'C': 10, 'class_weight': None, 'gamma': 0.01, 'kernel': 'rbf'}
Tuned SVM Model Accuracy: 0.8333333333333334
Confusion Matrix:
[[363  60]
 [ 40 137]]
Precision: 0.840177098159741
Recall: 0.8333333333333334
F1 Score: 0.8357719050640288
Sensitivity (Recall): 0.7740112994350282
Specificity: 0.8581560283687943
```

Fig.6

The result obtained in the performance metrics control after hyperparameter tuning is as shown in *Fig.6*. Interpretations of the new results can be made as follows:

Accuracy: The accuracy of the new model has increased significantly compared to the old model. This indicates that the model can generalize better and make more accurate predictions.

Confusion Matrix: When the complexity matrix of the new model is examined, it is seen that the number of misclassifications has decreased. In particular, there is a reduction in the number of false positives and false negatives, indicating that the model makes more stable and accurate predictions.

Precision: The precision of the new model has increased. This indicates that the rate at which samples predicted as positive are actually positive increases and the number of false positives decreases.

F1 Score: The F1 score of the new model is calculated as the harmonic average of precision and recall metrics. This metric indicates that the model is low in both false positives and false negatives.

Sensitivity (Recall) and Specificity:Sensitivity (Recall) and specificity values are indicators of the balance between classes. The sensitivity and specificity values of the new model show a more balanced performance, indicating that the model can classify both classes in a balanced way.

As a result, the new model's higher performance metrics such as accuracy, precision, sensitivity and F1 score indicate that it is a more balanced and reliable model. These improvements show that the model provides better learning and makes more accurate predictions.

6.1.3 Fairness Performance

Through literature review aimed at understanding various metrics and libraries available in Python to measure fairness, it has been concluded that the most common metrics, as identified by Köchling's (2020)study, are "Demographic Parity" and "Equalized Odds". Summarizing these concepts:

Demographic Parity:
Definition: Requires that positive outcomes (e.g., being selected for a job interview) are equally distributed among different demographic groups.
Explanation: Implies that the algorithm should provide similar opportunities or outcomes to individuals from different demographic groups.
Example: Achieving demographic parity in a job hiring algorithm would mean that the selection rates for job interviews are equal across different demographic groups.

Equalized Odds:
Definition: Aims to ensure equal treatment in terms of both false positives and false negatives across different demographic groups.
Explanation: Requires the algorithm to have similar false positive and false negative rates for different demographic groups.
Example: If equalized odds are achieved in a job hiring algorithm, it means the algorithm makes no biased decisions in any demographic group, avoiding both false positives and false negatives.

```
nographic Parity Ratio (Nationality Belgian): 0.7955801104972375
nographic Parity Ratio (Nationality Dutch): 0.921678533555522
nographic Parity Ratio (Nationality German): 0.7222099447513813
ualized Odds Difference (Nationality Belgian): 0.1046119235095613
ualized Odds Difference (Nationality Dutch): 0.006131549609810438
ualized Odds Difference (Nationality German): 0.11821705426356588
nographic Parity Ratio (Gender Female): 0.7256863162374974
nographic Parity Ratio (Gender Male): 0.704
nographic Parity Ratio (Gender Other): 0.5491841491841492
ualized Odds Difference (Gender Female): 0.0986895986895987
ualized Odds Difference (Gender Male): 0.10466988727858284
ualized Odds Difference (Gender Other): 0.14457831325301204
```

Fig.7

These concepts are crucial for promoting fairness in algorithmic decision-making processes and mitigating biases. Python fairlearn library was used to calculate these values on the established SVM model.The results are as in *figure 7*
Demographic Equality Ratio:
The demographic parity ratio of 0.726 for “Gender Female” indicates that women are slightly more likely to be hired than men.
The demographic equality ratio of 0.922 for “Nationality Dutch” indicates that the Dutch are more likely to be hired than other groups.

Equalized Probability Difference:
The equalized odds difference of 0.105 for “Gender Male” indicates that males are more likely to experience false positives than other groups.
The equated odds difference of 0.118 for “Nationality German” indicates that Germans are more likely to suffer false positives than other groups.
These metrics are important to ensure that algorithms treat demographic groups equally and that no group is privileged over others. Lower equalized probabilities difference and more balanced demographic equality ratios indicate the presence of a fairer algorithm, which is important to promote fair behavior in algorithmic decision-making processes.

6.1.4 Testing Model Performance

```
Test Accuracy: 0.8416666666666667
Test Confusion Matrix:
[[377  42]
 [ 53 128]]
Test Precision: 0.8393969448244413
Test Recall: 0.8416666666666667
Test F1 Score: 0.8402113765482434
Sensitivity (Recall): 0.7740112994350282
Specificity: 0.8581560283687943
```

Fig.8

The model has been tested on the test dataset, and its performance has been evaluated from various perspectives.

When looking at the confusion matrix, the important values in terms of fairness are *false negative and false positive.* The false negative rate measures the likelihood that candidates who are actually suitable for the job will be mistakenly rejected. The model's low false negative rate indicates that suitable candidates are less likely to be mistakenly rejected, thus providing equal opportunities to candidates. The false positive rate measures the probability that candidates who are not actually suitable for the job will be mistakenly hired. In this case, it is preferable to have a low false positive rate. The model's low false positive rate indicates that the hiring process is treated fairly and the likelihood of unsuitable candidates being mistakenly hired is reduced.Values seen as 53 and 42, respectively, are improvable.

The *accuracy* rate of 0.842 indicates that the model correctly classified approximately 84.2% of the samples in the test set. This is an accuracy rate above expectations, considering the requirements.The *sensitivity* value is 0.842, which indicates that the model correctly identifies approximately 84.2% of true positives. The *F1 score* is 0.840, which indicates that the model achieves good balance and has a balanced performance in terms of both precision and sensitivity. The *sensitivity* value is 0.774. The *specificity* value is 0.858, indicating that the model correctly identifies about 85.8% of true negatives. For details see *Fig.8*.

The results of the metrics used for fairness performance are presented in *Fig.9.*

Gender Comparison:

The demographic parity ratios for males are slightly higher compared to females and individuals with other genders,

indicating potential disparities in hiring rates.The equalized odds differences show variations across gender groups, with the "Other" category exhibiting a notably higher difference compared to females and males.

Nationality Comparison:

The demographic parity ratios for Belgian and Dutch nationalities are higher compared to German nationality, suggesting potential differences in hiring rates.

The equalized odds differences show relatively lower disparities across nationality groups, with Belgian nationality having the lowest difference.

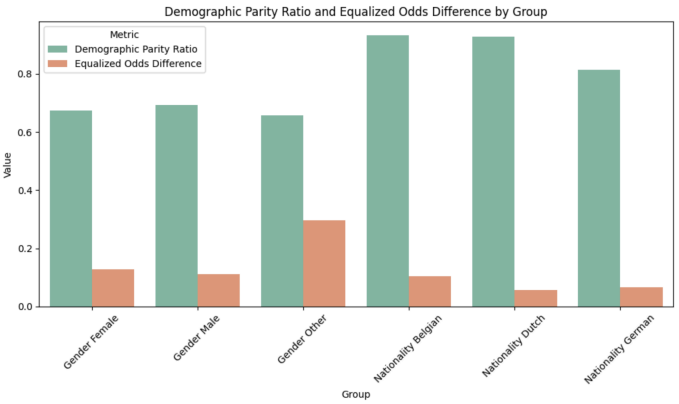
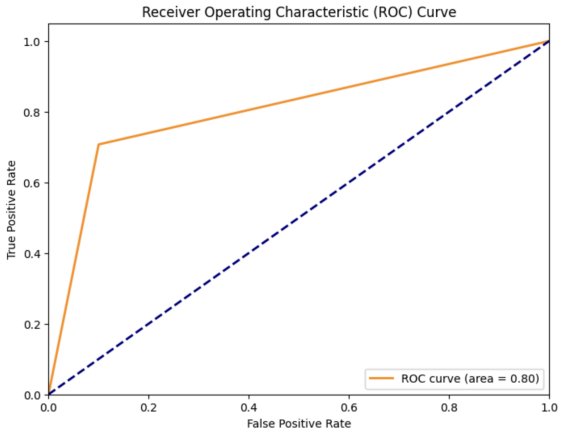


Fig.9

Overall, these metrics provide insights into the fairness of the model's predictions concerning gender and nationality. More comprehensive results will be obtained when the data is compared with the data of other models.



You see the ROC-AUC curve graph in *Fig.10.* We can interpret the performance of the model through this graph. The area under the ROC curve (AUC) value was calculated as 0.80. This value shows that the model has a high ability to distinguish positive and negative classes. It has been observed that the True Positive Rate (TPR) values in the ROC curve increase in the right direction. This indicates that the model classifies positive examples correctly and false positives are few. The ROC curve of the classification model shows that the model

successfully distinguishes positive and negative examples and has good overall performance.

The learning curve can be seen in *figure 11*. In this learning curve, the x-axis represents the 'Training Set Size' and the y-axis represents the 'Score' (could be accuracy, F1-score, etc.). The thick line indicates the 'Training score' and the dotted line indicates the 'Cross-validation score'. As the training set size increases, the training score also increases. This is expected, as a larger dataset allows the model to learn more patterns.The cross-validation score remains relatively stable as the training set size increases. This indicates that the model is not overfitting or underfitting the data, as the performance on the unseen validation set does not change drastically.Both the training and cross-validation scores are high (close to 1.00) for larger training set sizes. This suggests that the SVM model with a linear kernel performs well on the dataset. The gap between the training score and cross-validation score is small, indicating that the model generalizes well to unseen data. In summary, the learning curve suggests that the SVM model with a linear kernel is well-generalized and unlikely to benefit significantly from additional training data. The model has a high performance on both the training set and the validation set.

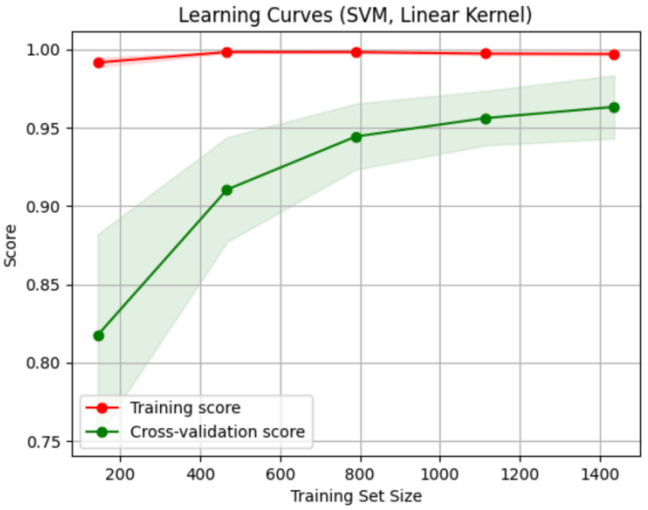


Fig.11

6.2 Random Forest

6.2.1 Model Building and Raw Performance

```
andom Forest Model Accuracy (Base): 0.865
onfusion Matrix:
[383  40]
[ 41 136]]
ensitivity (Recall): 0.768361581920904
pecificity: 0.9054373522458629
lassification Report for Random Forest Model (Base):
              precision    recall  f1-score   support

           0       0.90      0.91      0.90       423
           1       0.77      0.77      0.77       177

    accuracy                           0.86       600
   macro avg       0.84      0.84      0.84       600
eighted avg       0.86      0.86      0.86       600
```

Fig.12

Directly the same steps were applied in both machine learning models. The first step is to build the model and review the raw performance data. *See fig.12*

Accuracy: The model has a high rate of correct prediction (86.5%). However, if there is class imbalance (for example, one of the classes contains many more examples than the other), accuracy alone may not be sufficient as a performance metric. Class imbalance is evident in the dataset used. Therefore, accuracy alone is not a sufficient performance metric for analysis.

Confusion Matrix: The model predicted 383 true negatives (TN), 40 false positives (FP), 41 false negatives (FN), and 136 true positives (TP). High TN and TP values indicate that the model performs well in correctly classifying both negatives and positives. Considering that FP and FN values are important points in this study where fairness was examined, the model did a good job.

Sensitivity (Recall): Indicates that the model correctly identified approximately 76.8% of true positives. Sensitivity is a measure of cases where the model misses true positives.

Specificity: Indicates that the model correctly identifies approximately 90.5% of true negatives. Specificity is the measure of cases where the model misses true negatives over false positives.

Classification Report: By looking at class-based precision, recall and f1-score metrics, how each class is classified by the model can be examined in more detail. Performance is quite high for class 0 (precision, recall and f1-score are 0.90). However, for class 1 the performance is slightly lower (precision and recall 0.77, f1-score 0.77).

6.2.2 Hyperparameter Tuning

```
rom sklearn.model_selection import GridSearchCV

aram_grid_rf = {
    'n_estimators': [50, 100],
    'max_depth': [None, 10],
    'min_samples_split': [5, 10],
    'class_weight': ['balanced', None],
    'bootstrap': [True],
    'max_features': ['sqrt'],
    'min_samples_leaf': [2, 4],
    'max_leaf_nodes': [None, 10],
    'criterion': ['entropy']
```

Fig.13

GridSearchCV method was used in the Hyperparameter tuning process for the Random Forest model, as in SVM. In this process, it was aimed to optimize the model by selecting different parameters. Since the selected parameters vary depending on the model and problem, these parameters were chosen carefully. *See fig.13*. The adjusted parameters and the ***reasons for their selection*** are explained below:

class_weight: Used to address class imbalance in the data set. Especially in datasets with unbalanced class distributions, such as the recruitment dataset, this parameter has been used to help correct the imbalance and ensure that the model predicts each class equally well.

criterion: The criterion used to split trees affects the performance of the model. 'Entropy' or 'gini' metrics represent different information gains or classification accuracy and have been used to help the model make the best decisions.

min_samples_split: It is aimed to help the model learn in a more general way by determining the minimum number of samples in each node. This can help prevent overfitting and give the model a more balanced performance.

n_estimators: It is an important factor that determines the complexity of the model. Using multiple trees can ensure a more general form of the model and reduce the risk of overfitting.

max_depth: The maximum depth of trees is set to limit the complexity of the model. Deep trees can lead to overfitting, so maximum depth should be kept in check.

These parameters were chosen to ensure that the Random Forest model provides fair and balanced performance on a sensitive data set such as the recruitment dataset. Additionally, 'best_params' was used to properly tune the model to reduce the risk of overfitting and achieve good overall performance.The result obtained in the performance metrics control after hyperparameter tuning is as shown in *figure 14*. Interpretations of the new results can be made as follows:

```
Best Parameters for Random Forest: {'bootstrap': True,
Tuned Random Forest Model Accuracy: 0.8733333333333333
Classification Report for Tuned Random Forest Model:
              precision    recall  f1-score   support

           0       0.94      0.87      0.91       423
           1       0.74      0.88      0.80       177

    accuracy                           0.87       600
   macro avg       0.84      0.87      0.85       600
weighted avg       0.88      0.87      0.88       600
```

Fig.14

When switching to the Random Forest model with hyperparameter adjustment, the accuracy rate improved to 87.3%. When the classification report is examined, precision increased to 94% for class 0, while recall decreased to 87%. For Class 1, precision decreased to 74%, while recall increased to 88%.

According to this comparison, the hyperparameter-tuned model provided an increase in overall accuracy. However, it is noteworthy that the precision value decreases for class 1. Both models perform well, but further improvements may be needed to achieve balanced performance for certain classes.

6.1.3 Fairness Performance

```
emographic Parity Ratio (Gender Female): 0.6983716635890549
emographic Parity Ratio (Gender Male): 0.6814159292035399
emographic Parity Ratio (Gender Other): 0.6154649947753396
qualized Odds Difference (Gender Female): 0.08968058968058967
qualized Odds Difference (Gender Male): 0.09541062801932365
qualized Odds Difference (Gender Other): 0.10344827586206895
mographic Parity Ratio (Nationality Belgian): 0.8385093167701863
mographic Parity Ratio (Nationality Dutch): 0.9420289855072463
mographic Parity Ratio (Nationality German): 0.774796106287819
ualized Odds Difference (Nationality Belgian): 0.07855268091488564
ualized Odds Difference (Nationality Dutch): 0.021627647714604237
ualized Odds Difference (Nationality German): 0.04844961240310078
```

Selected specific performance metrics were applied on the Random Forest model to examine whether sensitive features were handled fairly. The results obtained are presented in *Figure.15*. These data can be interpreted as follows:

First, *demographic parity* rates were examined. It was determined that the demographic parity rate for women was 69.8% and for men was 68.1%. These values show that the probability of women and men

receiving positive results is above the general average. It was observed that the Other category had a slightly lower chance of being hired. The demographic parity rates for Belgians and Germans were determined to be 83.9% and 77.5%, respectively. These results indicate that certain demographic groups are less likely to experience positive outcomes than the overall average.

Then, *equal opportunity* differences were evaluated. It was determined that the equal opportunity gap for women was 9.0% and for men was 9.5%. These values indicate that the difference between false positive and false negative rates is high in certain demographic groups. Similarly, equal opportunity gaps for Belgians and Germans were found to be 7.9% and 4.8% respectively.

This analysis provides an important perspective to evaluate whether the model is fair to certain demographic groups. The results show a lack of equality and opportunity for certain demographic groups. It is important to take these findings into account to make the model more fair and balanced.

6.1.4 Testing Model Performance

```
st Accuracy (Random Forest): 0.735
st Confusion Matrix (Random Forest):
413   6]
153  28]]
st Precision (Random Forest): 0.7579926210766992
st Recall (Random Forest): 0.735
st F1 Score (Random Forest): 0.664181088419313
nsitivity (recall) (Random Forest): 0.15469613259668508
ecificity (Random Forest): 0.9856801909307876
```

Fig.16

The model has been tested on the test dataset, and its performance has been evaluated from various perspectives. Performance metrics of the tested model are presented in *Fig.16.*

First, the model performed quite well on the validation dataset. The accuracy in the training set was quite high at 0.87, and despite the imbalance between classes, it showed a very balanced performance.

However, we observe a decrease in the model's performance on the test set.The *accuracy* value in the test set was below the value obtained in the training set (0.735). This may indicate that the model has some difficulties in transferring its generalization ability from the training set to the test set. This may be due to factors such as overfitting or different data distribution in the test set.

When the *confusion matrix* is examined, we can see that the model has a particularly high number of false positive cases. This may indicate that the model is better at predicting the negative class, but poor at predicting the positive class.

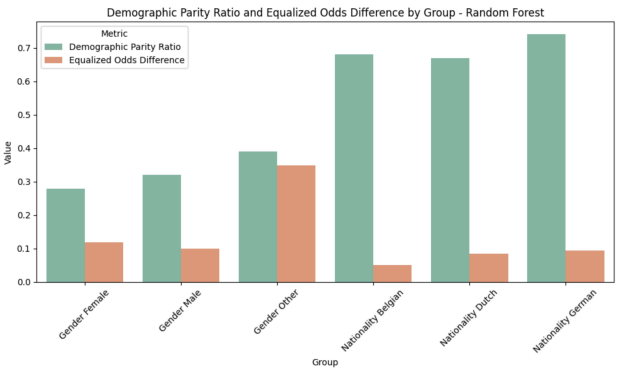
When metrics such as precision, recall and F1 score are examined, we can see that the model exhibits lower performance in predicting the positive class. This causes the sensitivity (recall) of the model to be low. In particular, the sensitivity value (15.47%) is quite low, indicating that the model is weak in detecting the positive class. Finally, the specificity of the model is quite high (98.57%). This indicates a high ability to accurately detect the negative class.

These results show that the model performed well in the training set, but could not show the expected success in the test set and was especially weak in predicting the positive class. To improve this situation, it may be necessary to revisit the model and prevent possible overfitting. It is also important to train and tune the model in a more balanced way, considering the imbalance of the dataset.

The results of the metrics used for fairness performance are presented in *Fig.17.*

Demographic Parity Analysis by Gender: The demographic parity analysis by gender reveals parity ratios of 0.278 for females, 0.321 for males, and 0.390 for other genders. These ratios indicate some challenges in achieving equal distribution of the target variable across different gender groups.

Equalized Odds Analysis by Gender: The equalized odds analysis by gender shows differences of 0.118 for females, 0.100 for males, and 0.349 for other genders. These values suggest significant disparities in misclassification rates across different gender groups.



Demographic Parity Analysis by Nationality: The demographic parity analysis by nationality yields parity ratios of 0.680 for Belgians, 0.670 for Dutch, and 0.741 for Germans. These ratios indicate substantial differences in the distribution of the target variable across different nationality groups.
Equalized Odds Analysis by Nationality: The equalized odds analysis by nationality reveals differences of 0.051 for Belgians, 0.084 for Dutch, and 0.094 for Germans. These differences highlight notable variations in misclassification rates across different nationality groups.

In this specific ROC curve, the x-axis is represented by the False Positive Rate (FPR), while the y-axis is represented by the True Positive Rate (TPR). The ROC curve, depicted as a blue line, illustrates the performance of the Random Forest model across various classification thresholds. The area under the ROC curve (AUROC) measures 0.57, signifying the model's capacity to differentiate between positive and negative classes. A value of 0.5 denotes random guessing, whereas a value nearing 1 indicates superior performance. With a value of 0.57, the model's performance slightly surpasses random guessing but still exhibits room for enhancement. The presence of the ROC curve above the red dashed line at 0.5 on the y-axis suggests that the Random Forest model outperforms a random classifier. Furthermore, the plot demonstrates that as the FPR increases, the TPR also increases, albeit at a slower pace. This observation indicates the Random Forest model's proficiency in identifying true positives amidst a higher occurrence of false positives, though the overall performance falls short of optimal levels. In summary, the ROC curve for the Random Forest model implies that while the model's performance marginally exceeds random guessing, opportunities for improvement remain. To elevate the model's effectiveness, potential measures include adjusting the classification threshold, exploring alternative machine learning algorithms, or implementing feature engineering techniques to refine the model's capacity to discern between positive and negative classes.

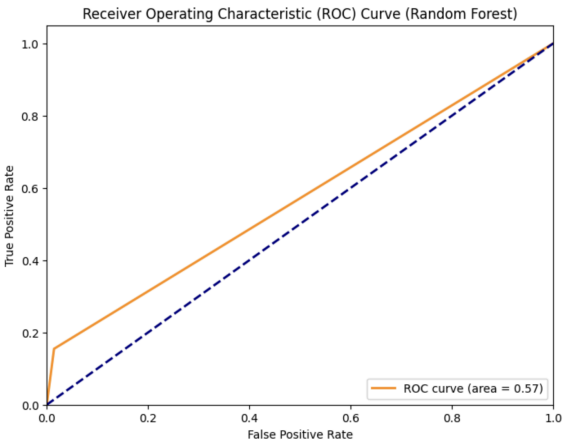


Fig.18

As the training set size increases, both the training score and cross-validation score decrease. This is unusual, as it is generally expected that the training score will increase and the cross-validation score will approach a value close to the training score. When the training set size is 1400, both the training score and the cross-validation score are around 0.7, indicating that the performance of the model is stable at this point.
The difference between the training score and the cross-validation score is relatively small, indicating that the Random Forest model does not overfit the data.

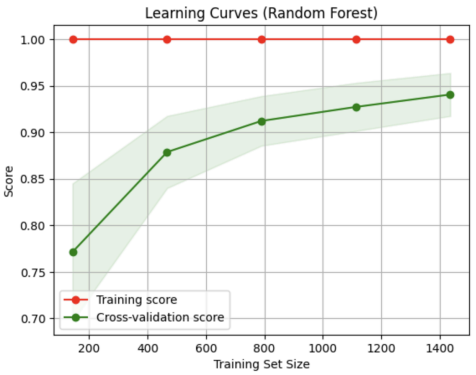


Fig.19

The learning curve shows that there is no significant improvement in the performance of the model as the training set size increases, indicating that the Random Forest model has reached its maximum performance on this dataset and further increasing the training set size will not yield better results.However, the overall performance of the model is not particularly strong (cross-validation score around 0.7)

6.3 Multi-layer Perceptron

6.2.1 Model Building

The MLP model is defined using the Sequential API. The model has three hidden layers and one output layer. The hidden layers contain 64, 128, and 64 neurons, respectively. The LeakyReLU activation function was used in each hidden layer. In the output layer, sigmoid activation function was preferred for binary classification problems.

6.2.2 Hyperparameter Tuning

```
MLP Model Accuracy (Tuned): 0.8533333333333334
Classification Report for MLP Model (Tuned):
              precision    recall  f1-score   support

           0       0.91      0.88      0.89       423
           1       0.73      0.79      0.76       177

    accuracy                           0.85       600
   macro avg       0.82      0.84      0.83       600
weighted avg       0.86      0.85      0.85       600

Confusion Matrix for MLP Model (Tuned):
[[372  51]
 [ 37 140]]
Sensitivity (Recall) for MLP Model (Tuned): 0.7909604519774012
Specificity for MLP Model (Tuned): 0.8794326241134752
```

Fig.20

Hyperparameters were determined using GridSearchCV. A parameter grid has been defined that allows hyperparameters such as 'batch_size' and 'epochs' to be searched among specific values. This method was made to reach the best combination of hyperparameter values. The ***reason for choosing these parameters*** as follows: While batch size indicates the number of samples the model will use during training, epoch refers to the number of times the entire data set will be used during the training process. These choices affect the overall performance of the model and its ability to produce fair results. In determining the batch size, it was taken into account that although larger batch sizes have the potential to obtain better generalizable results, excessively large batch sizes may cause unnecessary resource usage. Therefore, it is important that the batch size used in the recruitment process provides a reasonable balance. Similarly, in determining the number of epochs, the model's ability to perform adequately on the training set and generalize to new data was taken into consideration. The selection of the best model was made in the 5 fold by cross-validation. As a result, the best parameters were *'batch_size': 64, 'epochs': 10*.

The result of after tuning performance is presented in *figure.20*. These results show the performance of the MLP model after hyperparameter tuning. The accuracy of the model was measured as 0.853, which shows that the model achieved 85.3% success in correctly classifying the data. When we look at the classification report, a high precision and recall rate was obtained for class 0. However, for class 1 these rates are slightly lower, which may indicate that the model is less successful in recognizing class 1. Overall, the weighted avg metric shows that the model classifies classes in a balanced way. However, in case of class imbalance, other metrics should also be considered to evaluate the performance of the machine learning model.The model made 372 true negative (TN) predictions and 140 true positive (TP) predictions. However, there were 51 false positive (FP) and 37 false negative (FN) predictions. We can say that the model is more successful in identifying candidates who are not qualified to be hired. The Sensitivity (Recall) metric measures the rate of true positives (TP) and in this case it was calculated as 0.791. That is, the model correctly identified 79.1% of true positives.

The specificity metric measures the rate of true negatives (TN) and is calculated as 0.879. This shows that the model correctly identified 87.9% of true negatives.

These metrics show that the model performs well in classifying classes in a balanced manner and effectively identifies both positive and negative classes.

6.2.3 Optimization and Activation Function Analysis

```
Best Activation Function: relu
Best Accuracy: 0.8566666666666667
```

Fig.21

The model was compiled using the Adam optimizer. Adam optimizer is an adaptive momentum optimization algorithm that increases the efficiency of gradient descent optimization algorithms. The binary_crossentropy loss function is a commonly used loss function for binary classification problems.

Three activation functions are defined for the model. These functions are 'relu', 'sigmoid', 'tanh' and LeakyReLU(). With the created code, an environment was created in which all activation functions were tested. The results of the experiments are presented in *figure.21*. The high accuracy achieved with the relu activation function suggests that it was able to effectively capture the non-linearities present in the dataset, leading to better performance compared to other activation functions. Therefore, relu was selected as the best activation function for this MLP model, resulting in an accuracy of 85.67%.

6.2.4 Fairness Performance

```
graphic Parity Ratio (Nationality Belgian): 0.9941860465116278
graphic Parity Ratio (Nationality Dutch): 0.8337425971399681
graphic Parity Ratio (Nationality German): 0.7304327808471455

graphic Parity Ratio (Gender Female): 0.7616707616707616
graphic Parity Ratio (Gender Male): 0.7394957983193278
graphic Parity Ratio (Gender Other): 0.5666185666185667

ized Odds Difference (Gender Female): 0.040684838927686415
ized Odds Difference (Gender Male): 0.04961678096006454
ized Odds Difference (Gender Other): 0.12643678160919547

lized Odds Difference (Belgian): 0.170859538784067
lized Odds Difference (Dutch): 0.04080267558528429
lized Odds Difference (German): 0.12643678160919547
```

Fig.22

The MLP model was evaluated in terms of fairness metrics. The results obtained are seen in *figure.22.* If we summarize the result:

Based on *demographic parity*:

The ratio between men and women is quite close to each other, but the "other" category has a lower ratio. This indicates that the model behaves more balanced between men and women, but with lower balance among other genders. The Demographic Parity Ratio in Belgium is very close to almost indicating that the model is more balanced towards Belgian citizens. However, in the Netherlands and Germany these rates are slightly lower, which may indicate that the model is a little more unbalanced towards these groups.

If we evaluate on the basis of the *Equalized Odds* metric:

Equality Difference values are quite low, which shows that the model has a similar sensitivity between different genders. However, for the "other" category this value is slightly higher, which may indicate that the model has more imbalances regarding other genders. The Equality Odd for Belgium is quite high, which may indicate that the model has different sensitivity values among Belgian citizens. On the other hand, for the Netherlands and Germany these values are lower, which may indicate that the model exhibits more consistent behavior across these groups.

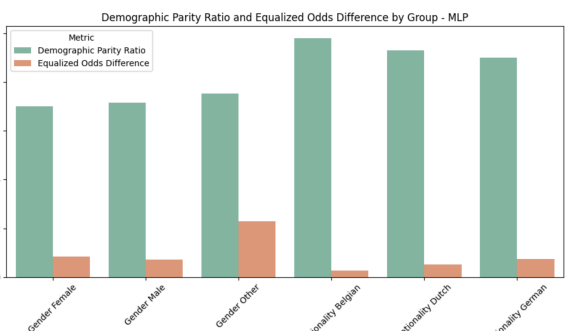
6.2.5 Testing Model Performance

```
19/19 [==============================] - 0s 3ms/step
Test Accuracy (MLP): 0.8416666666666667
Test Confusion Matrix (MLP):
[[365  54]
 [ 41 140]]
Test Precision (MLP): 0.8455095813654291
Test Recall (MLP): 0.8416666666666667
Test F1 Score (MLP): 0.8431636363636363
Sensitivity (Recall) (MLP): 0.7734806629834254
Specificity (MLP): 0.8711217183770883
```

Fig.23

The model has been tested on the test dataset, and its performance has been evaluated from various perspectives. Performance metrics of the tested model are presented in *Figure.23* The MLP model achieved an accuracy of 84.17% on the test set. While this shows that the model performs well on the new data, it is slightly lower than the 85.33% accuracy obtained on the validation set. Precision and recall values for the test set were determined as 84.55% and 84.17%, respectively, indicating the ability of the model to accurately identify and recall positive samples. However, a slightly lower precision value is observed compared to the validaiton set. The F1 score is 84.32%. Sensitivity (recall) and specificity values were calculated as 77.35% and 87.11%, respectively. This indicates that the model has relatively stable performance in correctly classifying positive and negative classes.

Since the accuracy rate on the test set is high, it can be said that the model has a generalizable performance. However, some decreases in precision and recall values are observed. These decreases indicate that the model has difficulty classifying certain classes or does not show a balanced performance. Therefore, the model requires improvement or further tuning in certain areas.



The results of fairness metrics of the model are examined for the Sensitive feature, obtained are presented in *fig.24.* These results show demographic parity ratios and equalized opportunity gaps on the MLP model's test data

set. On a gender basis, the demographic parity ratio was calculated as 0.701 for women, 0.715 for men and 0.752 for other genders. These rates show that there is no particular equality between different gender groups, but there is no major difference either. However, it is not at the desired demographic parity value. Additionally, when equalized opportunity differences are examined, it is seen that it is 0.085 for women, 0.073 for men and 0.229 for other genders. It can be said that there is negative discrimination against other genders. Similarly, on a national basis, the demographic parity rate was found to be 0.980 for Belgians, 0.930 for the Dutch and 0.899 for the Germans. However, when looking at the equalized opportunity gap, it is seen that it is 0.026 for the Belgians, 0.051 for the Dutch and 0.073 for the Germans. These results demonstrate differences in the model's performance across different demographic groups.

Fig.24

The ROC curve of the MLP model can be seen in *fig.25*. The features of the ROC curve are as follows:

ROC curve (blue line) shows the performance of the model at different classification thresholds. The area under the ROC curve (AUROC) is a measure of the overall performance of the model. An AUROC value of 1.0 indicates a perfect classifier, while a value of 0.5 indicates a completely random classifier. In this case, an AUROC value of 0.82 indicates that the MLP model is performing well, but there is still room for improvement

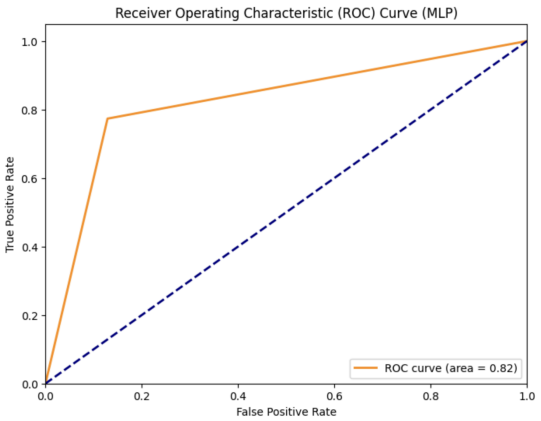


Fig.25

Looking at the graph, it can be seen that as FPR increases, TPR also increases. This shows that as the model makes its classification threshold more lenient, it correctly identifies more positive examples, but comes at the cost of increased false positives.

In summary, the ROC curve of the MLP model shows that the model performs well overall, with an AUROC of 0.82. However, there is still room for improvement, especially to adjust the classification threshold to optimize the balance between TPR and FPR.

Finally, the learning curve drawn for the MLP model can be seen in *Fig. 26.* From the MLP learning curve the following observations were made:

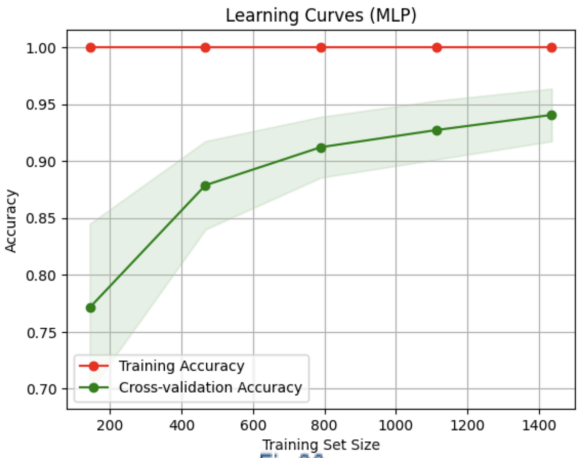


Fig.26

The x-axis represents the Training Set Size and takes values between 200 and 1400. The Y-axis represents the Accuracy value and takes values between 0.7 and 1.0. There are two lines in the learning curve: training accuracy and cross-validation accuracy lines. As Training Set Size increases, both training accuracy and cross-validation accuracy increase. Training accuracy is consistently higher than cross-validation accuracy, indicating that the model may be slightly overfitting. Overfitting occurs when the model is too complex and performs well on training data but does not generalize well to unseen data. The difference between training accuracy and cross-validation accuracy is quite small, especially when the Training Set Size is 1000 or more. This suggests that further increasing the Training Set Size may not significantly improve the performance of the model.

In summary, the learning curve for the MLP model shows that the performance of the model improves as the Training Set Size increases, but there is a slight overfitting issue. To address this, techniques such as regulation, release, or early stopping may be considered. You may also want to evaluate the model with other metrics, especially when accuracy is unbalanced, such as precision, recall, and F1 score, as accuracy may not be a reliable indicator of performance in these cases.

5. Model Comparison and Results

Self Reflection: *It is obvious that two of the established models are overfitting. These models are Random Forest and MLP models. Necessary precautions to avoid this situation will be taken in future studies.*

1. *Validation Performance:*

- The Random Forest model achieved the highest success with an accuracy rate of 87.3%. Then, the Multi-layer Perceptron (MLP) model and Support Vector Machine (SVM) model came second and third, respectively, with accuracy rates of 85.3% and 84.4%. However, these performances did not hold up when the models were faced with real-world data (Test Group).

2. *Fairness Performance:*

- In terms of fairness performance, the Random Forest model has the lowest equalized odds difference and the highest demographic parity rates. This indicates that the model performs more evenly across different demographic groups.
- The MLP model has equalized odds difference and demographic parity rates between different demographic groups. However, equalized odds difference appear to be slightly higher for other genders.
- The SVM model has lower equalized odds difference and demographic parity rates between certain demographic groups. However, the model appears to show imbalances for some demographic groups.

In this context, the Random Forest model has been the most successful model and one that minimizes discrimination between groups.

3. *Test Performance:*

- According to the evaluations made on the test dataset, SVM and Multi-layer Perceptron models were the best performing models. The accuracy values of both were measured as 0.8416. The difference between them is that in the SVM model, the model that passes from the measurement on the validation dataset to the real world data performs better, while in MLP the situation is the opposite. On the other hand, Random Forest has significantly reduced its performance with an accuracy value of 0.735. There may be various reasons for this situation. These reasons can be listed as overfitting, data snooping, or imbalance in data distribution.

-Comments on other metrics are explained in detail in the relevant sections.

4. *Optimization and Hyperparameter Tuning:*

- In terms of hyperparameter tuning and optimization, the Random Forest model gave the best results. Hyperparameter tuning using GridSearchCV increased the performance of the model and ensured a balanced performance.
- The MLP model has achieved good performance by tuning certain hyperparameters, but it is not as stable as the Random Forest model.

Overall, the Random Forest model gave the best results in terms of performance and fairness. However, other models have significant advantages in certain situations, and which model to choose will depend on the use case, the characteristics of the dataset, and the priorities of the projects.

6. Conclusion

This study aimed to establish some of the machine learning and deep learning models that can be used in recruitment processes and evaluate whether they are fair or not. First, the data set was loaded and the preprocessing steps were performed. Then, feature engineering operations were performed on the data set and sensitive features were determined.

Different machine and deep learning models (SVM, Random Forest and MLP) were selected and the performance of these models was evaluated in terms of both overall accuracy and fair decision making. Hyperparameter tuning for the models was performed using GridSearchCV. In particular, the results revealed that the Random Forest model showed a higher accuracy and fair decision-making ability. It was concluded that if overfitting is prevented, Random Forest may be the most suitable model for the dataset used. Additionally, the fairness of the models was measured using fairness metrics such as Demographic Parity and Equalized Odds. The results obtained show that the selected models provide fair treatment between demographic groups.

In conclusion, this study goes a step forward in assessing whether machine learning models used in recruitment processes are fair. In the future, the results of this study can be expanded by collecting more data and using different fairness metrics.

7. Future Work

The results of past work and the mistakes made provided important clues for future research. Lessons should be learned from these mistakes and steps should be taken to obtain better results in future studies.

First, more care should be taken when choosing a model. Considering the characteristics and size of the data set, the most appropriate model must be determined. Additionally, it is necessary to control the complexity of the model and take necessary precautions to prevent overfitting.

Second, a more detailed preprocessing should be performed on the data set. Steps such as handling missing data, dealing with unbalanced classes, and feature selection should be done more carefully and the dataset should be cleaned better.

Third, evaluation metrics must be better selected and ensure that they accurately reflect the real-world performance of the model. This is important when determining the model's success criteria and interpreting the results.

Finally, larger data sets should be used in future studies and care should be taken to ensure that the data set is more balanced. Additionally, a more comprehensive cross-validation strategy should be adopted and the reliability of the results should be increased.

It is hoped that by learning from these mistakes and making improvements, more solid results will be obtained in future studies. These steps are important to improve the performance of the model and make fair and reliable decisions.

References

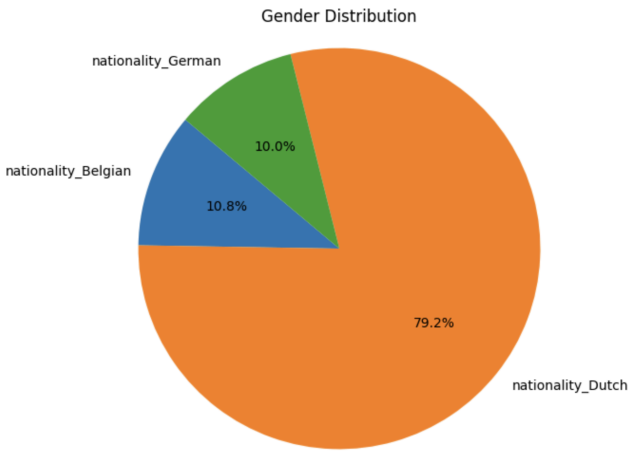
Grybauskas, A., Stefanini, A., & Ghobakhloo, M. (2022). Social sustainability in the age of digitalization: A systematic literature Review on the social implications of industry 4.0. *Technology in Society*, *70*, 101997. https://doi.org/10.1016/j.techsoc.2022.101997

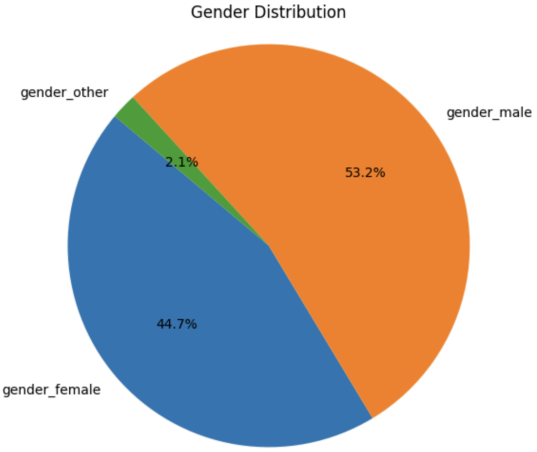
Lee, K. (2021). A Systematic Review on Social Sustainability of Artificial Intelligence in Product Design. *Sustainability*, *13*(5), 2668. https://doi.org/10.3390/su13052668

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2020). Highly accurate, but still discriminatory. *Business & Information Systems Engineering*, *63*(1), 39–54. https://doi.org/10.1007/s12599-020-00673-w

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Appendices





When analyzing the data included in the sensitive feature group, it becomes clear that there is a significant imbalance of data distribution between the groups. Imbalance between groups, such as shown in pie charts, can greatly affect the model's decision-making ability. Therefore, it is essential to create weight balance to ensure fairness to the data.