ICS5510 Assignment

Carmel Gafa

***Abstract*—TODO: abstract here**
***Index Terms*—TODO: keywords**

I. Introduction

This exercise will explore the well-known COMPAS dataset using several machine-learning techniques. We will also look into the ethical implications of predictive risk assessment models. We have taken the opportunity of this study to implement some of the techniques discussed in ICS5510, like imputation and encoding, to help in data preparation, linear regression, neural networks and others as the tools used for prediction.

Wherever possible, we preferred the manual implementation of some of the steps over the functionality available in popular Python libraries to appreciate the techniques implemented more thoroughly.

A. History of the COMPAS tool

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset and tool have a controversial history rooted in its use for assessing the likelihood of recidivism among criminal defendants. Developed by Northpointe, COMPAS gained widespread adoption in the U.S. judicial system for pretrial risk assessments and sentencing decisions. This tool is helpful in various stages of the criminal justice process, including bail, sentencing, and parole decisions.

However, in 2016, an investigative report by ProPublica revealed significant racial biases in the tool's predictions. The report found that COMPAS disproportionately labelled Black defendants as high risk for reoffending while underestimating the risk for white defendants, even when both groups had similar criminal histories. This revelation sparked a broader debate about using algorithmic tools in criminal justice and their transparency and fairness.

The COMPAS tool has not been directly the subject of lawsuits, but its use in judicial decisions has led to legal challenges. For instance, in State v. Loomis (2016), the Wisconsin Supreme Court upheld using COMPAS in sentencing. However, judges must be informed about its limitations, particularly its proprietary nature and potential biases. The case highlighted the broader tension between the utility of predictive algorithms and their application's need for accountability and fairness.

B. The COMPAS dataset

The dataset that originates from the COMPAS tool is widely used in criminology and machine learning studies.

The dataset contains attributes such as demographic information, prior charges, juvenile records, and risk scores, including the widely analysed 'decile score', which categorises individuals into ten different risk groups.

The decile score is a critical feature, assigning a numerical value to an individual's likelihood of reoffending. Other important features include the number of prior offences `priors_count` and the type of offence `c_charge_degree`, provide context for these predictions. At the same time, the label `two_year_recid` indicates whether an individual reoffended within two years of their COMPAS assessment.

While the dataset has been instrumental in research aimed at understanding and improving risk prediction models, it has also been the subject of extensive scrutiny due to its implications for fairness and equity in the justice system. A couple of thoughts resulting from this scrutiny include:

- Multiple studies, including the influential ProPublica investigation in 2016, have highlighted racial disparities in the COMPAS predictions. African-American defendants were found to be nearly twice as likely as Caucasian defendants to be labelled as high-risk for recidivism but not reoffend. Conversely, Caucasian defendants were more likely to be classified as low-risk but later reoffend, raising concerns about systemic bias embedded in the algorithm, which could exacerbate existing inequalities in the justice system.
- The COMPAS tool operates as a proprietary black-box model, meaning its internal workings and feature weights are not disclosed to the public or even to the defendants it evaluates. This lack of transparency prevents meaningful scrutiny and accountability, leaving users unable to fully understand or challenge the tool's predictions.
- The COMPAS algorithm relies on historical criminal justice data, which may reflect social and systemic biases. For example, law enforcement practices that can result in sentencing disparities can all influence the patterns observed in the data. Using such data as input, the COMPAS tool risks perpetuating these biases into an electronic tool.
- Some features in the COMPAS dataset, such as age and criminal history, are static and cannot change over time, as this data is based on the date of the COMPAS assessment. We can argue that these features in risk predictions without considering the period after the COMPAS assessment undermines the potential for individuals to reform and leads to insensible punitive outcomes.
- The ethical implications of using predictive algorithms

in high-stakes decisions, such as sentencing and parole, constitute a significant area of concern. The potential for false positives can lead to unjustly harsher treatment, while false negatives can impact public safety.
- The dataset available for research purposes is a reduced version of the original COMPAS data, with several features anonymised or removed. Missing important data introduces limitations for academic studies aiming to replicate or validate the findings from real-world COMPAS applications.

The criticism of the COMPAS tool emphasises the challenges of deploying machine learning systems in sensitive domains like justice. These challenges are not unique to COMPAS but highlight broader issues in applying algorithmic decision-making tools in socially important contexts. They highlight the need for transparency, fairness-aware modelling techniques, and careful ethical evaluations when designing and implementing such tools.

C. Objectives of this work

The main objectives of this study are:

- Analyse the COMPAS dataset and its predictions.
- Prepare the dataset for machine learning through cleaning, transformation, and feature engineering.
- Train and evaluate machine learning models for ethical analysis.
- Investigate potential biases and ethical implications in predictions.

II. Downloading and first look at the dataset

The COMPAS dataset used in this study is publicly available through ProPublica's GitHub repository. This repository contains the dataset and other assets used by ProPublica to investigate the biases present in the COMPAS risk assessment tool.

The file chosen for this analysis is **`compas-scores-two-years.csv`**, as it provides the cleanest and most relevant data for general recidivism prediction. This CSV file contains the key data required for our study, including several attributes related to demographics, criminal history, COMPAS risk scores, and the two-year recidivism outcomes that are important for exploring the predictive capabilities and the ethical implications of machine learning models in the context of recidivism prediction.

The dataset includes important information about individuals. Following an initial analysis, a list of the key fields in the dataset is below.

- Personal Information, includes attributes such as **`age`**, **`race`**, **`age_category`**, etc.
- Case and Event-Related Details are the fields prefixed with **`c_`** that provide a timeline and details of a person's interactions with the criminal justice system.
- Violence Risk Assessment are the fields prefixed with **`v_`** and are associated with the violence risk assessment in COMPAS. This dimension predicts violent recidivism risk.
- Case-Level Details for Violent Recidivism are the fields prefixed with **`vr_`**. These fields provide additional details specific to violent recidivism events.
- Juvenile Criminal Record are the fields prefixed with **`juv_`**. These fields capture information about an individual's juvenile criminal record, which is a key predictor of future adult criminal behaviour.
- Previous Charges and Severity can be deduced from fields such as **`priors_count`** and **`juv_`** fields.
- Additional fields, including **`r_charge_`**, **`r_offense_`**', **`vr_`** fields, **`c_charge_degree`**, and **`c_charge_desc`**, provide a broader perspective on criminal history and severity.
- Two-Year Recidivism, or the **`two_year_recid`** field in the COMPAS dataset, indicates whether an individual reoffended (recidivated) within two years of their initial assessment or release. This field is critical for evaluating the predictive accuracy of the COMPAS risk assessment tool.
- Decile Score is a standardized risk score in the COMPAS dataset. It categorizes an individual's likelihood of recidivism into ten equal groups (deciles) where 1 is the lowest risk, and 10 is the highest risk. Each decile represents approximately 10% of the sample when applied to a norm group.

Suppose that we observe the following data:

Field	Value
`priors_count`	5
`juv_felony_count`	2
`juv_misdemeanor_count`	3
`r_charge_degree`	Felony

We can interpret this as an individual who has five total prior charges, including:

- 2 juvenile felonies
- 3 juvenile misdemeanours
- The severity of previous charges includes felonies (**`r_charge_degree`**).

III. Preparing the data for further analysis and training

Before we can perform any analysis or apply machine learning techniques, it is important to pre-process and prepare the dataset so that we can handle missing values, encode categorical features, and split the data into training, testing, and validation sets. This step will produce a clean dataset for building accurate and unbiased models. The following steps outline the procedures to prepare the dataset for further analysis and training.

A. Initial look at data and missing values handling

The dataset has 7214 instances over 53 columns. The target of the dataset is **`decile_score`**, but the dataset also contains information about whether or not the person recidivised, most notably through the label **`two_year_recid`**.

The first step in data preparation is removing the features irrelevant to this exercise or with over 50% missing records. We removed all the COMPAS-administrative labels and additional recidivism information apart from **two_year_recid**, narrowing the dataset to 17 fields.

The difference between **c_jail_in** and **c_jail_out** was calculated into a new field, **days_in_jail** and the difference between **in_custody** and **out_custody**, in a new field, **days_in_custody**. We subsequently removed the features containing date information from the dataset, together with **days_in_custody**, as it contained no information. At this stage, the dataset contains thirteen features: eight numerical, four categorical, and one descriptive. It also contains two labels, decile_score, which we will treat as the leading label in this exercise and **two_year_recid**, which we are keeping to compare the prediction power of our models to the original one.

B. Imputation of missing data

While examining the resultant dataset, we noticed that **days_b_screening_arrest** has 6907 values that are not null. Whilst it is possible to eliminate the rows that contain the null values at this stage, we replaced the missing values using a KNN imputation technique by grouping the numeric values of this dataset so that we can calculate the missing values. We checked this process by plotting the distribution of **days_b_screening_arrest** before and after imputation to see if any variations occurred.

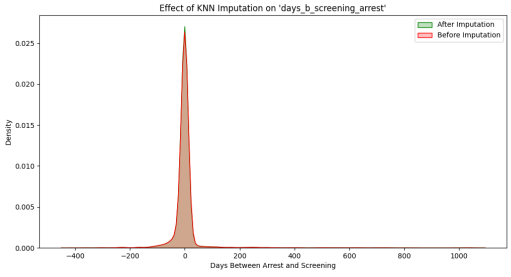


Fig. 1

At this stage, the dataset contains four categorical features that need encoding for machine learning algorithms. This section will focus on converting them into a numerical format using two encoding techniques. The categorical features and their values are listed below:

The following transformations are applied:

- One-Hot Encoding on **sex**, **race**, and **c_charge_degree**, transforming them into binary columns.
- Ordinal Encoding on **age_cat**. This encoding technique was preferred over one-hot in this case as it preserves order,thus respecting the inherent ranking of the category.

The original categorical columns were retained in the dataset for future use in the analysis steps.

Feature	Description	Unique Values
sex	Gender of the individual	[l]Male Female
race	Race of the individual	African-American Caucasian Hispanic Asian Native American Other
age_cat	Age category	Less than 25 25 - 45 Greater than 45
c_charge_degree	Degree of the criminal charge	F (Felony) M (Misdemeanor)

C. Splitting the data into train, test and dev

A stratified shuffle split technique is preferred to create the train, test, and dev datasets whilst ensuring that the splits are proportional by **race**. In the first split, 80% Train and 20% Test are created, whilst in the Second split, The 20% Test is further divided into 10% Test and 10% Dev.

IV. Data Exploration and Visualisation

This section will examine the dataset in more detail to understand the patterns, distributions, and relationships. In this exercise, we will use more of the visual tools available through several Python libraries to identify potential biases, explore correlations between variables, and uncover insights that may influence the outcomes of predictive models.

A. Demographic analysis

We begin this analysis by segmenting the dataset based by race and gender.

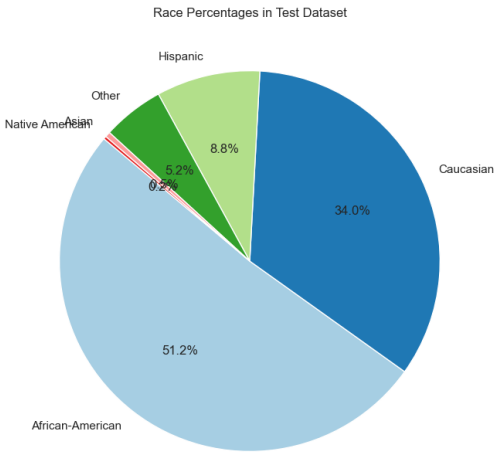
By examining the racial composition of the dataset, we observe the following:

- Over half of the test dataset is composed of African-American individuals, suggesting that the dataset may be imbalanced, with a disproportionate representation of one racial group.
- Asians and Native Americans each makeup only 0.2% of the dataset; this underrepresentation might raise some concerns as it may lead to challenges in statistical analysis or machine learning models. Such concerns include the lack of reliability or significance for these groups due to insufficient data.

Figure 2 also shows our dataset's male/female split, with females comprising only 19.8%. It is, therefore, evident that the female group is underrepresented, which can lead to biased models as models may overfit male patterns and underperform on females and misleading conclusions as insights derived might generalise poorly for the female subgroup.

B. Age distribution analysis

We used a boxplot to illustrate the age patterns across racial groups, helping to identify central values, spread, and any anomalies.



(a) Caption for the first image

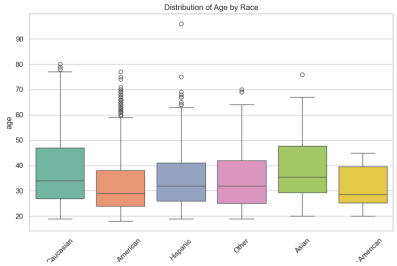
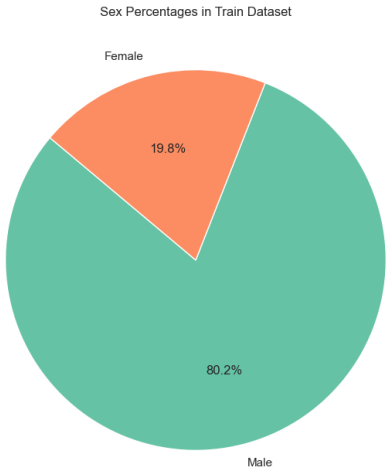


Fig. 3



(b) Caption for the second image

Fig. 2: Overall caption for the figure containing both images

Race	Median Age	Distribution Description
African-American	≈30 years	Concentrated in the 20–40 range, with a relatively narrow spread. We notice outliers above 60 years, indicating fewer older individuals.
Caucasian	≈40 years	Broader age range, from 20 to 70+ years. We notice more older individuals (upper outliers), making this group appear older on average.
Hispanic	≈30–35 years	Moderately broad spread, with most individuals between 20 and 50 years.
Asian	≈30 years	Narrow distribution, concentrated between 25 and 40 years. No outliers.
Native American	≈33 years	Very tight distribution, with all ages clustered closely around the median (little variability). It is important to note that this group accounts for a tiny portion of the population.
Other	≈35–40 years	Similar to Caucasians but with slightly fewer older individuals. The IQR shows a widespread.

REFERENCES

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

APPENDIX

Field Name	Description	Type	Options (if Categorical)	Used
id	Unique identifier for each individual.	Numeric	N/A	No
name	Full name of the defendant	Text	N/A	No
first	First name of the defendant (anonymized).	Text	N/A	No
last	Last name of the defendant (anonymized).	Text	N/A	No
compas_screening_date	Date of the COMPAS assessment.	Date	N/A	Yes
sex	Gender of the defendant.	Categorical	Male, Female	Yes
dob	Date of birth of the defendant.	Date	N/A	No
age	Age of the defendant at the time of assessment.	Numeric	N/A	Yes
age_cat	Age category of the defendant.	Categorical	Less than 25 25 - 45 Greater than 45	Yes
race	Race of the defendant.	Categorical	African-American Caucasian Hispanic Asian Native American Other	Yes
juv_fel_count	Number of juvenile felony offenses.	Numeric	N/A	Yes
juv_misd_count	Number of juvenile misdemeanor offenses.	Numeric	N/A	Yes
juv_other_count	Number of other juvenile offenses.	Numeric	N/A	Yes
priors_count	Number of prior offenses (adult and juvenile).	Numeric	N/A	Yes
days_b_screening_arrest	Days between arrest and COMPAS screening.	Numeric	N/A	Yes
c_jail_in	Jail booking date for the charge.	Date	N/A	No
c_jail_out	Jail release date for the charge.	Date	N/A	No
c_case_number	Case number associated with the charge.	Text	N/A	No
c_offense_date	Date of the alleged offense.	Date	N/A	No
c_arrest_date	Arrest date for the charge.	Date	N/A	No
c_charge_degree	Degree of the charge.	Categorical	F (Felony) M (Misdemeanor)	Yes
c_charge_desc	Description of the charge.	Text	Free text	No
is_recid	reoffended after COMPAS screening.	Binary	0 (No), 1 (Yes)	No
r_case_number	Case number for the re-offense.	Text	N/A	No
r_charge_degree	Degree of the re-offense charge.	Categorical	F (Felony) M (Misdemeanor)	No
r_charge_desc	Description of the re-offense charge.	Text	Free text	No
r_jail_in	Jail booking date for the re-offense.	Date	N/A	No
r_jail_out	Jail release date for the re-offense.	Date	N/A	No
two_year_recid	Label: offense within two years.	Binary	0 (No), 1 (Yes)	Yes
decile_score	COMPAS risk score (1-10).	Numeric	1–10	Yes
score_text	Risk category for general recidivism.	Categorical	Low, Medium, High	Yes
v_type_of_assessment	Type of COMPAS assessment conducted.	Text	Risk of Recidivism	No
v_decile_score	Violent recidivism COMPAS score (1-10).	Numeric	1–10	No
v_score_text	Risk category for violent recidivism.	Categorical	Low, Medium, High	No
start	Start date of the two-year recidivism period.	Date	N/A	No
end	End date of the two-year recidivism period.	Date	N/A	No
event	offense during the two-year period.	Binary	0 (No), 1 (Yes)	No