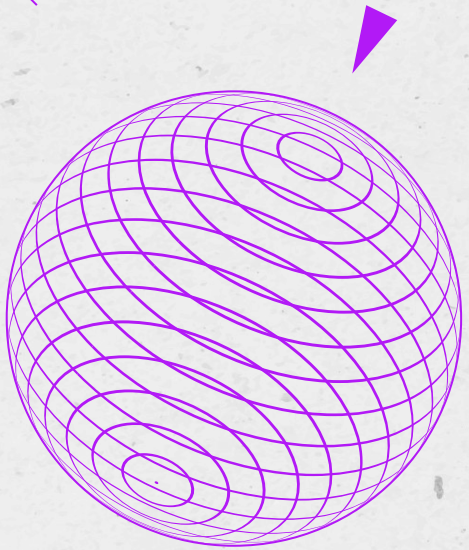
P5 Final Presentation

D. Paplauski, P.M. Ketterman,
R.M.A van Nee, V. Li - Group 14



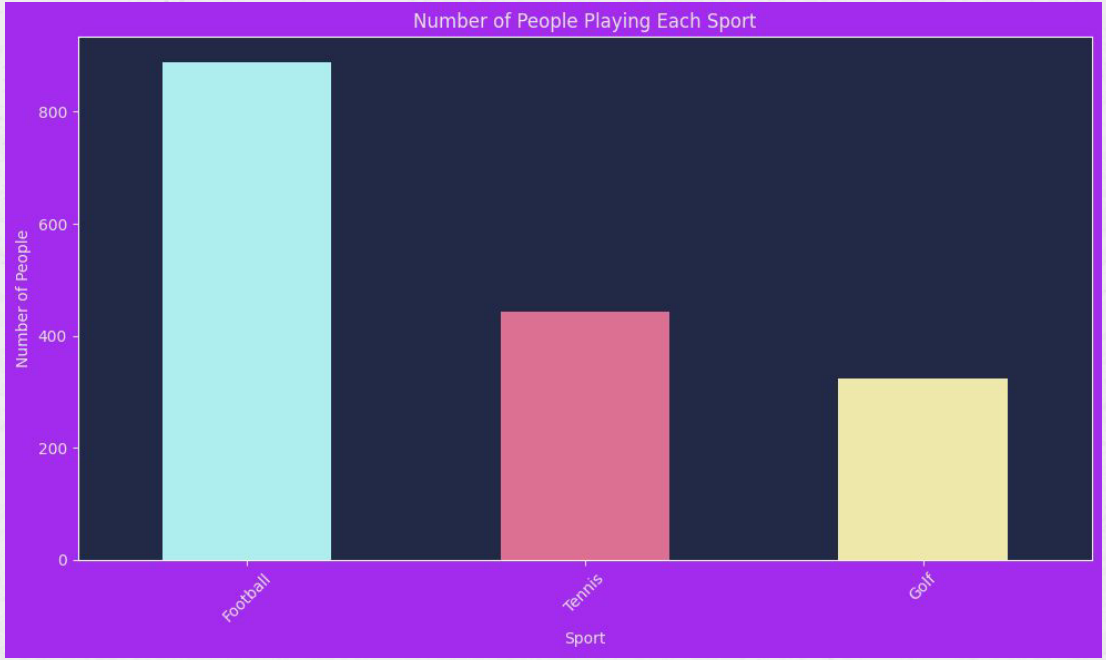
2023

Group 14

OUR TEAM

COMPANY AND DATASET

- Company B;
- Football (~800), Tennis (~400) and Golf (~300);
- 1655 people who play either football, tennis or golf.



TECHNIQUES USED

Decision Trees
Logistic Regression
k-NN

Our Motivation:

- Existing experience with k-NN;
- Simplicity and ease of implementation;
- It's very versatile, and it's not difficult to tune all the hyperparameters.

VALIDATION APPROACH

Data Splitting:

We divided the data into testing (30%) and training (70%) sets for each combination of features. This makes sure that our model is assessed using hypothetical data, giving us a more accurate picture of how well it performs.

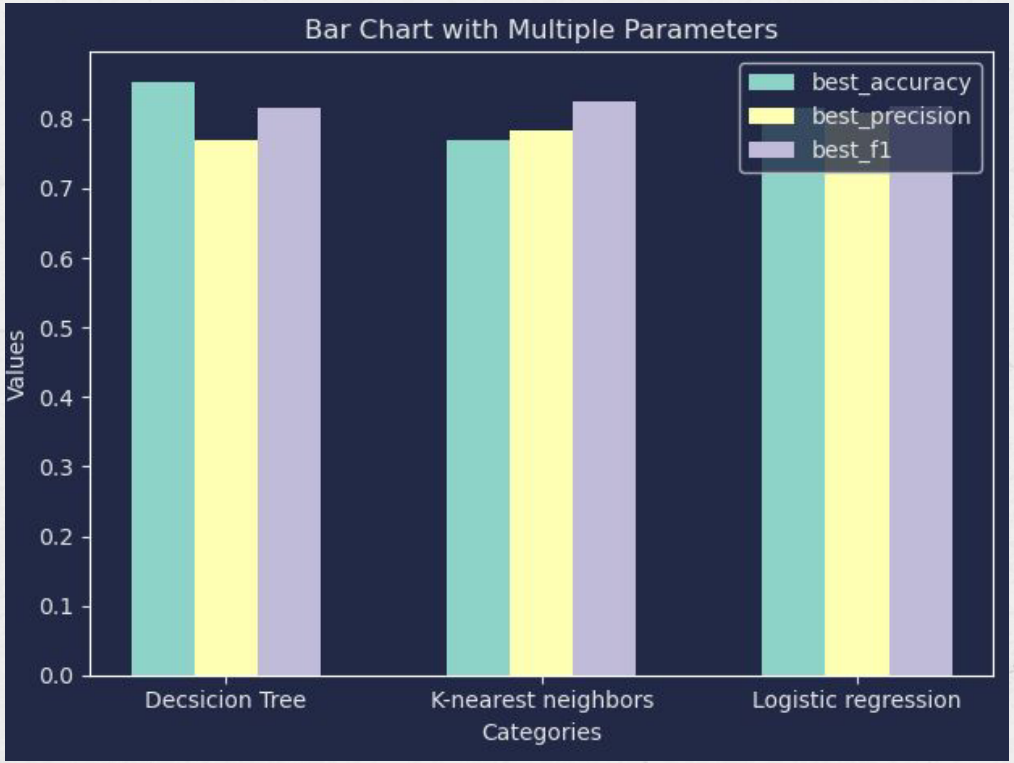
We decided to judge final performance of the models by F1 metric. The metric balance off precision and recall, since both, false negatives and false positives are important to the company.

INPUTS USED

Inputs choice:

We went through all the possible combinations of inputs for each model (but no more than 4, due to costs limits). Also we removed some features, since we assume that some of them are unethical(age, nationality), some of them are not important for the result (sport), and some of them are target features (decision).

MODELS COMPARISON

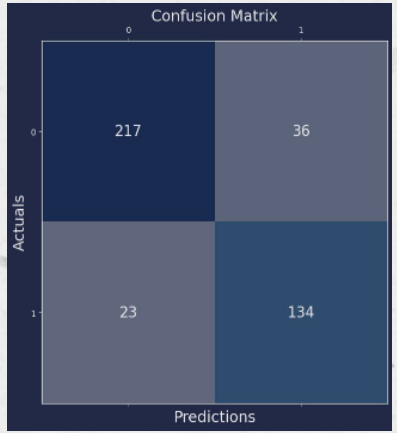


MODELS COMPARISON

Decision Trees

F1 score: 0.825

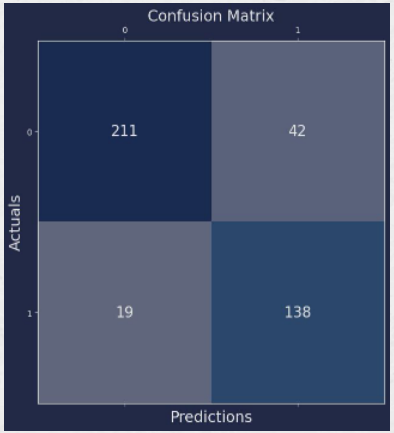
FP+TP = 59



Logistic Regression

F1 score: 0.816

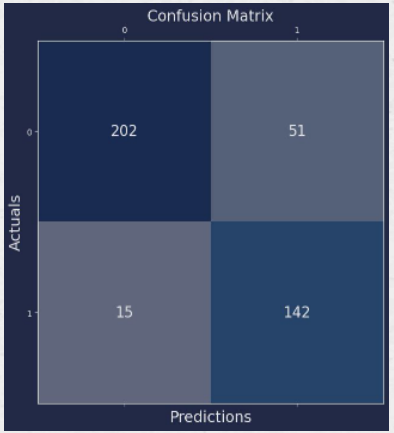
TP+FP = 61



k-NN

F1 score: 0.817

FP + TP = 66



2023

BIAS ANALYSIS

We chose decision tree according to the metrics. We tried to remove all the potential biases, but to avoid them in the future, several steps needed to be taken:

- **Feature selection**
  - Be sure that non of sensitive features are used for the model. Like age, gender or nationality.
- **Imbalanced data**
  - If some group is under-represented in the data, it potentially could cause discrimination towards them. Therefore pre-processing of the data is important.
- **Model complexity**
  - Make sure not to make the model too complex since it can overfit. Regular hyperparameter optimizing could help with it.

IMPLEMENTATION ADVICE

- **User training**
  - It's important to train potential users of the model so they could understand how to use model properly.
- **Keep all the documentation**
  - Keeping all the changes of the model could be useful in the future to analyze it.
- **Involve experts**
  - Involve all the stakeholders through the process.
- **Continuous Monitoring**
  - to identify any new patterns of bias and KPI control to be sure in the model success. Regular feedback also could be useful.