UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

MAURÍCIO HOLLER GUNTZEL

Fairness in Machine Learning: An Empirical Experiment about Protected Features and their Implications

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Dante Augusto Couto Barone
Coadvisor: Mr. Eduardo Gabriel Cortes

Porto Alegre
May 2022

“Eu agora — que desfecho!,
Já nem penso mais em ti. . .
Mas será que nunca deixo De lembrar que te esqueci?”
— Mario Quintana

DEDICATION

This thesis is dedicated to my professor Dante Barone and Eduardo Cortes who helped me during the creation of the work; Professor Maria-Esther for helping and clarifying doubts during the development of this work; to my friends and family for encouraging me when I was having a hard time.

ABSTRACT

Increasingly, machine learning models perform high-stakes decisions in almost any domain. These models and the datasets - they are trained on– may be prone to exacerbating social disparities due to unmitigated fairness issues. For example, features representing different social groups are known as protected features– as stated by Equality Act of 2010; they correspond to one of these fairness issues. This work explores the impact of protected features on predictive models' outcomes and their performance and fairness. We propose a knowledge-driven pipeline for detecting protected features and mitigating their effect. Protected features are defined based on metadata and are removed during the training phase of the models. Nevertheless, these protected features are merged into the output of the models to preserve the original dataset information and enhance explainability. We empirically study four machine learning models (i.e., KNN, Decision Tree, Neural Network, and Naive Bayes) and datasets for fairness benchmarking (i.e., COMPAS, Adult Census Income, and Credit Card Default). The observed results suggest that the proposed pipeline preserves the models' performance and facilitate the extraction of information of the models' to use in fairness metrics.

Keywords: Pipeline. fairness. machine learning. positive outcome. group fairness. Fairness Though Unawareness.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

COMPAS	Correctional Offender Management Profile for Alternative Sanctions
ML	Machine Learning
KNN	K-Nearest Neighbors
NN	Neural Network
DT	Decision Tree

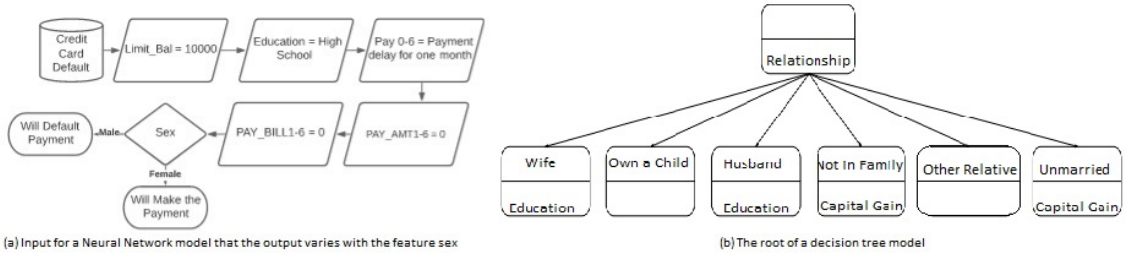
CONTENTS

1 INTRODUCTION

With the advance of big data, machine learning has played crucial roles in automated decision-making for various fields, including college admission, credit scoring, and criminal justice. As the practice areas of machine learning continue to increase, so does the concern regarding the potential bias that can be introduced or amplified by the model. For example, as reported by Reuters [1] the AI-based recruitment tool built by Amazon rated women poorly on the hiring scale based on historical data and tended to favor applicants that had more technical experience. The AI-based recruitment tool used a model that was trained observing resumes over ten years. Unfortunately, the model observed that men dominated technical positions during that time, which led women to be rated on the lower end of the scale. Another example is the algorithm COMPAS (Correctional Offender Management Profile for Alternative Sanctions), an artificial intelligence system used in several states in the United States. It was designed to predict whether a perpetrator is likely to commit another crime. According to a report published by ProPublica [2], this system was shown to have an implicit bias against African Americans, predicting twice as many false positives for African Americans as for whites. Based on the source [3], since this implicit bias was not detected before the system's implementation, many African Americans were unfairly and incorrectly predicted to re-offend. This case exemplifies the need for fair decision-making in algorithms that can be deployed in our daily lives.

1.0.1 Motivation

Figure 1.1: Motivating example. (a)Example of a case where the result of a neural network model varies with the gender. (b) Decision tree model trained with the Adult dataset shows that the most relevant feature of the trained dataset is the relationship of a person.



Source: The Author

[1] <https://ainowinstitute.org/discriminatingsystems.pdf>
[2] <https://www.propublica.org/work/how-we-analyzed-the-compas-recidivism-algorithm>
[3] <https://www.propublica.org/work/how-we-analyzed-the-compas-recidivism-algorithm>

Some datasets available for machine learning may contain features that characterize a person in a group (e.g., ethnicity, religion, or gender). Many of these groups can be sensitive to biases, and the feature(s) representing the distinct groups are considered protected. In fact, the United Kingdom has passed an act known as Equality Act [4] in 2010 providing a legal framework to define these features. Figure 1.1a shows a Neural network model decision input that makes the output change with the gender of the person, demonstrating that the model is biased towards genders. The case of Figure 1.1b, presents an example where the "relationship" is a protected feature with a high entropy value that exemplifies a problem using a protected feature with high importance during the classification process. Consequently, this model may present an undesirable bias considering the relationship status of the instances to be classified.

1.1 Problem Statement and Proposed Solution

The objective of this work is to propose a knowledge-driven pipeline that can facilitate the static evaluation of the fairness of the trained model with a dataset that contains sensitive attributes by removing these attributes from the training data and utilizing them to analyze the models. The method uses pre-processing and metadata, where the information of the sensitive feature is stored and used to verify the correlation of the sensitive attributes with other attributes stored. The verification of the method starts by defining fairness criteria for classification problems. Then, we explain that there are diverse types of discrimination, focusing on the inexplicable that is relevant to the approach, and within categorization algorithms, we utilized statistics metrics used in academia to define whether the trained model is fairer than the model trained without the knowledge-driven pipeline.

1.2 Contributions

Our experiments test different machine learning models for three datasets used for fairness benchmarks. We evaluate a *Fairness through Unawareness* method (Chen et al. 2019) by removing the protected features during the model's training and then removing the proxy variables of the protected attribute contained in the dataset. In sum, the contribu-

[4] <https://www.equalityhumanrights.com/en/equality-act>

tions of this work are the following:

1. A pipeline to identify and remove protected features to train the model while using metadata to track the changes.

2. An analysis of experiments using combinations of machine learning models and different datasets, the combinations of models trained with and without protected features and then without the proxy variables of the sensitive attributes are subjected to tests, and the results are then analyzed.

2 BACKGROUND

To the Cambridge Dictionary, fairness is the quality of treating people equally or in a right or reasonable way. This definition is too vague because there is no universal definition for fairness. After all, different preferences and outlooks in different cultures and times tend to change the concept of fairness. In this chapter, we explain some definitions of fairness and select one type of fairness definition to work in the empirical evaluation.

2.1 Types of discrimination

Discrimination is considered a source of unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally (Mehrabi et al. 2019). The work (Mehrabi et al. 2019) also explains types of discrimination and highlights that inexplicable discrimination can be divided in two types:

1. Direct Discrimination. Direct discrimination happens when protected attributes of individuals explicitly result in non-favorable outcomes toward them. (Chen et al. 2019) and (Mehrabi et al. 2019) used the "Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA)" as a basis to characterize the features as protected. The table 2.1 shows the attributes considered protected by the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA).

2. Indirect Discrimination. In Indirect discrimination, individuals appear to be treated based on seemingly neutral and non-protected attributes; however, protected groups or individuals still get treated unjustly due to implicit effects from their protected features.

The approach of this work focuses on direct discrimination taking into account the fact that the dataset used has at least one sensitive attribute enabling one to look directly in the model at the effects of the selected attributes.

Table 2.1: Protected features from Fair Housing and Equal Credit Opportunity Acts

Attributes	*FHA*	*ECOA*
Ethnic group	X	X
Color	X	X
National Origin	X	X
Religion	X	X
Sex	X	X
Familial Status	X	
Disability	X	
Exercised rights under CCPA		X
Marital Status		X
Recipient of public assistance		X
Age		X

Source: (Chen et al. 2019)

2.2 Fairness criteria for classification problems

In classification problems, an algorithm learns a function to predict a discrete characteristic Y, the target variable, from known characteristics X (2). Modeling A as a discrete random variable that encodes some features contained or implicitly encoded in X that we consider sensitive. Finally, we denote by R predictor of the classifier. There are three main criteria for assessing whether a given classifier is fair, that is, whether its predictions are not influenced by some of these sensitive variables:

1. **Independence**. The classification rate for each target class is equal for people belonging to different groups concerning sensitive characteristics.

2. **Separation** We say the random variables (R, A, Y) satisfy separation if the sensitive characteristics A are statistically independent of the prediction R given the target value Y. In some fields, separation (separation coefficient) in a confusion matrix is a measure of distance (at a given level of the probability score) between the predicted cumulative percent negative and predicted cumulative positive percent.

3. **Sufficiency**. The probability of actually being in each group is equal for two individuals with different sensitive characteristics, as they were expected to belong to the same group.

Given these three criteria for classification, we can analyze further, the work (Verma and Rubin 2018) collects and explains measures of fairness from different statics methods as an example would be based on the predicted outcome:

1. Group fairness. A classifier satisfies this definition if subjects in protected and unprotected groups have an equal probability of being assigned to the positive predicted class.

2. Conditional statistical parity. This definition extends the group fairness definition by permitting a set of legitimate attributes to affect the outcome. The definition is satisfied if subjects in both protected and unprotected groups have an equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L.

(Verma and Rubin 2018) also divide the fairness measures based in similarity

1. Causal discrimination. A classifier satisfies this definition if it produces the same classification for any two subjects with the exact same attributes X. An example would be if a male and female share the same attributes X, then the two need to share the same result of the classification.

2. Fairness Through Unawareness. An algorithm is fair as long as any protected attributes A is not explicitly used in the decision-making process.

3. Fairness through awareness. The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals.

2.3 Proxy Variables and Confounding Variables

From the Oxford reference (proxy variable): A variable used instead of the variable of interest when that variable of interest cannot be measured directly, as an example, per capita GDP can be used as a proxy for the standard of living. For statistics, a proxy variable is a variable that is not directly relevant but serves in place of unobserved or immeasurable variables. In order for a variable to be a good proxy, it must have a close correlation, not necessarily linear, with the variable of interest. This correlation might be either positive or negative.

(3) utilized postcodes and indicators of vulnerability from the census data to extract proxy variables for ethnic group. The gender was extracted from individual-level data on titles (e.x. Mr, Mrs, Ms, Miss). These facts highlight that confounding variables

are a problem that needs to be addressed by the model programmer. Confounding variables and proxy variables are related concepts: correlated predictor variables. But there's a difference between them:

1. Confounding variables affect your results in undesirable ways by not being included in the model. They are primarily a danger when you aren't aware of them during the analysis.

2. Proxy variables benefit your analysis. You know about and intentionally include them in the model to improve your results.

Related Work

The Case for Process Fairness in Learning: Feature Selection for Fair Decision-Making.(Grgic-Hlaca et al. 2016). The work consists of extracting human opinion on the features utilized in the model, opining in for each feature for three types of fairness: priory fairness where the user considers fair the atributte without knowing how it can be used. Accuracy fairness, where the user classifies the feature as fair if utilized to increase the classification accuracy and disparity fairness. 100 Amazon Mechanical Turk (AMT) workers have opined if it is just for each category for each feature. With the human judgments, it is then presented an accuracy trade-off with the process fairness. The work concluded that there are practical situations where a better outcome in fairness can be achieved with only a small cost to accuracy. This work works with three types of fairness compared to ours that focus on a priory fairness approach where the features considered are previously selected. The use of human judgment to classify each feature is inefficient for the large dissemination of the method, while we chose to select a criterion for the characterization of the features.

MultiFair: Multi-Group Fairness in Machine Learning(Kang et al. 2021). This study is inspired by two limitations found in the literature. It focuses on studying the problem of multi-group fairness, which aims to enforce statistical parity on multiple sensitive attributes simultaneously directly. The first inspiring limitation is that some works could only de-bias multiple distinct sensitive attributes (Bose and Hamilton 2019). They do not mitigate the groups of multiple sensitive features. The second is that the problems of optimization of other works are often subject to constraints of statistical parity (Feldman et al. 2015), (Kearns et al. 2018),(Zafar et al. 2017). The work proposes a generic end-to-end framework including a feature extractor, target predictor, sensitive

feature predictor, and density ratio estimator. Unlike ours, this method depends on an extractor to detect sensitive attributes selected in the pre-processing phase.

Bias Quantification for Protected Features in Pattern Classification Problems. (Koumeri and Nápoles 2021). This work proposes a bias quantification measure, called fuzzy-rough uncertainty, that acts as an individual fairness metric as it focuses on instance-wise inconsistencies. It quantifies the relevance of a protected feature in classification problems as a proxy for measuring fairness. This approach does not rely on any prediction model but on the distance functions utilized in the Fuzzy-rough regions to detect bias. The method's advantages are that no user intervention is needed to detect the discriminated subgroup. It does not depend on a machine learning model to compute its outcomes but on a solid mathematical foundation. While this method quantifies the importance of the feature to its accuracy and selects the features that will not create a significant disturbance inaccuracy for removing the atributter, unlike ours, it does not remove the protected attribute from the dataset for all inputs.

3 KNOWLEDGE-DRIVEN PIPELINE

In this work, we propose a pipeline able to exploit knowledge about the protected features in a dataset; it aims to facilitate the development of a fairer behavior in predictive models. A general overview of the pipeline architecture is shown in Fig 3.1. The general workflow of our proposed pipeline is divided into four phases. The first is the **Input** that consists of a dataset that contains metadata—this information is required in the subsequent phases. The second phase is the **Pre-processing**, where protected features are defined by choosing a criteria. The protected features can be chosen from different types of criteria an example would be from a data protection law viewpoint. An example of data protection law is the data protection law from Brazil[1] where sensitive data can only be used when the holder or his legal guardian consents and for specific situations that are in the constitution, the affected attributes are those who reveal racial or ethnic origin, religious or philosophical beliefs, political opinions, trade union membership, genetics, bio-metrics and a person's health or sex life. Another criteria that can be used to chose the protected features is a anti-discrimination viewpoint (Gellert et al. 2013) that, for example, in Brazil the affected attributes are race, color, ethnicity, religion and national origin.

With the definition of protected resources in the pre-processing step, with the help of metadata, two new datasets are created, the first dataset without the sensitive attributes and the second in addition to removing the sensitive elements from the proxy variables those characteristics contained in the original dataset are also removed from the created dataset found in the original dataset. The third step is **train and test**, where datasets are used for conventional machine learning, consisting of training and testing phases. In this phase, models go through train and validation methods such as cross-validation. As the validation method is used in a machine learning algorithm, the algorithm can be any machine learning algorithm that the user has expertise to use, then the metadata tracks the split data by inserting a new feature with the name split that has the value of train and test.

For the fourth and last step **Generating Results** the metadata is used in conjunction with the original data to reconstruct and enhance the train and test step result with the protected features. The steps of the pipeline can be summarized as:

1. Utilizing the metadata, we create a two new datasets from the original dataset that has at last one protected feature and utilizing methods and analysis a one dataset

[1] <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>

are created without key features like protected features and correlation features are utilized to create the last dataset without the sensitive attributes and it's proxy.

2. The datasets, without the key features, is divided in two. A training dataset and test dataset. The metadata is updated with the tracking of changes.

3. The models selected are then trained and tested.

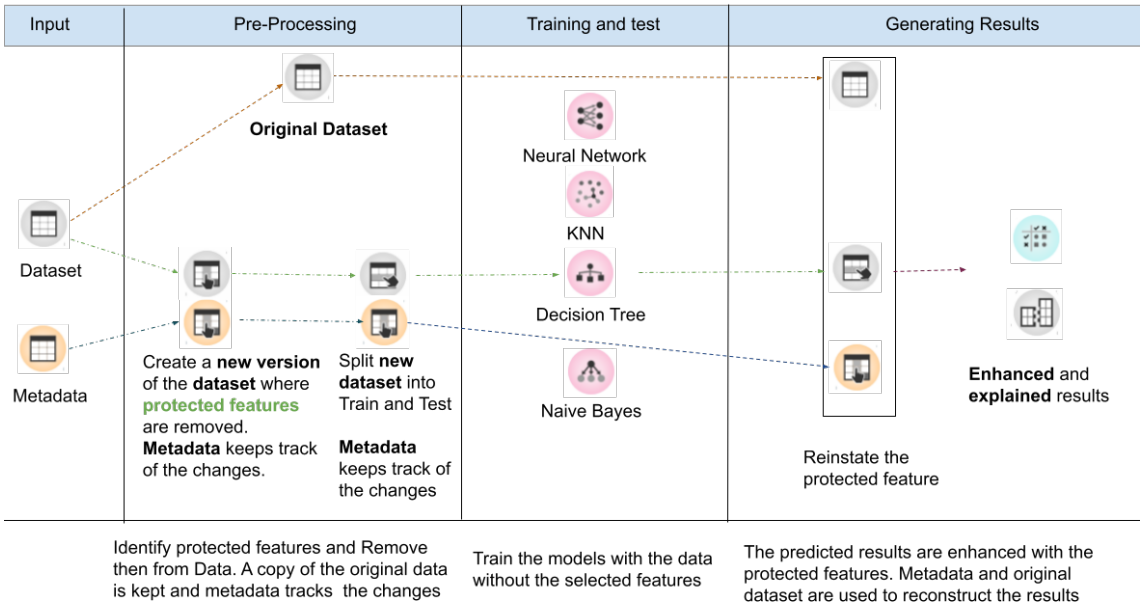
4. Metadata and original dataset are used to reconstruct the predicted results enhanced with the key features and then statistical analyse can be done in the data to classify the models in fairness.

To assess the impact of the knowledge-driven pipeline has on the fairness of predictive model we evaluate it empirically in the next chapter.

Our goal is to answer the following research questions:

RQ1) What is the pipeline impact on the fairness of the models? **RQ2)** Is the accuracy of the models used in the pipeline affected? **RQ3)** How can the pipeline help in analysing the fairness of models?

Figure 3.1: Demonstration of the Pipeline



Source: The Author

4 EMPIRICAL EVALUATION

To perform the empirical evaluation, we divided the experiments into three steps. First, we test the models using all the features provided by the datasets. We utilize the knowledge-driven pipeline to remove all the protection features from the learning, simulating, in this case, a more real user case of a dataset, and then, for the last test, utilizing the pipeline and the fact that the original dataset has at least one protected feature, we used feature correlation and information gain to discover possible proxy for the protected features of the dataset. The results of the models are empirically studied and compared.

4.1 Datasets

The three datasets chosen to realize the empirical evaluation are datasets that contains at least one protected attribute from 2.1 and they can be found in the internet [1], the three dataset are:

1. Adult Income (Dua and Graff 2017): US Adult Census data relating income to social factors such as age, education, race, etc. Barry Becker extracted the US adult income dataset from the 1994 US Census Database. The dataset consists of anonymous information such as occupation, age, country of origin, race, capital gain, capital loss, education, working-class and more. Each row is labeled as having a salary greater than ">50K" or "<=50K". The protected features are Ethnic group, gender, marital status, and Relationship. Utilizing Spearman correlation for the protected features.We found the features Occupation, Native Country and Education can be proxies for one or more sensitive attributes of the dataset.

   The distribution of the protected features are:

   - feature sex is: 10771 (33.08%) Female and 21790(66.92%) Male.
   - For the feature race the distributions are: White 27816(85.43%), Black 3124 (9.59%), Asian-Pac-Islander 1039(3.19%), Americ-Indian-Eskimo311(0.96%) and Other 271(0.83%).
   - For the feature martial-status: Divorced 4443 (13.65%), Married-AF-spouse 23(0.07%), Married-CIV-spouse 14976 (45.99%), Married-spouse-absent 418

[1] <https://www.kaggle.com/wenruliu/adult-income-dataset>, <https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset> and <https://www.kaggle.com/danofer/compass>

(1.28%), Never married 10683 (32.81%), Separated 1025(3.15%) and Widowed 993(3.05%).

- For the feature Relationship: Husband 13193(40.52%), Unmarried 3446(10.56%), Not-in-family 8305(25.51%), Other-relative 981(3.01%), Own child 5068(15.56%) and Wife 1568(4.82%).

feature sex is: 1395 (19.34%) Female and 5819(80.66%) Male. For the feature race the distributions are: 3696(51.23%) African-American, 32(0.44%) Asian, 2454 (34.02%) Caucasian, 637 (8.83%) Hispanic, 18(0.25%) Native-American, 377 (5.23%) Other.

Credit Card Customers Default (Yeh and Lien 2009): This dataset contains information about default payments, demographic factors, credit data, payment history, and account statements for Taiwanese credit card customers from April 2005 to September 2005. The purpose of this dataset is to predict whether the customer will default the payment next month. The protected features of this dataset are SEX and Marriage.

The distribution of the protected features are:

- For the feature Martial Status: Married 13659 (45.25%), Single 15964 (53.21%) and Others 323 (1.08%) .
- For the feature Gender: 18112 (60.37%) Female and 11888(39.63%) Male.

Utilizing Spearman correlation for the protected features we found the features Education and LIMIT-BAL can be proxy for one or more sensitive attribute of the dataset.

COMPAS (raw data) (Barenstein 2019): Correctional Criminal Management Profile for Alternative Sanctions is a popular commercial algorithm used by judges and probation officers in the US to score a criminal defendant's probability of recidivism (reoffense). The purpose of this dataset is to predict whether there has been a recurrence of the offense. The sensitive features of this dataset are Sex and Ethnic group. The distribution of the protected features are:

- feature sex is: 1395 (19.34%) Female and 5819(80.66%) Male.
- For the feature race the distributions are: 3696(51.23%) African-American, 32(0.44%) Asian, 2454 (34.02%) Caucasian, 637 (8.83%) Hispanic, 18(0.25%) Native-American, 377 (5.23%) Other.

feature sex is: 1395 (19.34%) Female and 5819(80.66%) Male. For the feature race the distributions are: 3696(51.23%) African-American, 32(0.44%) Asian, 2454 (34.02%) Caucasian, 637 (8.83%) Hispanic, 18(0.25%) Native-American, 377 (5.23%) Other.

Utilizing Spearman correlation for the protected features we found the features decile-score and vr-charge-desc can be proxies for one or more sensitive attributes of the dataset.

4.2 Predictive Models

We employ four supervised learning algorithms to realize the experiment of the empirical evaluation. The models to conduct the experiment were chosen due to the wide knowledge about the models they are:

1. **Naive Bayes**.probabilistic classifier based on Bayes' theorem with the assumption of feature independence.

2. **Neural Network**.A multi-layer perceptron (MLP) algorithm with back propagation. The parameters utilized in the experiment are: Neuron in hidden layers = 100; activation= ReLu; Regularization = 0.0001; maximum number of iterations = 200.

3. **Decision Tree**. A Decision Tree algorithm splits the data into nodes by class purity (information gain for categorical and MSE for numeric target variable). The parameters utilized in the experiment are: minimal number of instances in leaf is 2; the minimal subset is 5; the limit of maximal tree depth is 100.

4. **k-nearest neighbors algorithm (KNN)**. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. The parameter used in the evaluation is k=5.

4.3 Metrics

In this experiment, we use two types of metrics, the first type is utilized to measure the performance of each model, and the second type quantifies the fairness of a model. The metrics utilized to measure the performance of the model are:

1. **Area under curve AUC**. The area under the receiver-operating curve.

2. **Classification accuracy**. The proportion of correctly classified examples.

3. **F-1**. The weighted harmonic mean of precision and recall

4. **Precision**. The proportion of true positives among instances is classified as positive.
5. **Recall**. The proportion of true positives among all positive instances in the data.
6. **Standard deviation**. The measure of the amount of variation or dispersion of a set of value.

The metrics utilized for fairness in classification problems in the Evaluation are:

1. **Group fairness**. A classifier satisfies this definition if protected and unprotected groups are equally likely to be classified as a positive class.
2. **Fairness Through Unawareness**. An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process.

4.4 Implementation of the pipeline for the evaluation

The implementation of the pipeline was done in a naive approach, employing the metadata only to keep the history of changes in the data. The protected features criterion is used from Table 2.1. It provides the required knowledge to the pipeline to transform the original dataset into another without the protected features. Then, the transformed dataset is split into train and test data. The result of the models, trained on the transformed dataset, are then enhanced with the protected attributes by utilizing the metadata and the original dataset to reconstruct the results. These steps are repeated for the last test, where using the protected feature of the dataset, we test the binary correlation of each sensitive attribute with all attributes. The correlation above |0.1| is then removed from the dataset with the protected features. The correlation of |0.1| was chosen to eliminate the majority of proxy variable because (2) says that a slightly correlation on its own is too small to predict someone's gender with high accuracy. However, if numerous such features are available, as is the case in a typical browsing history, the task of predicting a sensitive attribute becomes feasible at high accuracy levels. The results enhanced with the key elements are then analyzed with the metrics for fairness and performance.

To better visualize the results of the tests, we created two types of graphs, the data that has more than two values in a feature is displayed as a radar chart and the feature that has two values or less as a bar plot. The Charts are used in the evaluation of the metric of group fairness and performance, to analyze the fairness metric we can look at the distributions of the values of positive outcome, the more balanced the chart is the

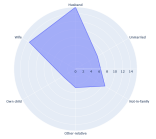
closer it is to achieving group fairness. In the Performance metric, the higher the value the better, the highest value one feature can achieve is one. For the test of Fairness through unawareness, we utilized two tables to exemplify the test first, we created an artificial instance that is put to be classified by all models trained without the pipeline, then the previously created instance is changed by modifying only the protected feature X, and then the result is put in a table. In constructing the second table, we redo the previous steps, only changing the models, the new models are trained in utilizing the pipeline to remove the sensitive attributes.

The following sections present the analysis of each model trained with all features, then utilizing the pipeline to remove only the protected elements and then excluding the protected attributes with the detected proxy features. Therefore for each protected characteristic, we provide three graphs. The first is the results of the experiments using all features. The second is the results with the sensitive attributes hidden from the trained model, and the last is removing the sensitive elements in the company of their detected proxies.

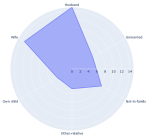
4.5 Results for the Adult dataset

The evaluation of the adult dataset is done by dividing the problem into two, the performance of the model 4.5 and by evaluating the fairness criteria for classification problems. We selected group equity and fairness through unawareness. Group fairness in this dataset is done by checking the probability of a positive random instance for each subgroup of sensitive attributes. An example in the case of the Adult Income dataset would be to predict the equal probability of males and females having more than 50k year income P(d = >50k| G = male) = P(D=>50k| G=female). The result of the probabilities of each attribute are in Figure 4.2, 4.4, 4.3 and 4.1. Figures 4.3, 4.4, 4.1 and 4.2 show that when utilizing the pipeline to remove the protected features, the distribution of positive outcomes becomes smoother, balancing the concentration of positive predictions in a sensitive attribute value in the case of Marital Status. When analyzing the removal of the proxy features, we see that the model tree for the sensitive attribute Ethnic group is better balanced than other models. The training of the models, utilizing the pipeline to remove the protected features, denotes a decrease in the odds difference for almost all tested models; this result can be seen as a step toward achieving group fairness. Tables 4.1 and **??** represents the fairness through unawareness test. In Table 4.1, we can see that

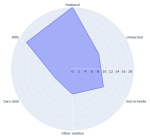
Figure 4.1: Radar Chart of Adult dataset with the percentage of positive outcome for the protected feature Relationship.



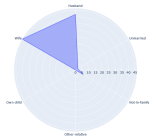
(a) KNN all features



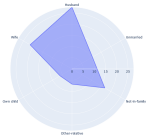
(b) KNN without protected features



(c) KNN removing protected features and proxy variables



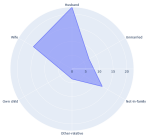
(d) DT all features



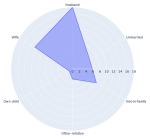
(e) DT without protected features

(f) DT removing protected features and proxy variables

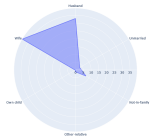
(g) NN all features



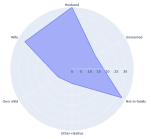
(h) NN without protected features



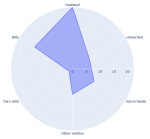
(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features



(k) Naive Bayer's without protected features



(l) Naive Bayer's removing protected features and proxy variables

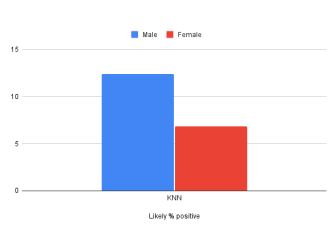
Source: The Author

in an instance modifying the protected feature, the outcome changes, but applying the pipeline, we can see the effects of fairness through unawareness in Table **??**. Figure 4.5 shows that the models do not suffer a drastic decrease in accuracy, recall, AUC, and other metrics when utilizing our approach, while it removes the criteria of using a protected attribute directly in the decision making, respecting fairness through unawareness, it does not mean that the models trained are fair.

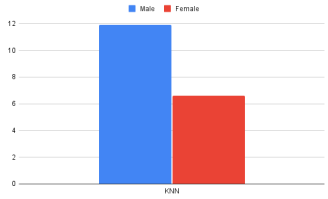
4.6 Results for the Credit Card Customers Default dataset

In the Credit Card Customers Default dataset, we aim to classify if the person that took the credit will default on the next month's payment. Checking group fairnessm was

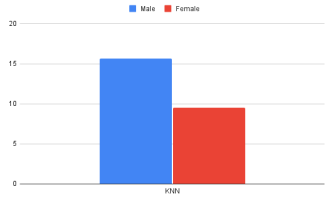
Figure 4.2: Bar Chart of Adult dataset with the percentage of positive outcome for the protected feature Gender.



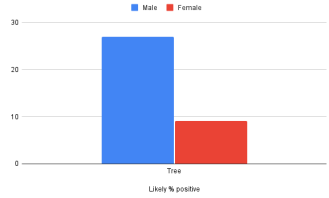
(a) KNN all features



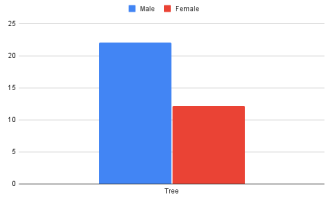
(b) KNN without protected features



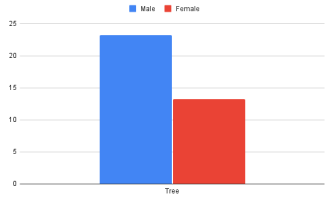
(c) KNN removing protected features and proxy variables



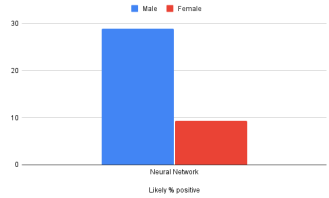
(d) DT all features



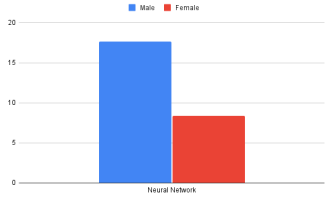
(e) DT without protected features



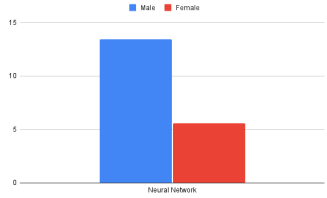
(f) DT removing protected features and proxy variables



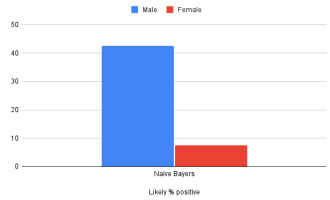
(g) NN all features



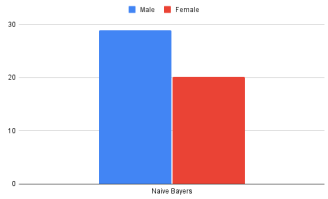
(h) NN without protected features



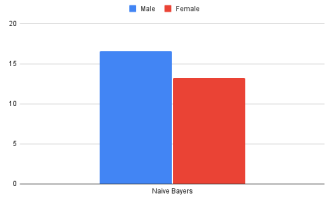
(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features



(k) Naive Bayer's without protected features



(l) Naive Bayer's removing protected features and proxy variables

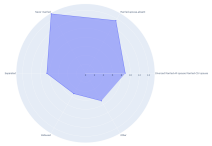
Source: The Author

Table 4.1: Adult Income dataset Fairness Thjorugh unawareness using the feature gender

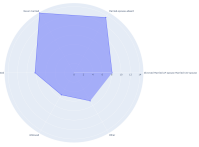
Models	*Male W/O Pipeline*	*Female W/O pipeline*	*Male using the pipeline*	*Female using the pipeline*
KNN	<= 50k	<= 50k	<= 50k	<= 50k
Neural Network	**>50k**	**<= 50k**	**<=50k**	**<= 50k**
Tree	<= 50k	<= 50k	<= 50k	<= 50k
Naive Bayes	**>50k**	**<=50k**	**<=50k**	**<=50k**

Source: The Author

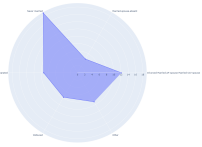
Figure 4.3: Radar Chart of Adult dataset with the percentage of positive outcome for the protected feature Marital Status.



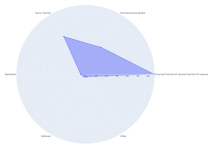
(a) KNN all features



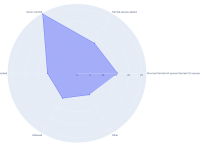
(b) KNN without protected features



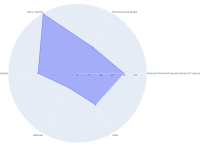
(c) KNN removing protected features and proxy variables



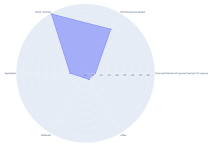
(d) DT all features



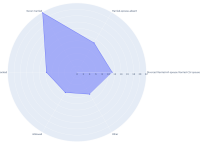
(e) DT without protected features



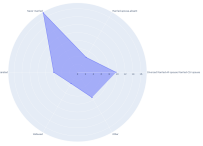
(f) DT removing protected features and proxy variables



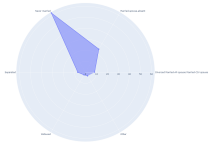
(g) NN all features



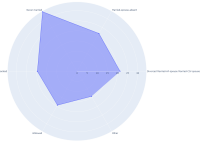
(h) NN without protected features



(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features

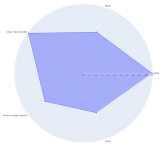


(k) Naive Bayer's without protected features

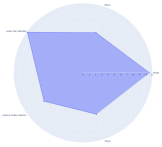
(l) Naive Bayer's removing protected features and proxy variables

Source: The Author

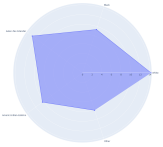
Figure 4.4: Radar Chart of Adult dataset with the percentage of positive outcome for the protected feature Ethnic group.



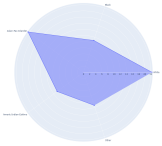
(a) KNN all features



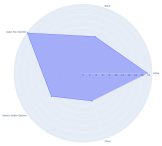
(b) KNN without protected features



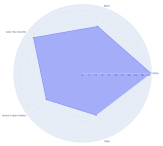
(c) KNN removing protected features and proxy variables



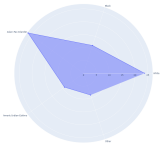
(d) DT all features



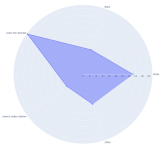
(e) DT without protected features



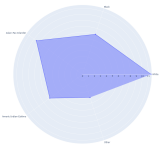
(f) DT removing protected features and proxy variables



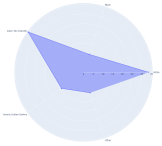
(g) NN all features



(h) NN without protected features

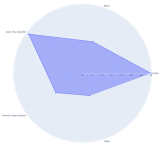


(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features

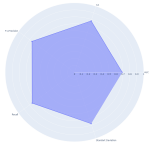
(k) Naive Bayer's without protected features



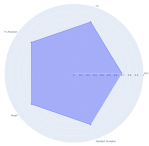
(l) Naive Bayer's removing protected features and proxy variables

Source: The Author

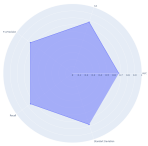
Figure 4.5: Radar Chart of Adult dataset with the performance of the models.



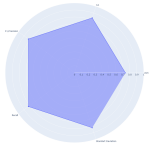
(a) KNN all features



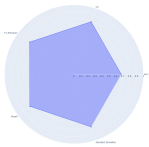
(b) KNN without protected features



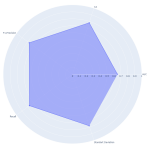
(c) KNN removing protected features and proxy variables



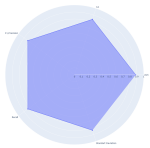
(d) DT all features



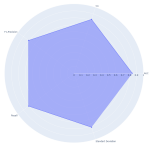
(e) DT without protected features



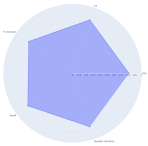
(f) DT removing protected features and proxy variables



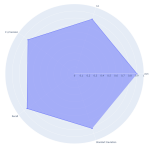
(g) NN all features



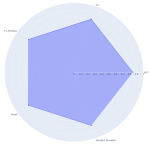
(h) NN without protected features



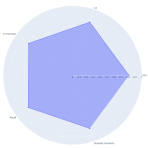
(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features



(k) Naive Bayer's without protected features



(l) Naive Bayer's removing protected features and proxy variables

Source: The Author

Table 4.2: Credit Card dataset outcome for Fairness Thorugh Unawareness using the feature gender

Models	*Male W/O pipeline*	*Female W/O pipeline*	*Male using the pipeline*	*Female using*
KNN	0	0	0	0
Neural Network	**0**	**1**	**0**	**0**
Tree	0	0	0	0
Naive Bayes	0	0	0	0

Source: The Author

done by calculating the probability of paying the next month for the feature marital status that has the values of single, married, and other. The other protected feature gender was also calculated, the gender attribute has the values of 'male' and 'female'. The dataset can pass the group fairness criteria If the probability of each attribute for the feature marital status is equal (P(d = 1| G = single) = P(D=1| G = married) = P(D=1| G = other)) and the probability of paying the next month is equal for both genders(P(d = 1| G = male) = P(D=1| G = female)).

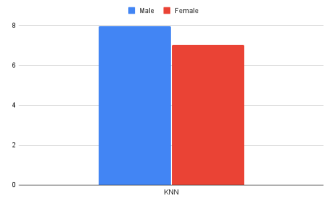
Analyzing Figures 4.6 and 4.7, we see the effects of the pipeline in the Naive Bayes model showing an approximation of the probabilities of a positive outcome. This approximation makes the model closer to group equity for the feature Marital Status. Removing the proxy variable, the decision tree model was the one with the most balanced positive outcome of the models for the Marital status feature. However, the gender feature has minimal impact on the distribution of the probabilities in removing the attributes and the identified proxy.

Figure 4.8 shows the performance of each model, and analyzing the figure can see that by removing the proxy features with the protected attributes, the performance of each tested model does not suffer a considerable impact. In Table 4.2 we can see in the Neural Network model that, for instance, modifying a protected feature, the outcome changes. Applying the pipeline, we can see the effects of fairness through unawareness for the same example in Table **??**.

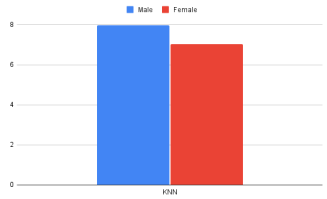
4.7 Results COMPAS dataset

The COMPAS dataset is used to classify if the criminal will recommit a felony between two years. Testing group fairness was done by calculating the probability of two years records for each sensitive attribute of the data. In the case of gender group equity is the probability of male and female being equal (P(d = 1| G = male) = P(D=1| G = female)).

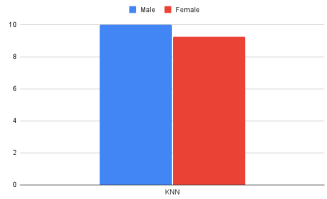
Figure 4.6: Bar Chart of Credit Card dataset with the percentage of positive outcome for the protected feature Gender.



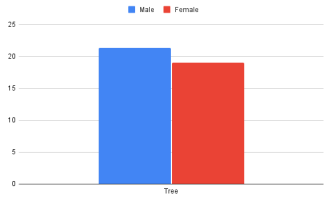
(a) KNN all features



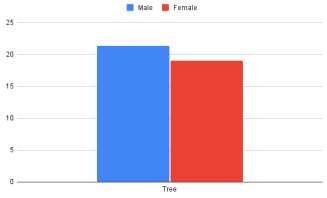
(b) KNN without protected features



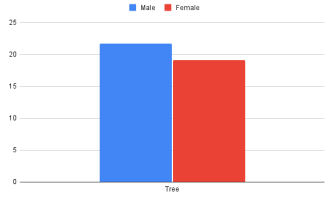
(c) KNN removing protected features and proxy variables



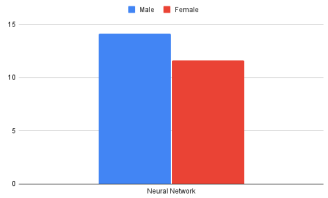
(d) DT all features



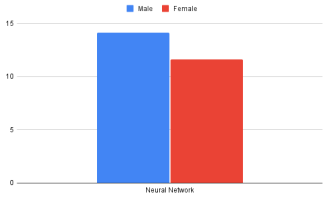
(e) DT without protected features



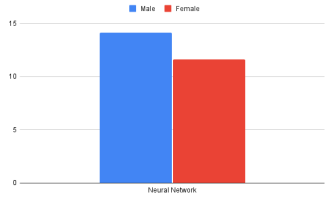
(f) DT removing protected features and proxy variables



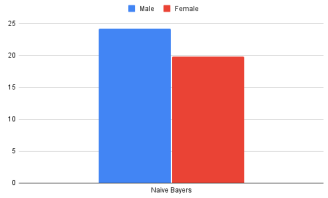
(g) NN all features



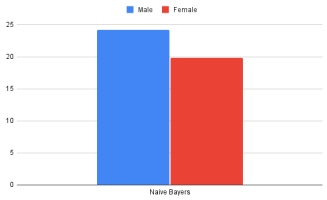
(h) NN without protected features



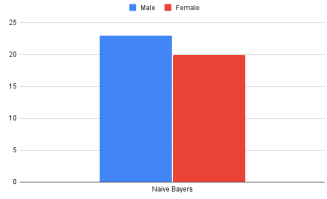
(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features



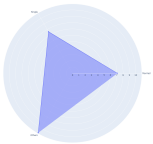
(k) Naive Bayer's without protected features



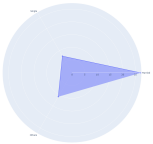
(l) Naive Bayer's removing protected features and proxy variables

Source: The Author

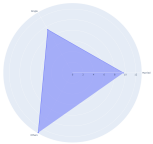
Figure 4.7: Radar Chart of Credit Card dataset with the percentage of positive outcome for the protected feature Marital Status.



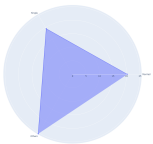
(a) KNN all features



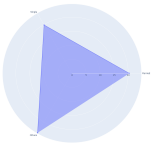
(b) KNN without protected features



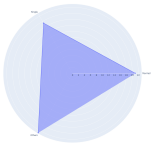
(c) KNN removing protected features and proxy variables



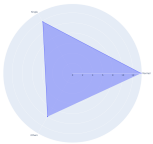
(d) DT all features



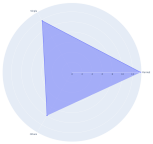
(e) DT without protected features



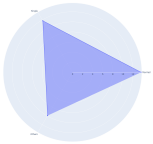
(f) DT removing protected features and proxy variables



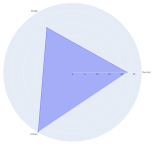
(g) NN all features



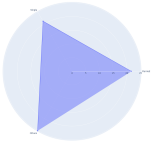
(h) NN without protected features



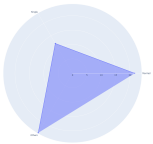
(i) NN removing protected features and proxy variables



(j) Naive Bayer's all features



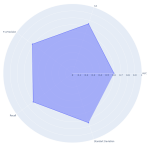
(k) Naive Bayer's without protected features



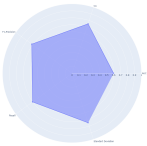
(l) Naive Bayer's removing protected features and proxy variables

Source: The Author

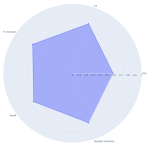
Figure 4.8: Radar Chart of the performance of Credit Card default dataset



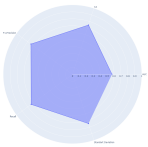
(a) KNN Performance with all features



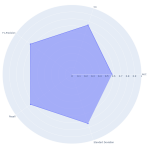
(b) KNN Performance Without protected features



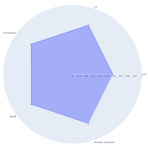
(c) KNN Performance without protected features and proxy variables



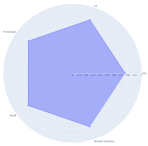
(d) DT Performance with all features



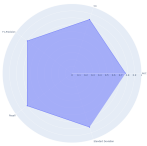
(e) DT Performance Without protected features



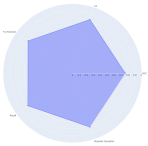
(f) DT Performance without protected features and proxy variables



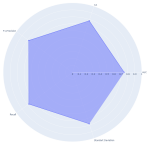
(g) NN Performance with all features



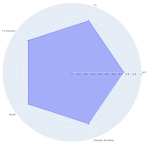
(h) NN Performance Without protected features



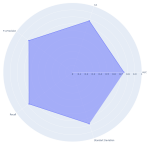
(i) NN Performance without protected features and proxy variables



(j) Naive Bayer's Performance with all features



(k) Naive Bayer's Performance Without protected features



(l) Naive Bayer's Performance without protected features and proxy variables

Source: The Author

The probability of the feature ethinic group is equal for each value if the feature ((P(d = 1 | g = AfricanAmerican) = (P(d = 1 | g = Asian) (P(d = 1 | g = Caucasian) (P(d = 1 | g = Hispanic) (P(d = 1 | g = Native American) (P(d = 1 | g = Other)).

We can use Figures 4.10 and 4.9 to verify the positive outcome for each protected feature, viewing the results of the graphs, we see that removing the protected attributes with the pipeline, the most affected model is the decision tree model for the feature ethnic group and the neural network model for the feature gender, analyzing the removal of the proxy features we can see a greater effect of the positive outcome for the value female in the neural network while others models suffered minimal impact from the removal. Analyzing the ethnic group, the removal of the detected proxies shows an impact in the following models: decision tree, Neural Network and K-Nearest Neighbors. Inspecting the balanced distribution of the probabilities for the group fariness criteria, the Neural Network model and Naive Bayer's are the closest in achieving group equity. Figure 4.11 we can analyze the performance of each model in the case of the Neural Network model the accuracy of the model makes it an unsuitably model even if it is the model closest to achieving group equity in both features.

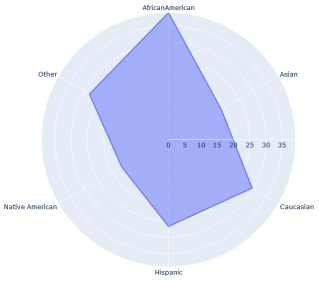
4.8 Overall analysis of results

The result of the datasets shows a variation in the accuracy of the models after utilizing the pipeline. In the Adult dataset, the performance metrics used on the model's tree, naive Bayes, and neural network have decreased, showing that the model's default used protected attributes to train and has a higher hit rate. The neural network model shows the pipeline's impact in removing the sensitive features on Credit Card Default's dataset. The metrics increase for the neural network model in the COMPAS dataset, making it more accurate.

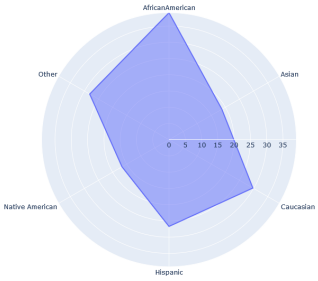
With the results, we conclude that, for this experiment, all models can be used, with the pipeline with minimal impact on accuracy. Regarding fairness, we can explore the pipeline to search each resource for the positive result of the protected resources contained in the dataset. We explore that by removing the protected attributes, the models stop being directly influenced by them, complying with the metric of fairness through unawareness.

Following the pipeline facilitates the analysis of the sensitive attributes contained within the dataset, verifying each protected element for the percentage of positives, true

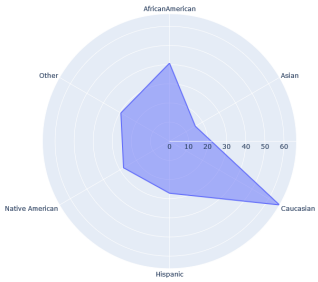
Figure 4.9: Radar Chart of the percentage of positive outcome for Ethinic group of the COMPAS dataset.



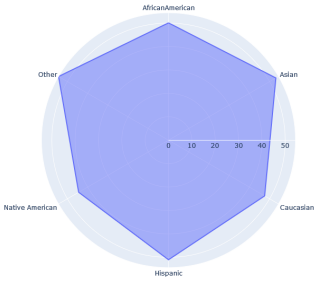
(a) KNN Performance with all features



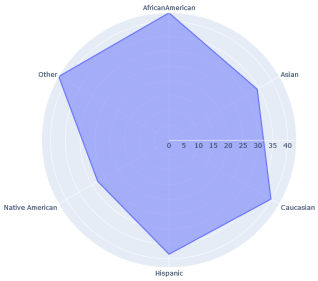
(b) KNN Performance Without protected features



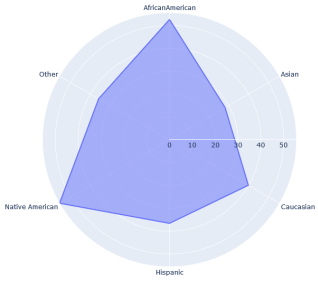
(c) KNN Performance without protected features and proxy variables



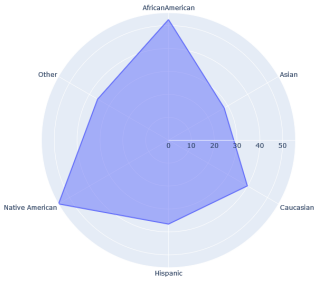
(d) DT Performance with all features



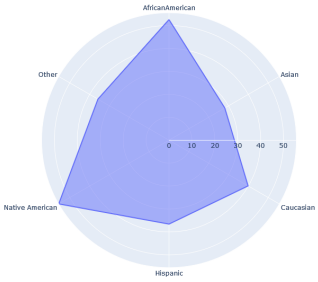
(e) DT Performance Without protected features



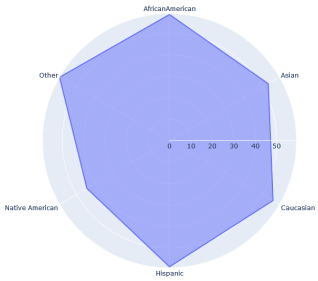
(f) DT Performance without protected features and proxy variables



(g) NN Performance with all features

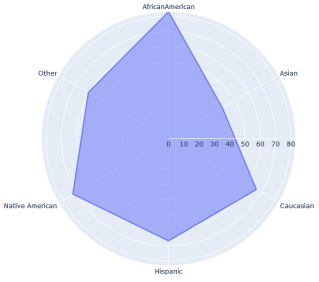


(h) NN Performance Without protected features

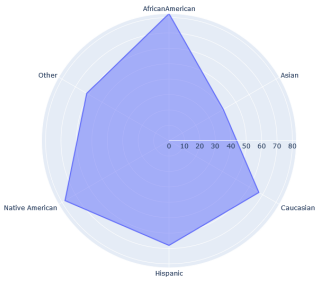


[]

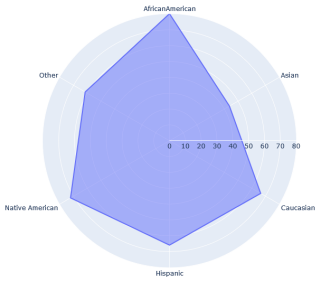
(i) NN Performance without protected features and proxy variables



(j) Naive Bayer's Performance with all features



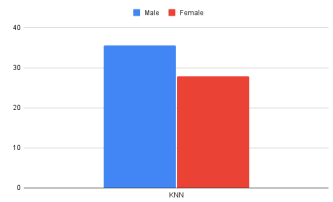
(k) Naive Bayer's Performance Without protected features



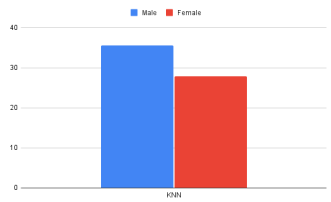
(l) Naive Bayer's Performance without protected features and proxy variables

Source: The Author

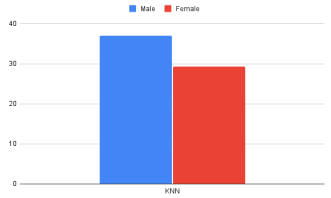
Figure 4.10: Bar Chart of the positive outcome for the Gender feature of the COMPAS dataset in KNN, NN (Neural Network), DT (Decision Tree) and NB (Naive Bayes)



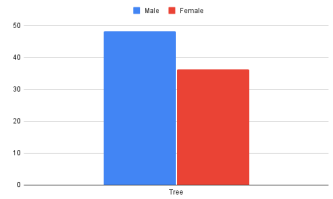
(a) KNN Performance with all features



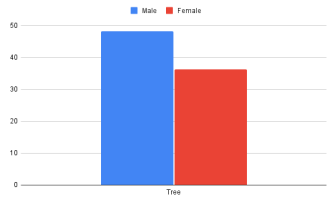
(b) KNN Performance Without protected features



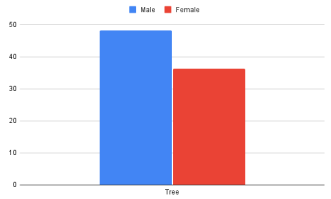
(c) KNN Performance without protected features and proxy variables



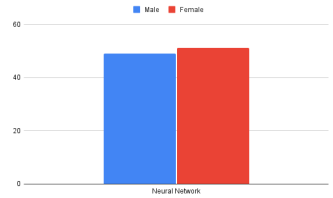
(d) DT Performance with all features



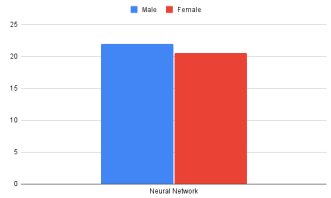
(e) DT Performance Without protected features



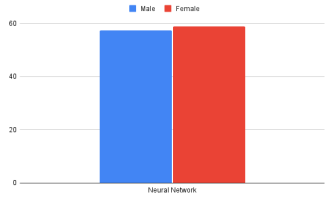
(f) DT Performance without protected features and proxy variables



(g) NN Performance with all features

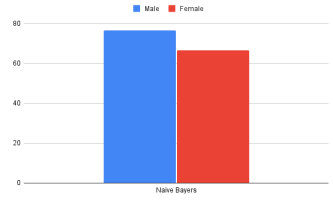


(h) NN Performance Without protected features

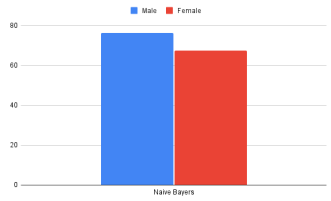


[]

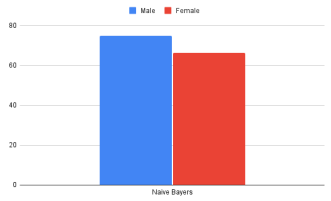
(i) NN Performance without protected features and proxy variables



(j) Naive Bayer's Performance with all features



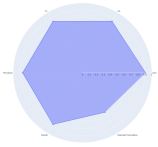
(k) Naive Bayer's Performance Without protected features



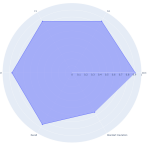
(l) Naive Bayer's Performance without protected features and proxy variables

Source: The Author

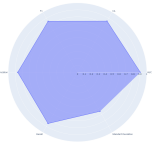
Figure 4.11: Radar Chart of the performance of the COMPAS dataset in KNN, NN (Neural Network), DT (Decision Tree) and NB (Naive Bayes)



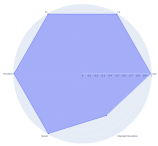
(a) KNN Performance with all features



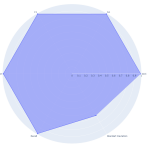
(b) KNN Performance Without protected features



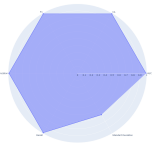
(c) KNN Performance without protected features and proxy variables



(d) DT Performance with all features



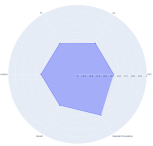
(e) DT Performance Without protected features



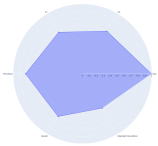
(f) DT Performance without protected features and proxy variables

(g) NN Performance with all features

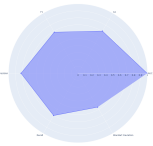
(h) NN Performance Without protected features



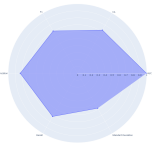
(i) NN Performance without protected features and proxy variables



(j) Naive Bayer's Performance with all features



(k) Naive Bayer's Performance Without protected features



(l) Naive Bayer's Performance without protected features and proxy variables

Source: The Author

positives, negatives, and false negatives. These analyses can then be utilized, with fairness criteria, invalidating the model at an acceptable level of fairness. The effects of the pipeline can be described as a fairness analysis facilitator where models can be trained and then studied, with statistical criteria for classification problems.

4.9 Limitations of the empirical study

The empirical study in utilizing the pipeline has some limitations. The generic pipeline was implemented with limited features, using the basic implementations of the metadata. The metadata was only in charge of keeping track and pointing the data utilized in the evaluation steps to its original source.

The criteria of protected features to generate the new dataset were extracted from Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA), exemplified in table[2.1]. The selected criteria did not factor others types of perspectives, like data protection. The chosen attributes also have another limitation, where only proxy variables with a binary correlation of the sensitive property in the dataset was considered, other types of proxy variables and confounding variables was not addressed.

Another limitation of the empirical study was the choice of the fairness criteria, we selected two benchmarks to work that we tough were easy to understand, test, and visualize the result, but in (Verma and Rubin 2018) there is a compilation of criteria used to analyze the fairness of classification problem.

5 CONCLUSION

With the advance of big data, machine learning has played crucial roles in automated decision-making for various fields in Job recruiting, sensitive characteristics query is being done by asking for the job seeker's consent, these responses can be, then used to create a dataset that can be used for machine learning as the dataset contains features that characterize a person in a group (e.g., ethnicity, religion, or gender), many of these groups can be sensitive to biases, and the feature(s) representing the distinct groups are considered protected.

We presented an approach to help users of machine learning algorithms make fairer classification decisions. Our presented approach was a pipeline used together with a dataset that contains at least one protected element. In the empirical evaluation, we evaluated the impact of the proposed pipeline on the performance of the models noticing that by removing the protected features and then their proxy, the reduction of the accuracy of the models was minimal, and in some instances, it increased the accuracy as seen in the neural network model of the COMPAS dataset.

The empirical evaluation was also evaluated concerning the effect of the pipeline on the fairness of the model. To evaluate fairness, the criteria used were fairness through unawareness and group fairness.

The test results show that applying the pipeline to remove the protected features makes the model comply with the criterion of fairness through unawareness and helps balance the distribution of the positive outcome for the values of the sensitive characteristics of the datasets. Removing the proxy variables in some cases makes the models more balanced for the positive outcomes making the models closer to approaching the criterion of group fairness and, in other cases, increasing the disparity of the positive results. Reviewing the empirical evaluation results, we can conclude that the proposed pipeline has a minimal impact on the accuracy of the models selecting features to remove from the train data. The model is influenced more in the accuracy, with more elements removed. In the case of removing the proxy variables together with the sensitive characteristics, the user has to select the approach to detect the proxy variable considering its impact on the models.

The effect of the pipeline on the fairness of the models is based on the criteria used for the experiment, the choice of problem classification criteria is of paramount importance, and the user must analyze case by case what his model wants to classify and

use the criteria that fit the dataset and model. This way, the pipeline can better display it's effect by using the metadata to analyze the trained model so the user can make the proper adjustment.

In the future, we will be working on creating a software implementation of the pipeline with several criteria and features with the metadata helping programmers and data analysts create fairer classifications models.

REFERENCES

BARENSTEIN, M. **ProPublica's COMPAS Data Revisited**. 2019.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. **Fairness and Machine Learning**. [S.l.]: fairmlbook.org, 2019. <http://www.fairmlbook.org>.

BONO, T.; CROXSON, K.; GILES, A. Algorithmic fairness in credit scoring. **Oxford Review of Economic Policy**, v. 37, n. 3, p. 585–617, 09 2021. ISSN 0266-903X. Available from Internet: <https://doi.org/10.1093/oxrep/grab020>.

BOSE, A. J.; HAMILTON, W. L. Compositional fairness constraints for graph embeddings. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). **Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 715–724. Available from Internet: <http://proceedings.mlr.press/v97/bose19a.html>.

CHEN, J. et al. Fairness under unawareness. **Proceedings of the Conference on Fairness, Accountability, and Transparency**, ACM, Jan 2019. Available from Internet: <http://dx.doi.org/10.1145/3287560.3287594>.

CHEN, J. et al. Fairness under unawareness: Assessing disparity when protected class is unobserved. In: **Proceedings of the Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2019. (FAT* '19), p. 339–348. ISBN 9781450361255. Available from Internet: <https://doi.org/10.1145/3287560.3287594>.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Available from Internet: <http://archive.ics.uci.edu/ml>.

FELDMAN, M. et al. Certifying and removing disparate impact. In: CAO, L. et al. (Ed.). **Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015**. ACM, 2015. p. 259–268. Available from Internet: <https://doi.org/10.1145/2783258.2783311>.

GELLERT, R. et al. A comparative analysis of anti-discrimination and data protection legislations. In: **Discrimination and privacy in the information society**. [S.l.]: Springer, 2013. p. 61–89.

GRGIC-HLACA, N. et al. The case for process fairness in learning: Feature selection for fair decision making. In: . [S.l.: s.n.], 2016.

KANG, J. et al. Multifair: Multi-group fairness in machine learning. **CoRR**, abs/2105.11069, 2021. Available from Internet: <https://arxiv.org/abs/2105.11069>.

KEARNS, M. J. et al. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: DY, J. G.; KRAUSE, A. (Ed.). **Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018**. PMLR, 2018. (Proceedings of Machine Learning Research, v. 80), p. 2569–2577. Available from Internet: <http://proceedings.mlr.press/v80/kearns18a.html>.

KOUMERI, L. K.; NáPOLES, G. Bias quantification for protected features in pattern classification problems. In: . [S.l.: s.n.], 2021.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. **CoRR**, abs/1908.09635, 2019. Available from Internet: <http://arxiv.org/abs/1908.09635>.

VERMA, S.; RUBIN, J. Fairness definitions explained. In: BRUN, Y.; JOHNSON, B.; MELIOU, A. (Ed.). **Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018**. ACM, 2018. p. 1–7. Available from Internet: <https://doi.org/10.1145/3194770.3194776>.

YEH, I.; LIEN, C.-H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. **Expert Systems with Applications**, v. 36, p. 2473–2480, 03 2009.

ZAFAR, M. B. et al. Fairness constraints: Mechanisms for fair classification. In: SINGH, A.; ZHU, X. J. (Ed.). **Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 54), p. 962–970. Available from Internet: <http://proceedings.mlr.press/v54/zafar17a.html>.