COMPAS Scores

Defendant Recidivism

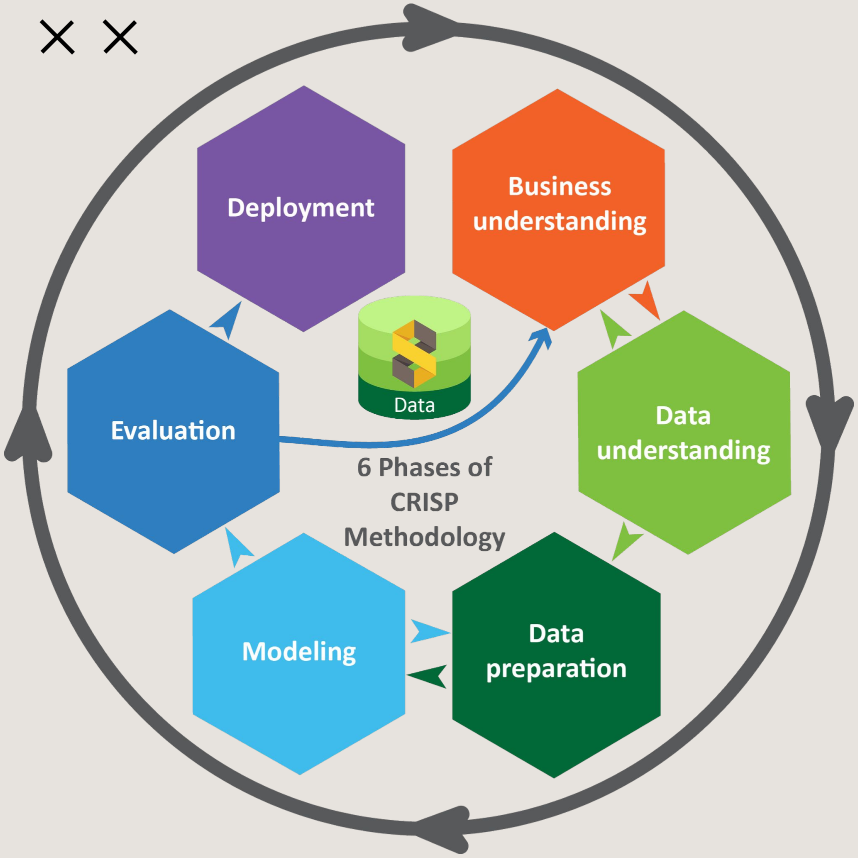
A Machine Learning Model that predicts if a defendant becomes a recidivist.

Group: 404 Name Not Found

Table of contents

Introduction

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.



CRISP-DM Methodology

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining is a process model that serves as the base for a data science process. It has six sequential phases.

Technologies, Tools and Libraries

python™

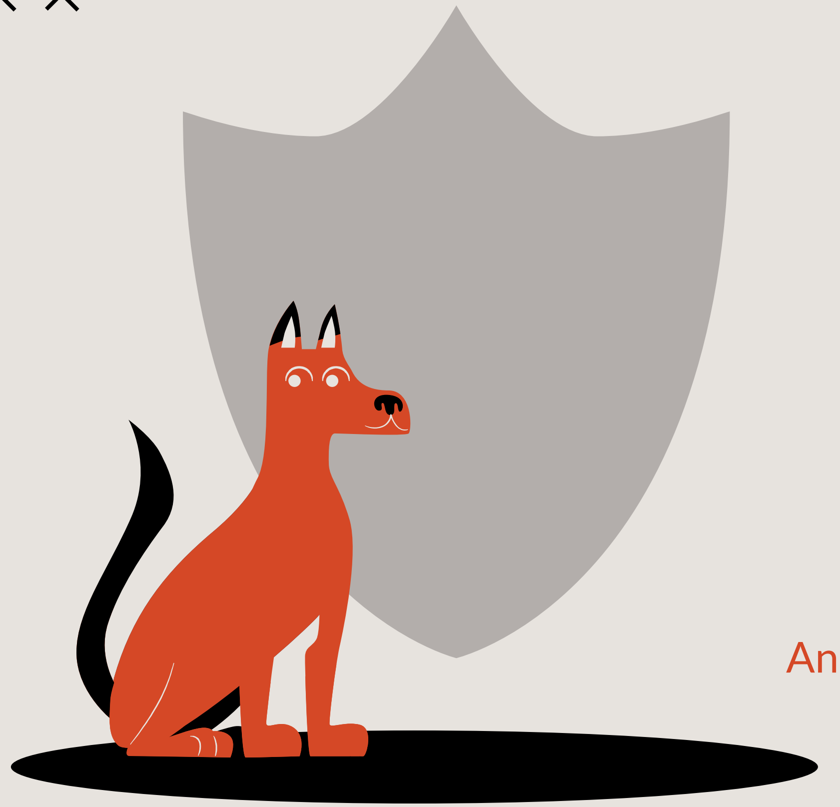
jupyter

pandas

seaborn

scikit learn

GitHub



01

Business Understanding

Analyze project background and main goals.

Background

The project is about **COMPAS Scores** dataset that collects over 11k records (registered from January 2013 to December 2014) and 47 attributes.

Each record represents a **criminal** with a name (first and last), date of birth, age, gender, ethnicity and other information about his/her arrest and screening date, type of offense and the recidivism.

Main Goals

Business Objectives

- Determine and **predict** if a defendant becomes a recidivist.
- It will be useful for **community corrections** applications.

Data Mining Goals

- Binary classification problem.
- If **is_recid** is equal to 0 → the Criminal is not a recidivist.
- If **is_recid** is equal to 1 → the Criminal is a recidivist.

Other Goals

In addiction, we decided to more sub-goals to our analysis.

Violent Recid

Predict if a defendant becomes a violent recid or not.

Days before becoming recid

Predict in mean the number of days that pass from the first crime.

Days before becoming violent recid

Predict the number of days for a violent criminal.

Success Criteria

Success Criteria

Achieve a good level of **accuracy**, **f-measure** (≥0.90).

Business Success Criteria

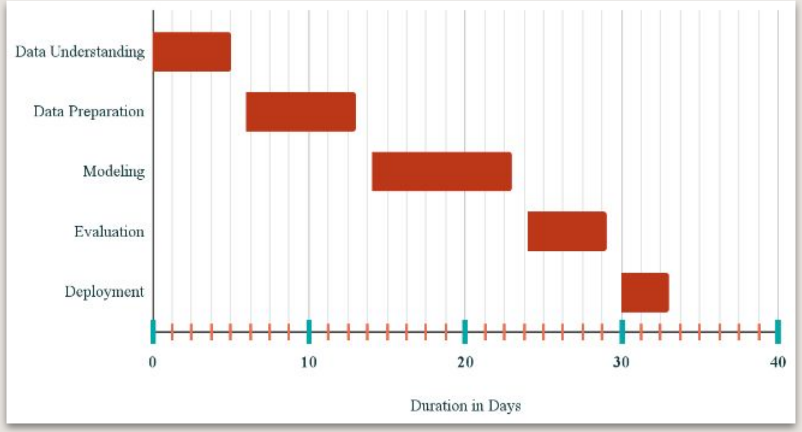
- **Improve** the re-education plan for the defendants
- **Decrease** the cost to maintain detention institutions full
- **Invest** in readmission to improve the community wellness

Project Plan

Project Status

Project Step	Duration	Status	Start	End	Pred.
01 Data Understanding	5 days	Completed	10/05/2022	15/05/2022	//
02 Data Preparation	7 days	Completed	16/05/2022	23/05/2022	01
03 Modeling	10 days	Completed	24/05/2022	02/06/2022	02
04 Evaluation	5 days	Completed	03/06/2022	08/06/2022	03
05 Deployment	4 days	In progress	09/06/2022	12/06/2022	04

Gantt Diagram



02

Data Understanding

- → Data Description
- → Data Exploration
- → Data Quality

Data Description

The final dataset is composed of **11,757 instances** and **47 attributes** that we decided to analyze by dividing in three main groups:

Numerical Attributes

- id
- age
- juv_fel_count
- decile_score
- juv_misd_count
- juv_other_count
- prisors_count
- days_b_screening_arrest
- c_days_from_campas
- is_recid
- num_r_cases
- r_days_from_arrest
- is_violent_racid
- num_vr_cases
- v_decile_score
- decile_score.1

Categorical Attributes

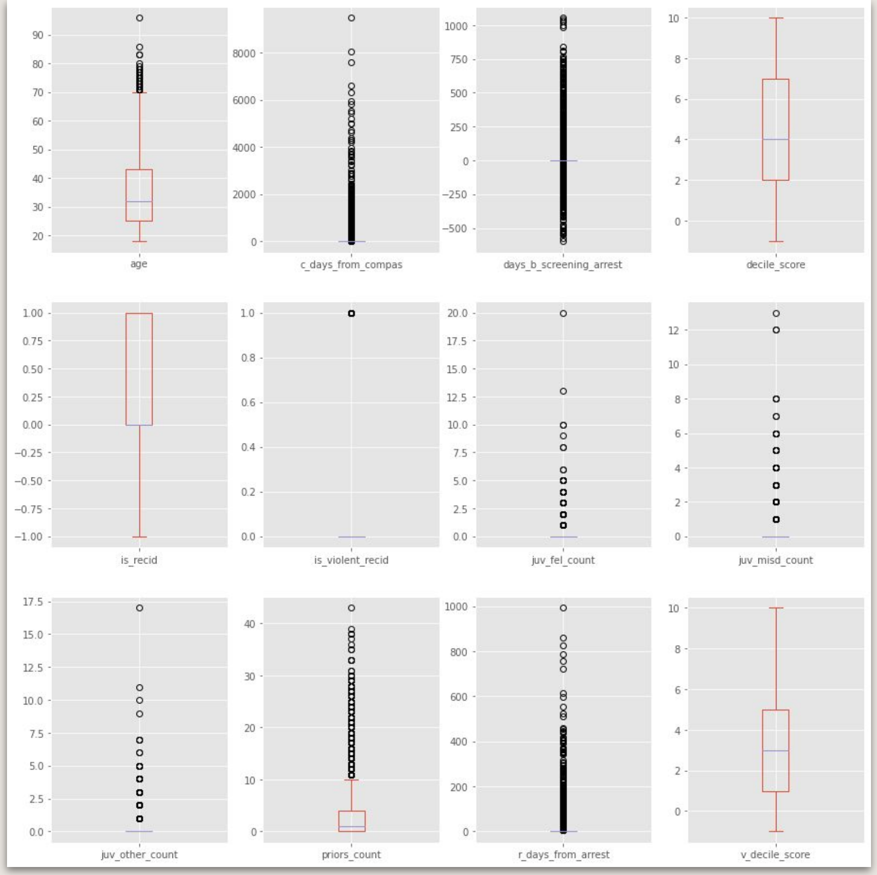
- name
- first
- last
- dob
- age_cat
- sex
- race
- compas_screening_date
- c_jail_in
- c_jail_out
- c_case_number
- c_offense_date
- c_arrest_date
- c_charge_degree
- c_charge_desc
- r_case_number
- r_charge_degree
- r_offense_date
- r_charge_desc
- r_jail_in
- r_jail_out
- vr_case_number
- vr_charge_degree
- vr_offense_date
- vr_charge_desc
- v_type_of_assessment
- v_score_text
- v_screening_date
- type_of_assessment
- score_text
- screening_date

Dates Attributes

- dob
- compas_screening_date
- c_jail_in
- c_jail_out
- c_offense_date
- c_arrest_date
- r_offense_date
- r_jail_in
- r_jail_out
- vr_offense_date
- v_screening_date
- screening_date

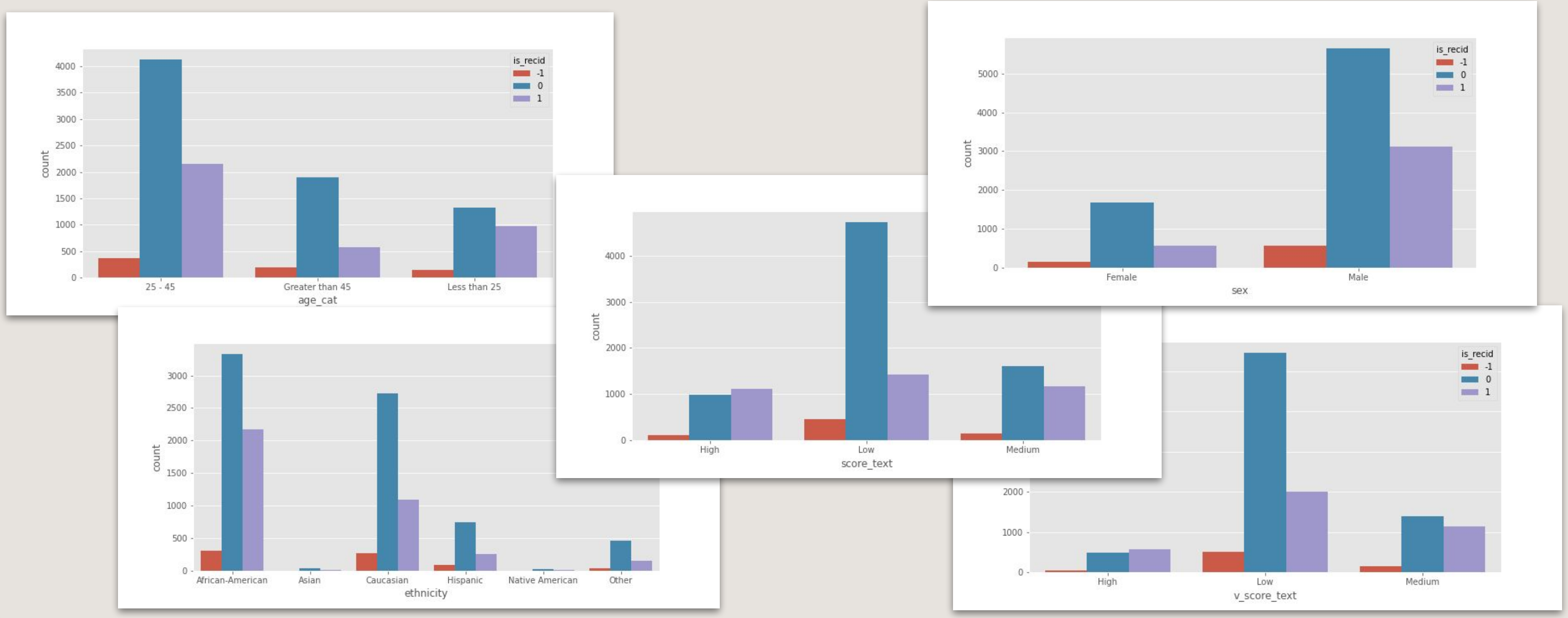
Data Exploration - Numerical Attributes

We have 16 numerical attributes. The following box plots, show better some linear values and few mainly characterized by outliers.



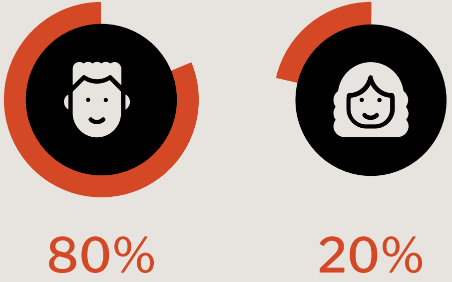
Data Exploration - Categorical Attributes

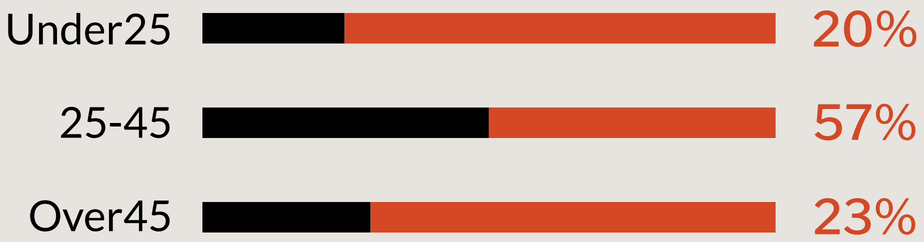
We have 31 non numerical attributes, most of them are text descriptions, identification codes, or dates. The most relative attributes are **sex**, **age_cat** and **race** (that we rename in **ethnicity** for moral reasons). Regarding the decile_score we decided to maintain the attributes **score_text** and **v_score_text.** The following plots shows the correlation between the categorical attributes and the class label:



Data Exploration - Categorical Attributes

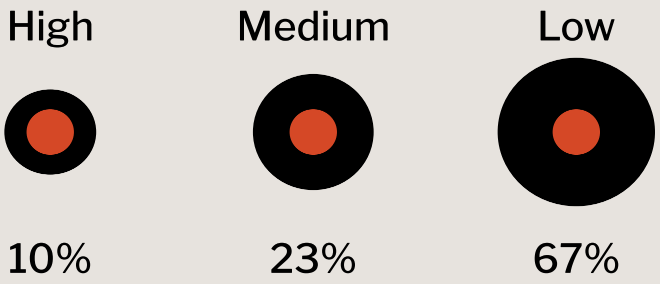
Gender and age





Ethnicity

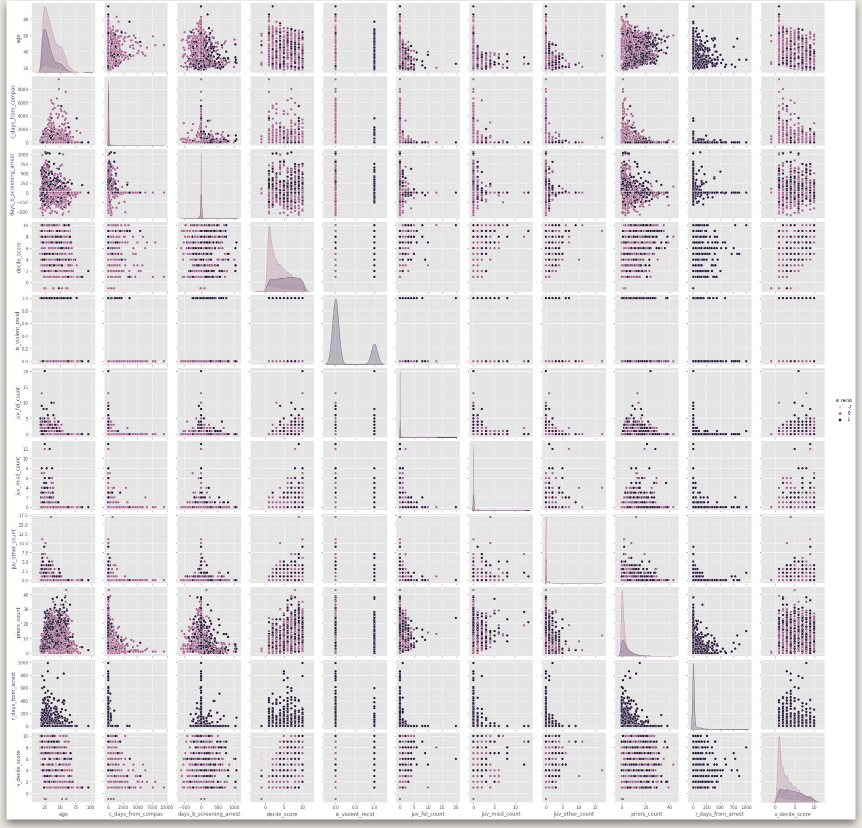
Score Text



Data Exploration - Dates Attributes

About the **12 dates attributes**, since we noticed a **high variability**, we had some problems in plotting them in a sustainable time range. So, we decided to show, on the report, the **values_count()** and their statistics description in order to complete our Data Understanding report.

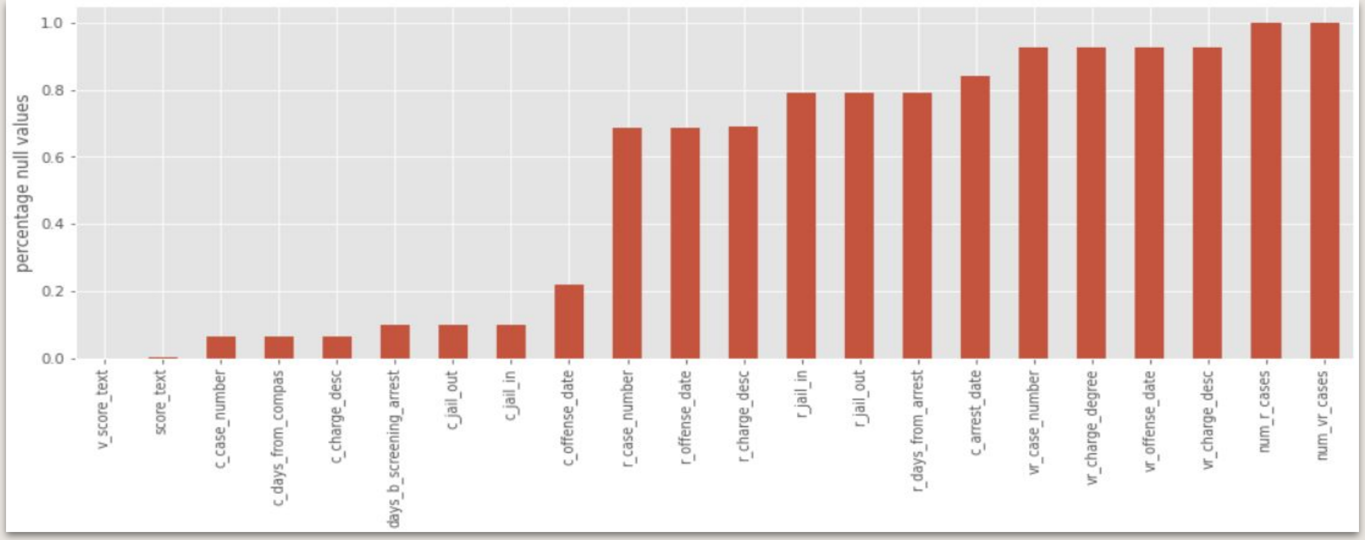
Data Exploration - Scatter Plots



Data Exploration - Correlation Matrix

Data Quality - Null Values

The following plot shows the percentage of null values (>0%):



By analyzing the percentage of NULL values for each attribute, we intuitively saw that **num_r_cases** and **num_vr_cases** are useless for our analysis because they are totally NULL.

Data Quality - Inconsistent Values

Duplicated attributes

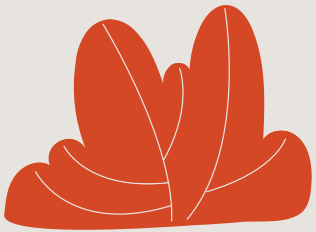
Another useless attribute is **decile_score.1** cause it is a duplicate, so we did not consider it for our analysis.

Incorrect values

- No coherence between **name** and the attributes **first** and **last**.
- Inconsistent values for the attribute **dob** (dates of birth) respect to the defendant's current **age** attribute.

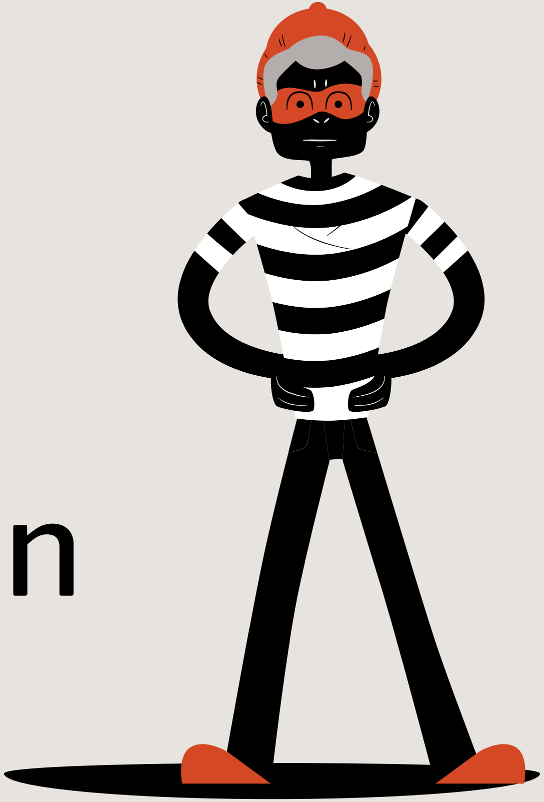
Other considerations

- The attributes **is_recid** and **is_violent_recid** have only **0** and **1** as possible values. We will erase the record with values **-1** (assuming that this value indicates an unknown value).



03

Data Preparation



Prepare our Dataset for the Modeling phase.

This is an important phase in which we can apply strategies and mechanism in order to select, clean and transform our data.

Data Selection

Based on the previous step done, we identified the most relevant information to reach our goal.

Included attributes

Included Attributes
• age_cat • c_offensive_date • is_recid • is_violent_recid • r_offensive_date • race • sex • score_text • v_score_text • vr_offensive_date

We decided to delete the attributes:

- with high percentage of null values
- with high variability
- with inconsistency
- with outliers
- regarding dates
- less relevant to our analysis

Important Steps

Data cleaning

- **columns.difference()** to clean dataset.
- Some is_recid value are “**-1**”, assumed that are unknown values.
- Not binarized **is_recid** and **is_violent_recid**.

Formatted data

- Renamed race to **ethnicity** for ethical reason.
- Renamed the values of age_cat:
  - Less than 25 => **young**.
  - 25 - 45 => **adult**.
  - Greater than 45 => **senior**.

Data Integration

- We didn’t integrate or merge anything in our dataset.

Data Construction

- No need to add new data for the analysis.

04 Modeling

After the Data Understanding and the Data Preparation steps, we are ready to proceed with the Modeling phase.

Classifiers Description

Since we worked on a binary classification task and since the final dataset contains mainly categorical attributes, some of the most common algorithms used to solve this type of problem are the following:

Naïve Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.

Decision Tree Classifier

A DT is a flowchart like tree structure with:
- nodes → test on an attribute;
- branches → outcome of the test;
- terminal nodes (leaves) → class label.

K-Nearest-Neighbors Classifier

It is non-parametric, that means no underlying assumptions about the distribution of data.

Random Forest Classifier

It is basically a set of decision trees (DT) from a randomly selected subset of the training set.

AdaBoost Classifier

AdaBoost classifier combines weak classifier algorithms to form strong classifiers.

Test Design

We splitted into Training Set and Test Set, respectively with a percentage of 75 and 25. Then we performed Undersampling or Oversampling in order to balance the recid dataset and the violent recid dataset.

Build Model

To find out the best parameters for each model's algorithm we used a function called GridSearchCV. The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

Model Assessment

We fitted the models and, based on the predictions set, we got the values of the quality measure: accuracy, recall, precision and F-measure. We decided to look at F-measure to select the best model for our dataset.

	recid dataset	violent recid dataset
Naïve Bayes	Accuracy: 0.5934 Recall: 0.6670 Precision: 0.4315 F-measure: 0.5240	Accuracy: 0.5929 Recall: 0.3805 Precision: 0.2663 F-measure: 0.3133
Decision Tree	Accuracy: 0.6402 Recall: 0.5816 Precision: 0.4707 F-measure: 0.5203	Accuracy: 0.5702 Recall: 0.3850 Precision: 0.2514 F-measure: 0.3042
KNN	Accuracy: 0.3442 Recall: 0.9903 Precision: 0.3374 F-measure: 0.5033	Accuracy: 0.6771 Recall: 0.1593 Precision: 0.2483 F-measure: 0.1941

	recid dataset	violent recid dataset
Random Forest	Accuracy: 0.6380 Recall: 0.6151 Precision: 0.4699 F-measure: 0.5328	Accuracy: 0.5713 Recall: 0.3850 Precision: 0.2522 F-measure: 0.3047
AdaBoost	Accuracy: 0.6369 Recall: 0.5838 Precision: 0.4671 F-measure: 0.5190	Accuracy: 0.6199 Recall: 0.3584 Precision: 0.2812 F-measure: 0.3152

05

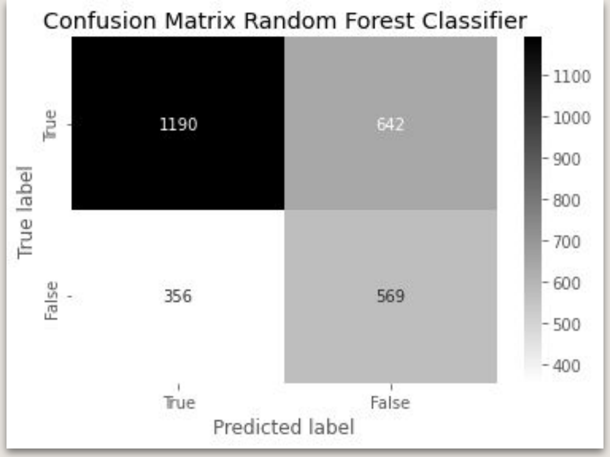
Evaluation

In this final phase, we evaluated the best model according to the previous steps.

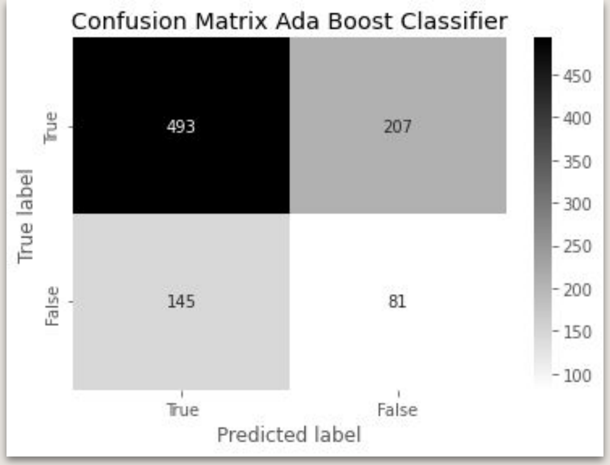
Results Evaluation

According to our results, the best models are the **Random Forest Classifier** (f-measure>0.53) for recid prediction and the **AdaBoost Classifier** (f-measure>0.30) for the violent recid dataset.

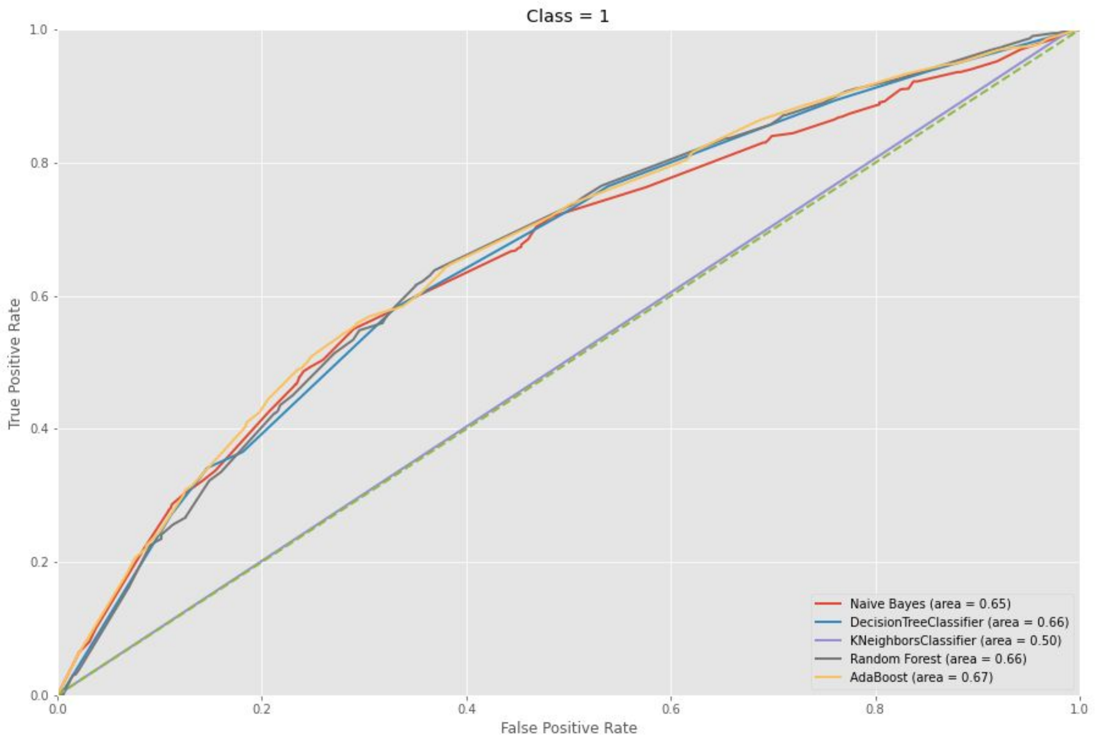
Recid Dataset



Violent Recid Dataset



ROC Curve - Recidivism



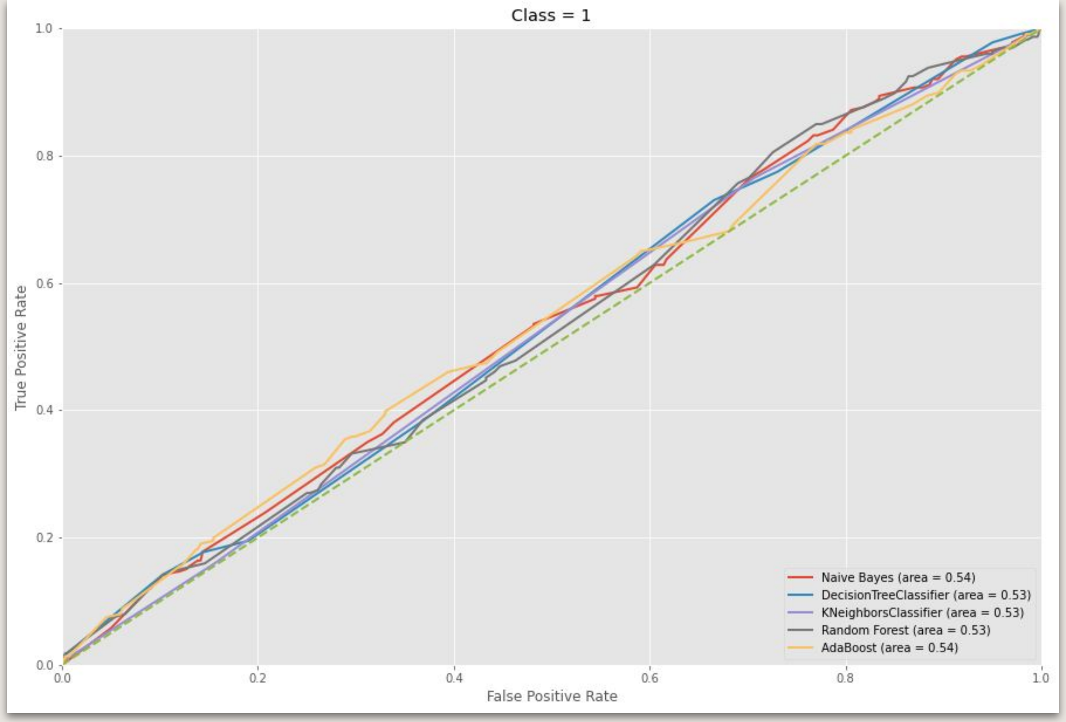
Classification Report - Random Forest Classifier

	precision	recall	f1-score	support
0	0.77	0.65	0.70	1832
1	0.47	0.62	0.53	925
accuracy			0.64	2757
macro avg	0.62	0.63	0.62	2757
weighted avg	0.67	0.64	0.65	2757

Classification Report - Ada Boost Classifier

	precision	recall	f1-score	support
0	0.76	0.66	0.71	1832
1	0.47	0.58	0.52	925
accuracy			0.64	2757
macro avg	0.61	0.62	0.61	2757
weighted avg	0.66	0.64	0.64	2757

ROC Curve - Violent Recidivism



Classification Report - Ada Boost Classifier

	precision	recall	f1-score	support
0	0.77	0.70	0.74	700
1	0.28	0.36	0.32	226
accuracy			0.62	926
macro avg	0.53	0.53	0.53	926
weighted avg	0.65	0.62	0.63	926

Classification Report - Naive Bayes Classifier

	precision	recall	f1-score	support
0	0.77	0.66	0.71	700
1	0.27	0.38	0.31	226
accuracy			0.59	926
macro avg	0.52	0.52	0.51	926
weighted avg	0.65	0.59	0.61	926

Sub goals

Predict the number of days from the first crime to the recid offense date.

Recid Dataset

	coefficient
age_cat_senior	47.857954
age_cat_young	-17.088324
ethnicity_Asian	89.822557
ethnicity_Caucasian	-27.749853
ethnicity_Hispanic	-4.900140
ethnicity_Native American	-0.566236
ethnicity_Other	-6.337896
score_text_Low	26.047265
score_text_Medium	15.459139
sex_Male	-0.796888

Violent Recid Dataset

	coefficient
age_cat_senior	124.789211
age_cat_young	-8.889943
ethnicity_Asian	-42.569574
ethnicity_Caucasian	-75.149561
ethnicity_Hispanic	45.466648
ethnicity_Native American	364.672239
ethnicity_Other	-84.999001
v_score_text_Low	-32.905414
v_score_text_Medium	23.736487
sex_Male	-45.110255

Meaning of the Coefficients

× ×	Age Cat	Sex	Ethnicity	Score
1st Dataset	young	male	caucasian	high
2nd Dataset	young	male	other	low

	Age Cat	Sex	Ethnicity	Score
1st Dataset	senior	female	asian	low
2nd Dataset	senior	female	native american	medium

Process Review

Accuracy

We did not reach a high level of accuracy

Other Models

Explore and analyze the datasets with more types of models

Model Parameters

Investigate and find best parameters

Predicting Days

Find a better way to explore linear regression models

Balance Dataset

Try to balance the datasets in a better way

Assessment

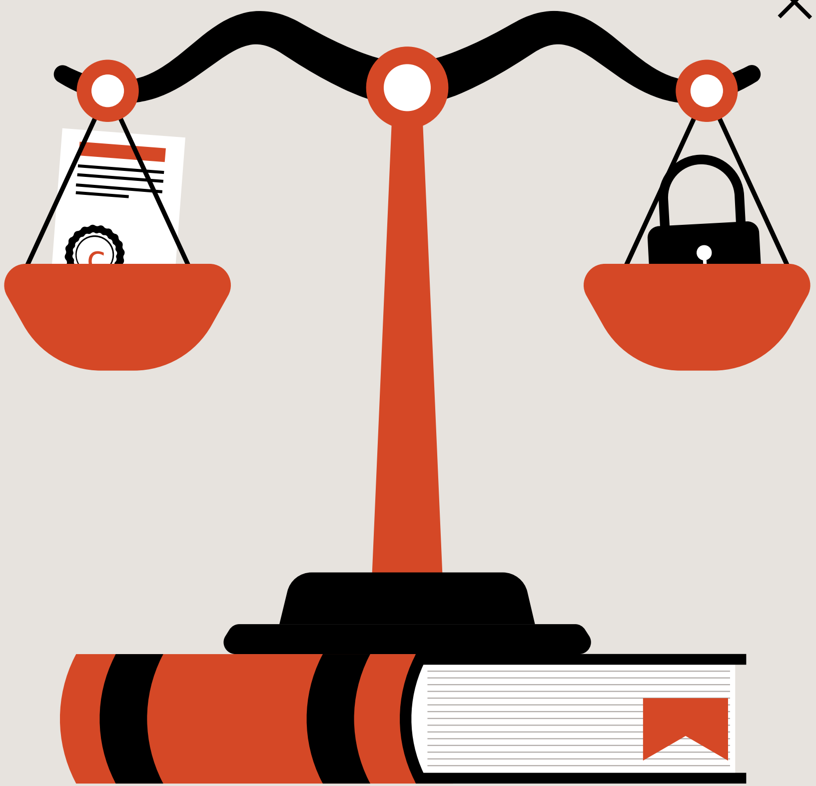
Try new statistics libraries and technologies

Conclusion

Since this dataset is not well studied in literature, we wanted to analyze it in a simple and readable way.

- We reached a discrete accuracy level (0.60)
- High biases present in this unbalanced dataset
- Represents a deeper social issue
- We analyzed different sub goal in order to explore different point of views
- Do our best to reach a high degree of inclusiveness for the wellness of the whole society

Thanks for your attention



Group **404 Name Not Found**

- Canonaco Martina [231874]
- Gabriele Giada [235799]
- Gena Davide [231873]
- Morello Michele [223953]

Data Analytics (Machine Learning) Project - a. y. 2021/2022