

# Fairness in Machine Learning Algorithms

**Rutvik Rahul Nate**

BSc H Computer Science

Supervisor: Dr Stef Garasto

Due Date: 25<sup>th</sup> April 2022

Word Count:10,200

# Abstract

The aim of the project is to identify fairness in the decision making of machine learning algorithms. Decisions play a crucial role in our lives, whether choosing a degree or determining a specific goal. Hence, these decisions must be correct and unbiased. The dataset utilised in this research is the adult income dataset from the University of California, Irvine repository, where the target values are the individual's income. Consider the following hypothetical situation with a bank that wishes to accept loan applications depending on an individual's income. If the borrower's income exceeds 50,000 dollars, the bank will approve the loan application. In this project, a streamlit application was constructed that allows users to input data and determine whether their annual income is higher than or less than \$50,000. The application is built on a threshold optimizer algorithm that considers 'sex' as a sensitive feature. Another tool used in this study is Fair-learn, which is a tool for analysing and reducing bias from the accuracy prediction by using fairness metrics like demographic parity in this case. In this project, threshold optimizer is an additional layer that is used on top of machine learning algorithms to ensure fairness. The threshold optimizer uses a fairness parameter called demographic parity to ensure that selection rates are equal.

# Preface

# Acknowledgement

I would like to express my sincere thanks to my personal supervisor, Dr Stef Garasto, and my second supervisor, Dr Yasmine Arafa, for their unwavering support, direction, and encouragement to ensure that my project is a success. Additionally, I would want to express my gratitude to my parents, family and friends for their constant support and aid during this process.

# Table of Contents

## Contents

<b>Abstract</b> .....	ii
<b>Preface</b> .....	iii
<b>Acknowledgement</b> .....	iv
<b>Table of Contents</b> .....	v
<b>1. Introduction</b> .....	1
<b>2. Literature Review</b> .....	2
2.1 Job Recruiting or Retaining Employees .....	2
2.2 Criminal Risk Assessment .....	3
2.3 College admissions .....	3
2.4 Loan approvals .....	4
<b>3. Analysis</b> .....	4
3.1 Methodology .....	4
3.2 Impact of Legal, Ethical, Social and Professional Issues .....	6
<b>4. Requirement Specification</b> .....	7
<b>5. Design</b> .....	8
5.1 Initial Design Stage 1 .....	8
5.2 Initial Design Stage 2 .....	8
5.3 Final design Stage 1 .....	9
5.4 Final design .....	9
5.5 Gantt chart time distribution for the project .....	10
<b>6. Implementation</b> .....	11
6.1 Proposed Product .....	12
6.2 Data collection and Preprocessing .....	12
6.3 Machine learning Models .....	12
<b>7. Testing and Integration</b> .....	12
7.1 Initial integration .....	13
7.2 Testing and Interpreting the results .....	13
7.3 Final testing .....	15
<b>8. Product evaluations</b> .....	19
<b>9. Further Development Options</b> .....	20
<b>10. Conclusion</b> .....	20
<b>References</b> .....	20

**Appendix- Code for the product ..... Error! Bookmark not defined.**

# 1. Introduction

Researchers acquire data in structured and unstructured formats every day and now have thousands of datasets. Datasets contribute to the development of Machine Learning by acting as a fuel for it. Because datasets are at the core of modelling, researchers begin by exploring these first. Identify critical relationships in the dataset and then use that data to solve problems. Cleaning, wrangling, and pre-processing of the data is required before introducing the data to our models. ML is very powerful as it enables cost savings and more efficient decision-making. Also, ML contributes to the enhancement of the consumer experience.

Decision making is essential as it could affect other people's lives, whether related to offering a job to a specific person or accepting a loan application for a client. Incorrect decisions can hinder a company's success and its employee's capacity to reach higher ambitions. Wrong decisions could also affect an individual's motivation regarding a specific task. Human decision making depends on intuition or experience from the past, which brings bias into the process. Bank loans, medical issue, hiring in businesses and organizations, and performance-based promotion are just a few instances of how external choices may affect people's lives.

Machine learning can improve decision-making by incorporating aspects important to decision-making that humans often neglect (Barocas, et al., 2019). Companies have opted for machine learning algorithms for the decision-making process. Having a highly qualified staff is critical since it allows the organisation to maintain a competitive advantage in the market. Schools often prefer to keep their best-achieving students and provide them with scholarships for further continuation in the same university. When making these decisions, schools are neutral or fair, and their decisions are based on a student's previous performance or engagement in other activities. On the other hand, companies rely on data that a human has extracted. Human judgments are used to train machine learning algorithms, and since human judgments are biased, bias is incorporated into the final system (Harris, 2020). Machine Learning can uncover statistical patterns in data comprising text and pictures that people may be unable to detect.

The decision-making process's quality is dependent upon the dataset's quality. Educational departments are under tremendous pressure to make correct admission decisions. But the data stored is not always sufficient; thus, it is critical that the data stored is evaluated and only relevant information is utilised to support the admission decision-making process (Nieto, et al., 2019).

Although algorithms may not recall instances of unfair prejudice, consumers do, which is why it is critical that these algorithms remain fair. These choices must be accurate and consistent across all individuals and groups. In the article (Angwin, et al., 2016), the authors discussed about the machine learning model that was used to calculate risk assesment for bail contenders using historical data. It was discovered that the data was skewed in predicting bail and was dependent on an indivisual's 'race'. Machine learning has the potential to increase bias in existing data, resulting in a viscous loop in the models. The authors in (Angwin, et al., 2016) gathered risk ratings for a significant number of persons convicted in Broward County, Florida, in 2013 and 2014, and investigated how often these individuals were charged with further offences the following year. The author applied the same benchmarks as the judicial community's machine learning systems. The findings were alarming, since just 20% of those forecasted to commit violent crime really did (Angwin, et al., 2016). This demonstrates the need of introducing fairness into the court system, since it may effect those who are not deserving of punishment as well.

Fairness can be applied during the pre-processing which is to change the data by keeping the sensitive features protected so that the machine learning models cannot learn from this specific attribute. This project code was useful in identifying the differences in accuracy, selection rate, and true positive rate between two groups. This is the difference between the possibilities that the model selects as a result of its decision. Consider the following scenario: in order to make predictions, you have the option to choose between a man and a female. The selection rate is defined as the difference between how frequently the model chose men against how often it chose females within a certain time period. It was possible to construct a plot after applying the same code to the final model and using the threshold optimizer as well. The project was developed using a proper gantt chart to note and manage the timing for every different parts of the project.

## 2. Literature Review

The bias of algorithms is now being studied by a significant number of experts worldwide in a variety of domains, including law, economics, and computer science (Bird, et al., 2019). Machine learning fairness has developed as a critical subject in computer science, as shown by the increase in the number of researchers working on this problem and the number of conferences organised in this domain (Bird, et al., 2019). As the number of articles on this subject has grown, several distinct concepts of fairness have developed (Marvin, et al., 2021).

Examples of instances in which there may be a lack of fairness. A business that uses machine learning to calculate insurance premiums inadvertently discriminates against the elderly. A credit card business utilised monitoring data to drive minorities to higher customised items rates.

In machine learning, fairness may be separated into two components. The first component consists of two tasks: detection of discrimination and eradication of discrimination, and in the second component, fairness can be classified as fairness in ranking, regression, and reinforcement learning (Marvin, et al., 2021).

Real-World situations or previous research in which fair decision making could be applied :

### 2.1 Job Recruiting or Retaining Employees

Academically skilled individuals are always seen as valuable assets to a business team. A bad recruitment decision may significantly affect an organisation's productivity and morale. As a result, it is crucial to attracting individuals with these potentials (Wang, et al., 2019). It also helps the company maintain an advantage in the market. The authors process data by examining the employee's background and educational status. The results provided by (Wang, et al., 2019) indicated that their models provided higher accuracy and better recall and precision scores by including personal background of an employee. Additionally, the authors stated that personal background characteristics are critical for classifying and forecasting employee growth, which is beneficial for prediction performance (Wang, et al., 2019). However, what if an employee performed better than predicted, wouldn't that affect the company because they might have fired the employee? This may have resulted in a significant loss for the corporation if a large number of similar employees with potential were dismissed. Thus, the loss may be minimised by integrating fairness to the machine learning models, since this would guarantee that employees are evaluated based on their historical performance data. The algorithm would predict and move on but the employee would still remember the outcome in the future.



Predicting employee retention is also a hot topic among researchers, as seen by the many studies on the subject. Another similar example where researchers worked on the employee retention problem and provided machine learning models with higher accuracy is in the paper (Marvin, et al., 2021). The researchers suggested a machine learning classification technique for predicting employee retention based on candidate-relevant features. They discovered that random forest classifiers and KNN provided the best accuracy, over 90% (Marvin, et al., 2021).

## 2.2 Criminal Risk Assessment

Machine learning algorithms have been popularly used to assess the risk of criminal activity by a defendant (Angwin, et al., 2016). These algorithms help the police or the judges to make a judgement based on a defendant's historical data. The algorithms are used for many other reasons as well such as facial recognition, predicting crime based on social media accounts or individual risk assessment. Individuals arrested in Broward County, Florida, were analyzed to the algorithm using the same parameters as in the ProPublica study to determine how many would face charges in the coming years. And as discussed earlier only 20 per cent of those projected actually went on to commit crimes (Angwin, et al., 2016).

According to the paper (Angwin, et al., 2016), two individuals, Prater and Borden, were compared with the algorithm's predictions of their future actions and behaviors. Prater (a white male) was projected to have a low probability of committing crime in the future, despite the fact that he had previously committed armed robberies and stealing on several occasions (Angwin, et al., 2016). By evaluating Borden's (a Black female) past criminal behaviour, the algorithm in the study projected that she would have a high risk of committing crime in the future where her past crime was she took but also returned the kids bike after police arrived (Angwin, et al., 2016). The findings are shocking: after two years, Borden was found not guilty of any crime in the future, but Prater was condemned to eight years in jail for grand theft (Angwin, et al., 2016). The algorithm was incorrect in its predictions. No matter how awful this sounds, it is true in certain circumstances, which is why machine learning fairness should be considered in compliance with the legislation, as is the case with artificial intelligence.

In the paper (Makhlouf, et al., 2021), the authors discussed how machine learning algorithms are biased and are based on historical data. To find the most appropriate way of noting fairness in each particular real world example the authors followed the following steps:

- To begin, (Makhlouf, et al., 2021) defined the collection of characteristics associated with fairness in relation to real-world instances.
- Second, the authors (Makhlouf, et al., 2021) examined how different notions of fairness behaved.
- Finally, combining steps one and two the authors were able to identify the most suitable fairness notions for a variety of real-world examples.

## 2.3 College admissions

Admission to college is a crucial component of a student's career. This period establishes the student's future. Students must pass examinations in order to apply to colleges. Previously, the SAT (standardised assessment exam) was the only and most significant test used by institutions in the United States to evaluate whether a student has a required level of proficiency. However, the number of tests and examinations has grown, and they differ per institution. Why are universities concerned about these tests? The solution is straightforward: similar to companies, they want to be ranked

among the best institutions, which is very reasonable. This is not the situation at all institutions. Unknowingly, their algorithms are built on data impacting minorities, either because of the place in which they reside or because their historical data indicates that their groups have had a scarcity of intellectuals (Makhlouf, et al., 2021). This may cause long-term economic injustice and slowly affect the higher education's role in society and cause imbalance (Makhlouf, et al., 2021). College admissions should be explicitly based on test results, extracurricular activities involvement, interviews, and aptitude testing. To do this, educational institutions should evaluate the dataset upon which machine learning algorithms are built. This way, the student may feel secure that they are being examined fairly and without jeopardising their mental health.

## 2.4 Loan approvals

For many years, banks and money lending businesses have chosen machine learning algorithms. Banks desire the maximum level of trust when lending money, and they place a great deal of confidence in machine learning algorithms. When applying for loans, you must complete an application that includes information about your income, age, gender, ethnicity, marital status, and credit score. While banks strive to provide the best possible client experience, they fall short in some areas. An example of a machine learning algorithm applied to a bank system is in the paper by (Luo, et al., 2020) where the authors apply K-means algorithms to the online banking system.

Numerous researchers are examining ways to enhance the banking system by focusing on the loan prediction problem. On the Kaggle platform, researchers could access a huge number of publicly accessible datasets to evaluate their machine learning models. Researchers often utilise these datasets to construct a model that is efficient and provides a slightly better forecast for banks about whether to approve or reject an individual's loan application.

A bank's profit or loss is entirely based on loans. A group of researchers (Sheikh, et al., 2020) that used machine learning algorithms to the Loan approval prediction issue were able to employ logistic regression to ensure that their model accurately predicted which individuals are likely to fail on a loan. Their model was built using all available customer data, including age, gender, marital status, credit history, and employment. The researchers (Sheikh, et al., 2020) achieved an accuracy of 81% on the dataset using their trained model. However, their model was introduced by demographic data such as an individual's gender. Although their (Sheikh, et al., 2020) model functioned admirably, the researchers did not investigate the possible impact of fairness in this scenario. Bear in mind that although machine learning algorithms learn from data and provide a result and move on, clients will always remember the outcome. As a result, it is critical for banks to consider incorporating fairness into their loan prediction process. As a result, this project is developed around this topic and will seek to address the issue of incorporating fairness in machine learning models for loan prediction.

# 3. Analysis

## 3.1 Methodology

The algorithm will be entirely focused on achieving high accuracy while minimising prediction disparity amongst demographic groups. The algorithm would weigh the cost of fairness against the expense of accuracy. Increased accuracy implies decreased fairness, but in this project, the algorithm would attempt to achieve a balance between the two, delivering an accuracy of more than 80% with a disparity of less than 10%. Additionally, hyperparameter optimization is performed to ensure high accuracy.

The dataset used in this prediction problem is the adult census income dataset from the UCI repository (Kohavi & Becker, 1994). A clean form of this dataset containing the same values is in the scikit-learn library and it was easy to perform the algorithms and pre-processing tasks hence the adult dataset from the scikit library is used in the final product which is the same as the UCI adult income dataset.

In this dataset clean records were extracted having age more than 16 years. The dataset determines whether an individual makes more than 50,000 a year or less than 50k. The attributes in this dataset are age, workclass, fnlwgt, education, marital-status, occupation, relationship, race, sex, capital loss, capital gain, hours per week and native country. Several researchers have used this dataset for classification purposes and many have found that decision tree and logistic regression have performed well with this dataset. Hence in this project, both algorithms will be tested in the testing phase and the one which provides better results would be used for the final product. Banks need good machine learning algorithms that would predict the likelihood of the individual being capable of paying the loan. Thus, consider a hypothesis scenario for this project: if an individual's income is more than \$50,000 or 50K, the bank will approve the loan application; if the individual's income is less than \$50,000, the bank will reject the loan application. The protected attributes in this project would be a person's sex and ethnicity. Now, keeping protected attributes does not necessarily guarantee that the system is fair. Bias in the system may persist for a variety of additional reasons, such as unintentional correlations between gender and the capital gain or loss columns contained in the dataset. As a result, only using protected attributes does not solve the fairness problem. Sample bias depends on the sample we choose to train our model, if the training data does not represent the entire population or does not have exposure to every feature then it is said to be sample biased. Label bias comes into the system due to human errors. Humans collect data but unknowingly have some bias in their dataset because the data is labelled incorrectly. The data that is used may not reflect the emergence of societal values.

Many researchers have come up with solutions based on how to interpret and implement fairness notions. An example of group researchers who worked on similar problem is (Chakraborti, et al., 2020). In their paper (Chakraborti, et al., 2020) they produce mathematical equations to address contrastive fairness in decision making algorithms. The researchers state that to consider if the decision making process is fair, the decision for one individual the decision taken should have a high probability score than the alternative decision and vice versa (Chakraborti, et al., 2020). Also, while making fair system the person should make sure that even if the sensitive features are protected the prediction for one individual should have a higher value than the other. The researcher states that if these two points are satisfied that one can say that the system is fair.

Both the decision tree technique and the random forest classifiers produced good accuracy for the models under evaluation in this project. The random forest classifier, on the other hand, performed the best, with an accuracy of more than 85 percent. By functioning as a layer on top of the random forest model, the threshold optimizer contributed to the fairness component of the issue as well. Despite the fact that it lowered the accuracy by 6%, it also reduced the disparity and the selection rate difference between the male and female groups.

As discussed in the project proposal this project would follow five stages of SEMMA:

1. **Sample:** Obtaining the Dataset from Kaggle that has the appropriate amount of data that is adequate to train the model and is relevant to the research project.
2. **Explore:** In this stage, the dependency between the data components will be compared, and the data gaps will be explored with the use of visualisation. Data pre-processing steps would be taken in this stage.

3. **Modify:** This stage involves cleaning up the data, which might take a significant length of time due to the possibility that some information is missing.
4. **Model:** training the model with 2 different supervised machine learning models and comparing their result to choose the one that performs the best.
5. **Assess:** The last stage will be to evaluate the product and the entire project with industry standards and compare if the project was able to meet the baselines.

## 3.2 Impact of Legal, Ethical, Social and Professional Issues

Machine learning algorithms do make the task of humans easier but do these algorithms have any ethical or legal impact, To answer this research was carried out by a group of people in (Jameel, et al., 2020). The authors emphasise the serious importance of using high-quality data while training and testing AI algorithms. Like humans have government, police and laws to keep an on our behaviour and reasoning, Machine learning algorithms need to be set up with some fundamental ethics in order to avoid any further consequence. Artificial intelligence has been widely used in practically every sector in recent years. Take a look around you; we now have Tesla's automated automobiles. Although artificial intelligence has just recently been introduced in this industry, this machine learning algorithm has the potential to affect people's ethics, since automated automobiles would indicate that individuals with driving abilities would lose their jobs (Jameel, et al., 2020). In certain fields, where individuals are paid to translate or act as a tour guide for a specific location, such individuals will now lose their jobs as a result of chatbots or bots that translate within seconds using machine learning algorithms. While an automated world sounds appealing, we need to establish certain guidelines for ethics before using these machine learning algorithms. Machine learning algorithm can raise issues regarding the freedom of users as AI could control a persons thoughts or decisions.

Machine learning algorithms that are being used in the social media platforms need to support democracy. The algorithms should be transparent and not be able to create fake news or leak any private data that may cause a risk to a specific country. Algorithms learn from data and the data could be biased. Bias can creep in from data, humans or in algorithm design. Data can be inherently biased without any intention. Biased data can be carried unknowingly from one human to another human. But the machine carries out these bias judgement and learns from this data to later create a more biased prediction. In this project, one major part of Legal, ethical, social and professional issues is being tackled which is by introducing fairness in machine learning algorithms to reduce bias. As noted by (Angwin, et al., 2016) algorithms were often skewed and targeted certain populations. These groups were significantly disadvantaged, and the predictions were often inaccurate, perhaps causing distress to those harmed by unfair algorithms.

To ensure fairness has been applied in this project fairness metric like demographic parity is used. Demographic parity is a fairness metric which has been applied to the algorithm which ensures that both genders have an equal probability of an outcome. Both Male and females would be kept at the same equal level for testing and then the machine learning algorithm works on the other features to present the prediction.

Another important factors that machine learning models need to follow are during the data mining process the researchers should make sure that their data is not leaked from one user to another.

## 4. Requirement Specification

The dataset is the most critical component of the project. The kind of dataset used to train machine learning models is crucial since the algorithms learn from the data and provide predictions. A dataset from the scikit-learn library was selected for this research. The data comes from the adult census income dataset, which includes information on individuals such as their age, gender, marital status, employment, relationship, and capital loss and gain. This dataset forecasts an individual's income based on these factors, with the target column containing two outputs being >50K or ≤50K. As a result, this is a classification problem, and classification methods will be used in this project. The dataset also ensured to follow the legal, social and ethical implications because the authors made sure not to include any individuals name.

After identifying a dataset, the following step is to determine how fairness might apply to our problem. The initial step is to identify all of the dataset's sensitive features (Li, et al., 2021). The second stage determines the sensitive feature's fairness in relation to the result (Li, et al., 2021). The sensitive features would be stored as protected attributes, but this does not resolve the issue of unfairness. Sensitive attributes have distinct consequences on the machine learning model and hence it influences the result. Thus, the model is said to be unfair but the degree of unfairness is what determines how unfair it is. The researchers from (Li, et al., 2021) explained this theory by considering two attributes as A and B, then they calculated the probability of the outcome which was  $P(A \cap B) = 1$  which means that no matter what we choose A or B the probability of the outcome should be same for both A and B. This is also called as equalised odds.

Decisions and predictions are made using machine learning algorithms. It is critical to choose the appropriate machine learning algorithm for the project.

The decision tree method was selected for this project initially based on its accuracy and efficiency when used in combination with the threshold optimizer. But then later opted to use the random forest classifier with the threshold optimizer as the final model for this project. Threshold optimizer is a layer that is added on top of the machine learning algorithms to introduce fairness. Threshold optimizer has fairness metric called as demographic parity which has equal selection rates. In this project, demographic parity group is the gender group hence it equalises odds for both males and females.

In the machine learning community, many researchers are working on minimising the tradeoff between fairness and accuracy. A group of researchers in (Kenfack, et al., 2021) tried to examine the effects of fairness on machine learning models. Loss in accuracy can also cause serious troubles especially in the medical field hence it is important to find a fine balance between fairness and accuracy (Kenfack, et al., 2021). The researchers (Kenfack, et al., 2021) used 3 different fairness metrics which would be applied in this project:

- Demographic Parity/ Statistical Parity: Here the model should not be dependent on the protected attributes and when this is successful, the classifier would not be unfair towards any individual from any unprotected group. The classifier would predict positive outcome even if the feature member have a slightly negative stats ( for example they might not be able to pay back the loan in given time) (Kenfack, et al., 2021).
- Equalised Odds: In this fairness metric, the false positive rate and the false negative rate should be the same across all the groups (Kenfack, et al., 2021). Hence the researcher states

that if the probability of false positive rate and the probability of false negative rate is equal we may get a somewhat fair model. In this project, confusion matrix would be plotted to show the true positive and false positive values for both the genders (male and female).

- Equal opportunity: equal opportunity sounds similar to equalised odds because it also requires the false negative rates to be equal across all groups (Kenfack, et al., 2021). But in this case that's the only true value it requires to be considered as equal opportunities.

## 5. Design

The Design of the project had many ups and downs. In this project, many algorithms were tested and datasets were changed multiple times the following stages will explain how each stage help to build the project overall:

### 5.1 Initial Design Stage 1

This was the most fundamental stage of the whole project. The topic had just been determined, and the next step was to locate an appropriate dataset that had at least one sensitive feature column. At this point, the primary emphasis was on interpreting how fairness in machine learning could be portrayed using the dataset as a basis for analysis. Therefore, the first initial idea was to work with the Kaggle HR employee dataset, which included information about employees and projected whether or not the employee would quit the organisation based on the features of the employees. The dataset contained columns like job involvement, salary increase, gender, monthly rate which would have been easy to interpret and analyse.

Now, this dataset nearly met all of the criteria for being chosen for the project, but the number of entries in the dataset was insufficient to be considered. If this data had been used to train a machine learning model, it would not have been feasible to explain where and how bias occurs in machine learning. It is also true that this little data may project high accuracy practically every time, which is not true in the real world, which is why this dataset and the concept were discarded. The main point was to prove fairness is required in real world example. Which real-world example should be used was still a question?

### 5.2 Initial Design Stage 2

An important field where machine learning is highly used is the medical field. As everyone is aware that cancer is a life taking disease and has affected millions of people around the world. Researchers are using machine learning to examine and analyse different cancer problems and are coming up with better ideas to solve them. One of the most popular research is done on the breast cancer dataset (Fatima, et al., 2020). Breast cancer is the one of the most common type of cancer found in women. Therefore, it is one of the most important issues that could be solved using machine learning techniques. Machine learning could help solve this problem by predicting in advance if the individual would have cancerous cells. Machine learning is often used in predictive modelling and pattern recognition applications.

International medical standards have increased and machine learning techniques are known to increase everyday. Early detection and treatment of breast cancer may significantly reduce the severity of the disease's impact on the patients quality of life.

Hence the idea of solving cancer dataset problem with machine learning would be excellent and useful for the real world problem. But in this case, removing the protected attributes from the models reach, does not make sense and wrong predictions could affect womens lives. In any similar medical dataset if we decide to assign gender as the protected attribute and then do not use it to obtain our algorithms predictions, would cause a serious problem, as in this case the protected attributes are important to classify the problems solution. So the idea of using medical dataset had to be dropped too.

### 5.3 Final design Stage 1

The adult income dataset from the UCI machine learning repository was best suited for this project because it had the features that were needed to classify the problem. The same dataset was available on the scikit-learn library and was easy to load and work on in the loading phase. Just two lines of code was required to call the dataset in the python file which was

```
from sklearn.datasets import fetch_openml  
  
data = fetch_openml(data_id=1590, as_frame=True)
```

The scikit learn library has special inbuilt command called as `fetch_openml` which helps the use to load any dataset that is stored in this library by just entering the data id. This was really beneficial since it eliminated the need for an external excel file and enabled testing personnel to simply access the information and experiment with their own machine learning models.

The initial idea was to use the fairlearn dashboard and present the disparity in the predictions. However in the testing phase it was found that dashboards were removed from the fairlearn package and the entire plan to output the results had to be changed.

### 5.4 Final design

The dataset (Kohavi & Becker, 1994) adhered to legal, social, and ethical standards since it was uploaded to a well known repository and was used by many researchers. In this project, the data originates from the adult census income dataset, which contains information on people such as their age, gender, marital status, job status, relationship status, and capital gain and loss. This dataset anticipates an individual's income using these variables, with the target column including two outputs: >50K and 50K. As a consequence, this is a classification challenge, and this project will use classification techniques.

Now after the dataset and everything was finalised, the next step was to create a product that would be useful for the real world problem. In this project a hypothetical case was created in which we have to imagine that a bank wants a application which would require the users to input their data and would predict if they are successful for the loan application or not. Now the Loan prediction would be based on Income of the individual if the income was more than 50K the application would result an success message output saying that the individual has an income of more than 50k. The trained threshold optimizer model was saved as 'Trained\_model.sav'

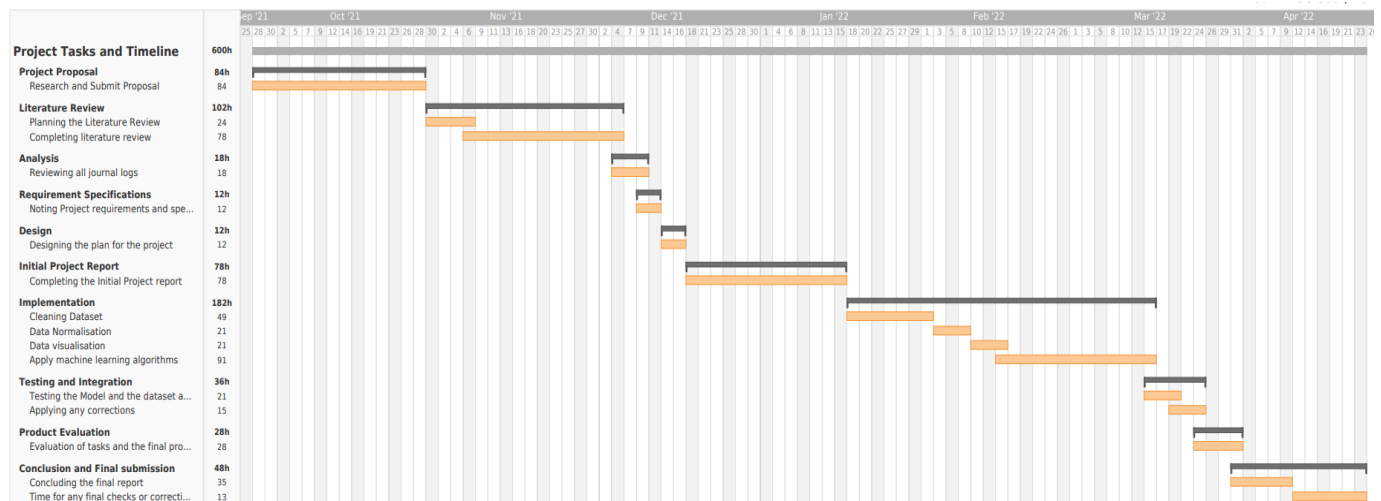
To design a user-friendly interface for this project, streamlit was used. Streamlit has a simple web interface that integrates efficiently with machine learning algorithms. Using streamlit, stakeholders or businesses can quickly test machine learning via a web interface without digging into complicated code files. Another similar online interface that could have been used is flask, however flask requires some understanding of HTML and CSS, which is not intuitive for a business. Streamlit is also compatible with Python, which simplify the developer's task by eliminating the need to purchase additional licences or download other software to execute the code. The trained data was kept in a sav file and



could be easily accessed using the streamlit python code. The next step is the demonstration of how time was allotted for every stage for this project.

## 5.5 Gantt chart time distribution for the project

The following chart illustrates the time allocation for the project, beginning with the research and delivery of the proposal and ending with the final submission report. The projected time required to complete this project is around 600 hours, with each distribution detailed below.



This gantt chart was prepared during the proposal submission for this project. However, a few modifications were required throughout the actual execution of the design and testing processes. The design process took longer than the 12 hours specified above. The design phase took around 20-30 hours, although this lost time could be regained. The dataset cleaning was scheduled to take approximately 40+ hours, but also because the dataset was pre-cleaned, and was from the scikit-learn library. and only required some minimal data pre-processing steps, a significant amount of time was saved and the project was back on track to be submitted before the deadline. The testing step took an additional four hours since the dashboard for Fairlearn that was chosen previously was no longer accessible in the scikit learn library and a few modifications were essential to demonstrate the plots in a latest trend. Finally, everything was completed on time, and there was still time to double-check any remaining code mistakes. Gantt charts are quite beneficial since they allow researchers to track how much time they spent on the project throughout implementation. Other courseworks and external factors also affect the timing of the project parts. For example, there was a time where priority to the other coursework had to be given due to its early deadline time.

Challenges faced by supervised machine learning algorithm are as follows:

1. Training supervised machine learning models make take a significant amount of time if compared to unsupervised machine learning.
2. In some datasets, bias could be incorporated due to human errors and which may make the supervised model to be biased.
3. In contrast to unsupervised machine learning models, supervised machine learning models are unable to cluster or categorise data on their own.



## 6. Implementation

Fairness can be applied during the pre-processing which is to change the data by keeping the sensitive features protected so that the machine learning models cannot learn from this specific attribute. Another way to introduce fairness in machine learning is by incorporating explicit fairness constraints in the machine learning algorithms, which basically means applying demographic parity or equalised odds to the algorithms. Third way to integrate fairness in machine learning is to incorporate it with the existing machine learning model such as applying a threshold optimizer layer on top of our model. All 3 fairness steps were taken into application for this project.

The dataset looked like the figure below:

Figure1:

Index	nu	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
0		Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States
1		Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States
2		Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States
3		Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States
4		Never-married	nan	Own-child	White	Female	0	0	30	United-States
5		Never-married	Other-service	Not-in-family	White	Male	0	0	30	United-States

The sensitive features in this dataset are the columns 'sex' and 'race'. Both these columns were kept as protected attributes to integrate fairness at the pre processing stage. Both the column race and sex were dropped from the dataset before splitting it into a testing and training dataset.

Machine learning models are often split datasets into test/train subsets. The sklearn package has a function that is utilised by importing the test train split. However, the target column had to be assigned to a variable first. As a result, the variable 'Y' had an array of 0's and 1's representing the income column data. The variable 'X' contained data from all columns except those that were protected. This was the first step in ensuring that fairness was introduced to the problem. However, having protected attributes alone does not resolve our issue.

Supervised machine learning : supervised machine learning, is a subfield of machine learning and artificial intelligence that includes both deep learning and reinforcement learning (IBM Cloud Education, 2020). Supervised machine learning uses labelled datasets to predict the classification problems. Supervised learning can be divided into two parts which is classification and regression. In this research 2 of the supervised machine learning models were used namely decision tree and random forest classifiers.

Decision tree algorithms: The structure of a decision tree is similar to that of a tree. Each internal node in the decision tree represents an attribute test, while the branches reflect the test's conclusion and the leaf represents the class name (Geek for Geeks, 2021).

Random Forest classifier: It is another form of supervised machine learning used for classification problems. The random forest classification model is a collection of decision trees which are combined together in order to reduce the variance and provide a high accuracy of predictions (IBM Cloud Education, 2020).

Hence due to higher accuracy of random forest classifier this model was opted for the final product. In the random forest model some hyperparameter tuning was done in order to get maximum possible

accuracy. The number of estimators were changed from 10 to 99 which increased the accuracy from 81 percent to 85 percent for the random forest classification model.

## 6.1 Proposed Product

The product idea is simple. There is a bank application made for the bank so they can use machine learning algorithms to check if their client is eligible for loan or not. In the machine learning algorithms both decision tree and random forest were tested and the one with highest accuracy was the random forest classification model. The final model was a combination of randomforest model and the threshold optimizer with the sensitive features being 'sex' of an individual. The accuracy of the final model suggests that the model performed good. There will be a compromise of accuracy in the final model but the disparity in the predictions between male and female candidates has been reduced to less than 1%.

## 6.2 Data collection and Preprocessing

As discussed earlier the dataset is from the UCI machine learning repository by (Kohavi & Becker, 1994). For the data preprocessing, data mining technique were used to turn the data into a more useable form for this specific problem. Feature scaling has been used to convert the features that contained string to int data type. The model is not able to read the column names or the features that are written in characters hence feature scaling was required. Sklearn library has a inbuilt function called as Label encoder. The categorical data is converted into a numerical format without losing any information. In this project the data is split into a test train split. Additionally, the sensitive features are also passed into the test train split function to create a separate dataframe called SF\_test and SF\_train which would contain the training and testing data set for the sensitive feature which is 'sex' in this scenario.

```
X_train, X_test, Y_train, Y_test, SF_train, SF_test = train_test_split(X_test, Y, SF, test_size = 0.20,
random_state = 0, stratify=Y)
```

## 6.3 Machine learning Models

In this project, supervised machine learning models like decision tree and random forest classifiers are employed. Decision tree is a popular tool used by many researchers for classification and prediction. The structure of a decision tree is similar to that of a tree. Each internal node in the decision tree represents an attribute test, while the branches reflect the test's conclusion and the leaf represents the class name (Geek for Geeks, 2021). Classification problems may be solved using decision trees without needing complex calculations. They indicate clearly which features are important in predicting the classification problem. Decision trees do not perform well with a limited number of training data sets, however the training data in this project, called X train, has almost 300,000 values. The primary issue with the decision tree in this project was its accuracy, which was less than 80% even after some hyperparameter tuning. Therefore, random forest classifiers were examined again and demonstrated to perform very well with the training and testing data.

# 7. Testing and Integration

The testing and integration stage took around 35 hours. The following stages would explain the number stages it took to get a final product and the errors at each stage.

## 7.1 Initial integration

The dataset was successfully loaded into the python file and by using commands like `.head()`, the first five rows of the dataset was printed. This was done to check if the dataset has the correct number of required feature columns. Then the dataset was checked for any null values by using the command `.info()`. This command tell us the data type of each column/feature and also if there are any null values.

Figure 2: Data columns and its data types

#	Column	Non-Null Count	Dtype
0	age	48842 non-null	float64
1	workclass	46043 non-null	category
2	fnlwgt	48842 non-null	float64
3	education	48842 non-null	category
4	education-num	48842 non-null	float64
5	marital-status	48842 non-null	category
6	occupation	46033 non-null	category
7	relationship	48842 non-null	category
8	race	48842 non-null	category
9	sex	48842 non-null	category
10	capital-gain	48842 non-null	float64
11	capital-loss	48842 non-null	float64
12	hours-per-week	48842 non-null	float64
13	native-country	47985 non-null	category

This was the most impotant step as here sensitive features identified. The columns age,fnlwgt, education-num,capital gain,capital loss and hours per week were float values hence there was no need of feature scaling for these columns. The categorical values such as workclass, education,marirtal status, occupation,relationship,native country was required to be encoded with the help of Label encoder because the algorithm would print a error stating float values not recognised.

Then was the most important phase to check the accuracy of the decision tree algorithm

```
Training accuracy for Decision Tree= 0.9998976275177233
Testing accuracy for Decision Tree= 0.8178933360630566
Training accuracy for Random Forest Classifier= 0.9998464412765848
Testing accuracy for Random Forest Classifier= 0.8581226328180981
```

Figure 3: Accuracy of models

The training accuracy of the decision tree algorithm and random forest classfier was somewhat similar to each other but the testing accuracy is the one that is important for the project. The testing accuracy for the decision tree algorithm is good (81%) but after applying the threshold optimizer layer to the decision tree the accuracy of our final model decreased to 75%. Hence it was neccessaary to go back to the threshold optimizer and to test it this time by applying it to the random forest classifier and the results were excellent as accuracy of the final model was around 80% which means the performance of the model increased and the tradeoff difference could be reduced between the accuracy and the fairness of the model.

## 7.2 Testing and Interpreting the results

It was time to put both algorithms through thier tests before and after the fairness measure was applied. It's was facilitated by presenting plots for both models in relation to their sensitive features. The first thought was to make use of the fairlearn dashboard, which would make it simple to evaluate the results and track the change in disparity between the predictions over time. Fairlearn community had discontinued providing dashboard services and had taken the dashboard down,

which made this impossible. Fairlearn metrics plotting methods were the only means to see the difference between groups. Fortunately, the code to plot the graph was provided in their website user manual, but it needed to be altered in order to produce a result that was suitable for this project.

Figure 4: Plotting technique using Metric Frames

```
from sklearn import metrics as skm
metric_frame = MetricFrame(metrics={"accuracy": skm.accuracy_score,
                                   "selection_rate": selection_rate,
                                   'true positive rate': true_positive_rate,
                                   },
                           sensitive_features=SF_test,
                           y_true=Y_test,
                           y_pred=randomforest_model.predict(X_test))

print(metric_frame.overall)
metric_frame.by_group.plot.bar(
    subplots=True, layout=[2, 2], legend=False, figsize=[12, 8],
    title='Accuracy and Selection rate difference between male and female')
```

This code helped to identify the difference in accuracy, selection rate and true positive rate. The selection rate is the difference between the options that the model choose. For example in this case if there is a option to choose between a male and female for making predictions. The difference between how often did the model choose males vs how often did it choose females is the selection rate. After applying the same code to the final model with the threshold optimizer as well a plot was created.

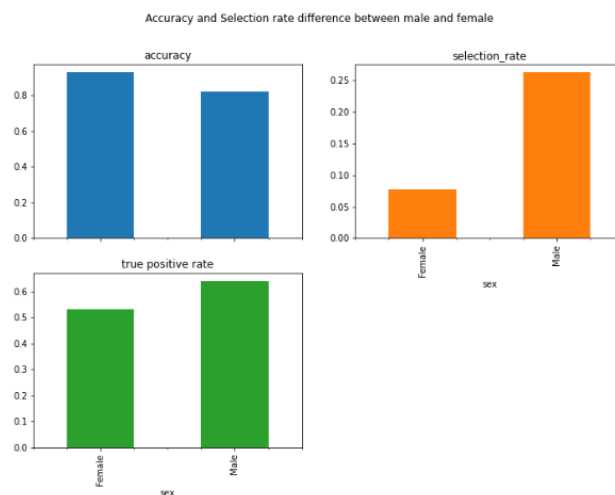


Figure 5: showing the accuracy and selection rate difference between male and female for the random forest model with minimal or no fairness metrics applied

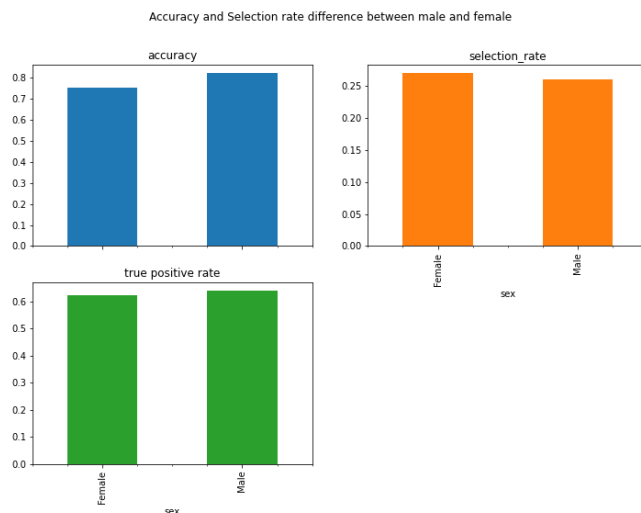


Figure 6: showing the accuracy and selection rate difference between male and female for the threshold optimizer model applied to the random forest model.

As the plots clearly indicate that the model in figure(5) which was the random forest model had a huge difference between the selection rates for male and female. However this was improved in the threshold optimizer model (figure(6)) after the selection rate difference was minimal. The accuracy of the random forest model for selection of male and females differed too. The true positive rate should be same for the males and females and if its true it's a clear indication that the model is fair because equalised opportunity requires true positive rate to be the same for both the features. The main topic for this entire project was to prove how one can find a balance between tradeoff and accuracy and integrate fairness in the machine learning models.

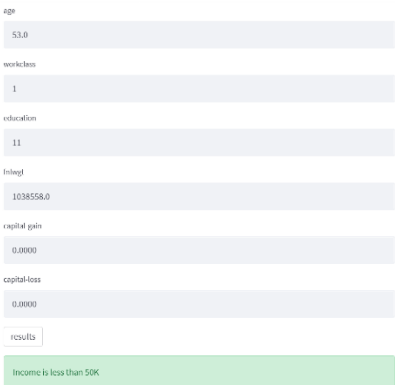

### 7.3 Final testing

In this stage the code was running without any compilation errors and now was the time to check if the streamlit application is running and if the predictions could be made using the threshold optimizer model.

The screenshot shows a Streamlit web interface with the following inputs: age (53.0), workclass (1), education (11), fnlwgt (1038558.0), capital-gain (0.0000), and capital-loss (0.0000). A 'results' button is located below these inputs. The output, displayed in a green box, is 'Income is less than 50K'.

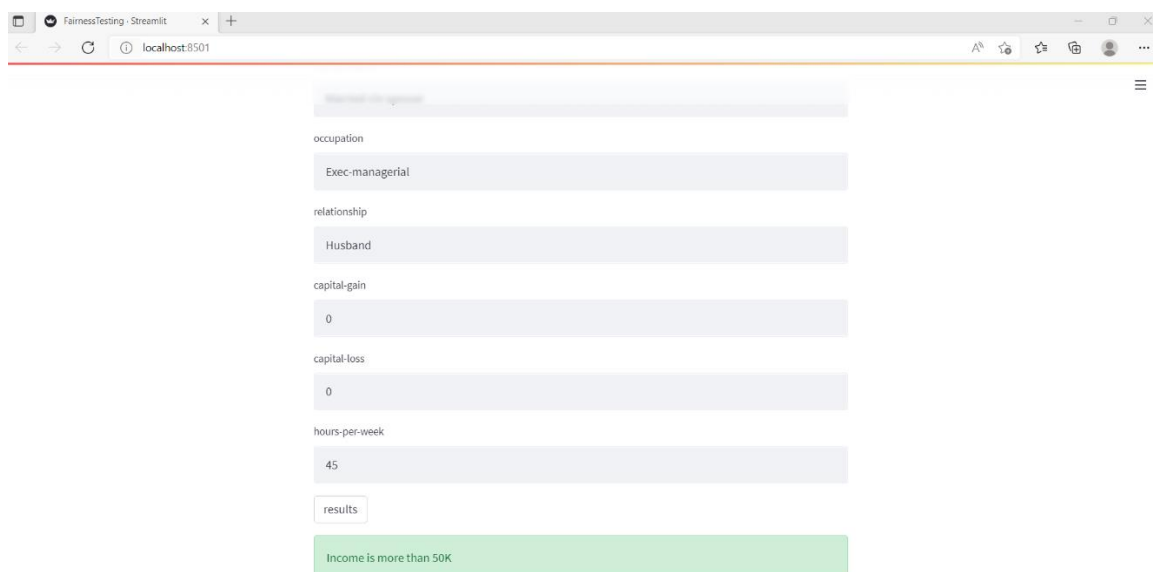
This was the initial streamlit page. There were many problems in this due to which this design style had to be dropped. The first major problem was the user would not be able to understand how and what values to enter in workclass education as these values were different from the initial columns in the dataset. The second major problem faced with this initial design was that the results were incorrect and every time it resulted in the same output.

The problem was identified which was due to the iteration in the for loop.

The code changed from this	To this	Original value in the dataset for corresponding age
<pre>if ((predictions == 0):     return "Income is less than 50K"     break else:     return "Income is more than 50K"     break</pre>	<pre>for i in predictions:     if ((predictions[i+1]) == 0):         return "Income is less than 50K"         break     else:         return "Income is more than 50K"         break</pre>	Income more than 50K
		Income more than 50K

As you can see above the prediction were correct and matched the original dataset values.

More features were added in the end to the final design like hours per week



The streamlit application was able to take text box input which would allow the users to select the values from the original dataset and check them by entering the values in the application. After pressing the results button the application runs the threshold optimizer model and predicts if the application would have an income more than or less than 50K based on the training data.

The important features for this application to predict were and they were added to the streamlit application as followed :

```
age = st.text_input('age')
workclass = st.text_input('workclass')
fnlwgt = st.text_input('fnlwgt')
education = st.text_input('education')
marital_status = st.text_input('marital-status')
occupation = st.text_input('occupation')
relationship = st.text_input('relationship')
capital_gain = st.text_input('capital-gain')
capital_loss = st.text_input('capital-loss')
hours_per_week = st.text_input('hours-per-week')
```

An example of how the application works is shown below:

Lets consider that you want to check if the applicant is eligible for loan or not. You need to enter the applicants age,workclass, fnlwgt, education, marital status, occupation, relationship, hours per week, capital gain and capital loss.

So if the applicant has the values as listed below

```
age = 42
workclass = Private
fnlwgt = 159449
education = Bachelors
marital_status = Married-civ-spouse
occupation = Exec-managerial
relationship = Husband
capital_gain =5178
capital_loss = 0
hours_per_week = 40
```

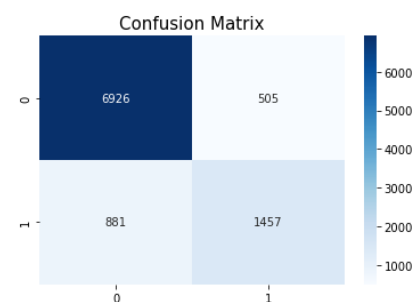
The value predicted in the dataset is that this individual will have a income more than 50K and now we check if our application which was built with a fairness addition can predict the same or not.

occupation	Exec-managerial
relationship	Husband
capital-gain	5178
capital-loss	0
hours-per-week	40
results	Income is more than 50K

As you can see we were able to get the correct predictions without involving the protected attributes. Now sometimes this application can predict a false result this is because the accuracy of the model is 80% if the accuracy was more than 95 percent we would have always got correct predictions. An example of similar input data but where the model predicted incorrectly is shown below.

Married-civ-spouse	
occupation	Exec-managerial
relationship	Husband
capital-gain	5178
capital-loss	0
hours-per-week	40
results	Income is less than 50K

Confusion matrix for the random forest classifier



The confusion matrix tells us about the true positive, true negative, false positive and false negative values.



## 8. Product evaluations

Making correct decisions is important since they may have a direct impact on other people's life, whether it is approving a loan application for a customer or accepting a student for college admission. Incorrect judgments might jeopardise a business's profitability and an employee's ability to pursue greater goals. Wrong judgments may also have an effect on an individual's motivation for a certain endeavour. While the usage of artificial intelligence and machine learning has increased significantly, industry should ensure that machine learning algorithms follow certain guidelines to avoid causing social, legal, ethical, or professional problems. Datasets may be skewed, and it is the researchers' responsibility to ensure that they attempt to eliminate bias from the data by applying fairness approaches to it wherever feasible. In certain instances, removing protected attributes is not feasible such as medical data that could impact the person medical condition. In order to integrate fairness into machine learning algorithms, a layer called a threshold optimizer is built on top of them. The 'demographic parity' is a metric, which is used by the threshold optimizer to ensure fairness, has equal selection rates for all candidates.

The product was evaluated on if it checked out the following functional requirements

Requirement	Requirement Met Yes/No
Creating a fully functional application to help the real world problems	Yes
Machine learning algorithms used to provide efficient solutions	Yes
Fairness metrics were applied to the model/algorithms	Yes
Comparing the accuracy for the models before and after fairness metrics were applied	Yes
Got a final product with a working web application	Yes
Application had quick response time	Yes
Application was user friendly	Yes
Having a feedback system to get feedback from users	No
Integration of cloud based system	No

The priority of the requirements and the time remaining decided which requirements were supposed to be completed first. Some functions requirements were completed without any errors and some function requirements provided errors during the testing phase. The streamlit webpage has working buttons and text boxes but the application could have additional features like select boxes where users could select from a list of features for data input instead of manually entering all the data by themselves.

## 9. Further Development Options

Although the model was able to get a accuracy of more than 80% and performed efficiently, research could be held in the future rearding how to improve the performance even more and work around neural networks. If there was more time for this project research an indept explanation of fairness notions could have been examined. By including more accurate real world data this project idea could be combined with a real world company and research could be done further in this area.

## 10. Conclusion

Fairness in machine learning is a key problem that many researchers should consider while developing their machine learning models. It may contribute to the minimization of model bias and contribute to the development of a more fair system for the industry. Yes, there is a trade-off between the model's accuracy and its fairness. However, the fine balance is inherent in the situation and could be resolved by the use of appropriate fairness metrics. In this study, the random forest classifier achieved an accuracy of 85.7 per cent without incorporating any fairness to the problem, and 79.8 percent with the fairness measures applied, which is a positive indicator. It is now up to banks to decide whether to continue using the same high-precision machine learning model or to opt for a more equitable machine learning model with a variation in accuracy of less than 6%. Additional research may be undertaken in the future to compare fairness to neural networks and to determine ways to improve accuracy even further. Both Decision tree algorithm and Random forest classifiers provided a high accuracy for the models. But the random forest classifier performed the best having an accuracy of more than 85%. Threshold optimizer helped the fairness aspect of the problem by acting as a layer on top of the random forest model. It did bring down the accuracy by 6 % but it also reduced the disparity and the selection rate difference between the male and female group which was proven by the plots.

## References

- Angwin, J., Larson, J., Mattu, S. & Kirchner, L., 2016. *ProPublica*. [Online]  
Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>  
[Accessed 20 01 2021].
- Barocas, S., Hardt, M. & Narayanan, A., 2019. *Fairness and Machine Learning*. s.l.:fairmlbook.org.
- Bird, S. et al., 2019. *Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned*. San Francisco, USA, Association for Computing Machinery. <https://doi.org/10.1145/3308560.3320086>.
- Chakraborti, T., Patra, A. & Noble, J. A., 2020. Contrastive Fairness in Machine Learning. *IEEE Letters of the Computer Society*, 3(2), pp. 38-41.
- Fatima, N., Liu, L., Hong, S. & Ahmed, H., 2020. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, 8(doi=10.1109/ACCESS.2020.3016715), pp. 150360-150376.

- Geek for Geeks, 2021. *Decision Tree*. [Online]  
Available at: <https://www.geeksforgeeks.org/decision-tree/>  
[Accessed 2022].
- Harris, C. G., 2020. *Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making*. New York (USA), Association for Computing Machinery, <https://doi.org/10.1145/3366424.3383562>.
- IBM Cloud Education, 2020. *Supervised Learning*. [Online]  
Available at: <https://www.ibm.com/cloud/learn/supervised-learning>  
[Accessed 2022].
- Jameel, T., Ali, R. & Toheed, I., 2020. *Ethics of Artificial Intelligence: Research Challenges and Potential Solution*, 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). s.l., doi=10.1109/iCoMET48670.2020.9073911.
- Kenfack, P. J. et al., 2021. *Impact of Model Ensemble On the Fairness of Classifiers in Machine Learning*. doi=10.1109/ICAPAI49758.2021.9462068 ed. s.l.:pages 1-6.
- Kohavi, R. & Becker, B., 1994. *UCI Machine learning repository*. [Online]  
Available at: <https://archive.ics.uci.edu/ml/datasets/adult>  
[Accessed 20 01 2022].
- Li, Y., Huang, H., Guo, X. & Yuan, Y., 2021. An Empirical Study on Group Fairness Metrics of Judicial Data. *IEEE Access*, 9(doi=10.1109/ACCESS.2021.3122443), pp. 149043-149049.
- Luo, G., Li, W. & Peng, Y., 2020. Overview of Intelligent Online Banking System Based on HERCULES Architecture. *IEEE Access*, 8(doi=10.1109/ACCESS.2020.2997079), pp. 107685-107699.
- Makhlouf, K., Zhioua, S. & Palamidessi, C., 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 102642(ISSN 0306-4573).
- Marvin, G., Jackson, M. & Alam, M. G. R., 2021. *A Machine Learning Approach for Employee Retention Prediction*. s.l., IEEE Region 10 Symposium (TENSYP), doi={10.1109/TENSYP52854.2021.9550921}} .
- Nieto, Y. et al., 2019. Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions. *IEEE Access*, 7(doi={10.1109/ACCESS.2019.2919343}), pp. 75007-75017.
- Sheikh, M. A., Goel, A. K. & Kumar, T., 2020. *An Approach for Prediction of Loan Approval using Machine Learning Algorithm*, 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). doi=10.1109/ICESC48915.2020.9155614 ed. s.l.:IEEE.
- Wang, H. et al., 2019. *Predicting Employee Career Development based on Employee Personal Background and Education Status*. Seoul, Republic of Korea, Association for Computing Machinery, <https://doi.org/10.1145/3352411.3352451>.