U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Prever Resultados de Futebol

Aprendizagem Supervisionada - **Classificação**

João Abelha - up201706412

João Varela - up201706072

Vítor Barbosa - up201703591

Inteligência Artificial

Definição do Problema

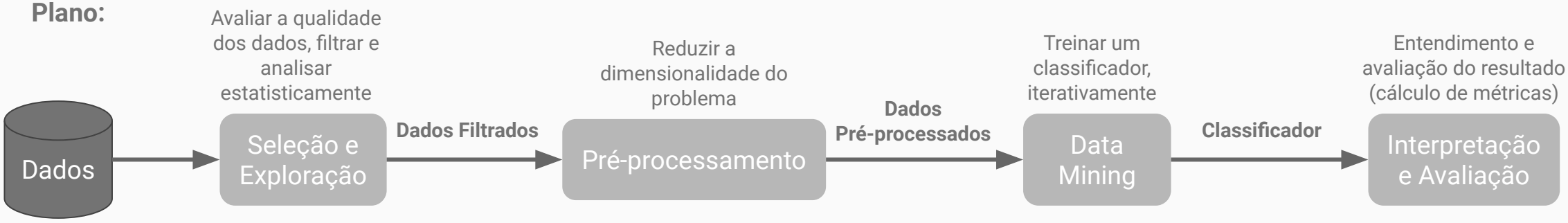
Objetivo:

- Prever o resultado de jogos de futebol, construindo classificadores que calculam a probabilidade do resultado de cada jogo (vitória, empate ou derrota).
- Comparar o desempenho de diferentes algoritmos de aprendizagem e compreender a influência dos dados no mesmo.

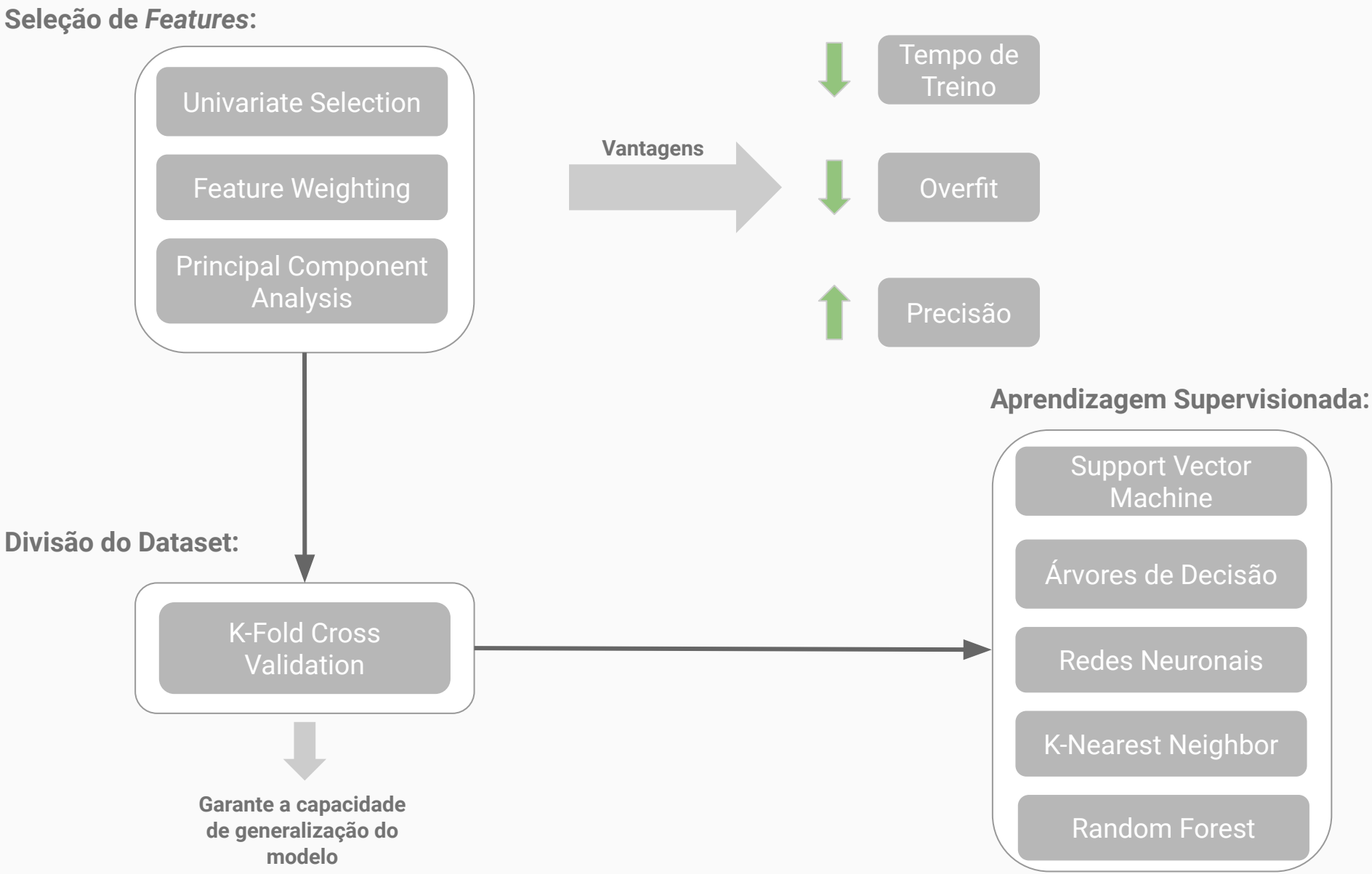
Dados Iniciais:

- Estatísticas retiradas de jogos de futebol.
- Atributos dos jogadores (retirado do jogo *FIFA*).
- *odds* de cada jogo (retiradas de 10 sites de apostas diferentes).

Plano:



Algoritmos e Técnicas



Ferramentas

Python - não só é uma linguagem com uma grande simplicidade e legibilidade, mas contém também poderosas frameworks e bibliotecas que tornam mais simples o desenvolvimento de projetos de aprendizagem supervisionada, cujos algoritmos associados são normalmente bastante complexos.

- **Scikit-learn -** extensa biblioteca com um enorme leque de algoritmos úteis à aprendizagem computacional.
- **Keras** - biblioteca dedicada a redes neuronais.
- **Pandas -** biblioteca usada para análise de dados.
- **Numpy -** biblioteca usada para análise de dados.
- **Matplotlib -** biblioteca para visualização de dados.

Jupyter Notebooks - ambiente de desenvolvimento ideal para o desenvolvimento de projetos, permitindo um *workflow* rápido graças às suas funcionalidades que permitem intercalar código com visualização de dados e correr rapidamente *snippets* de código.

Seleção e Exploração

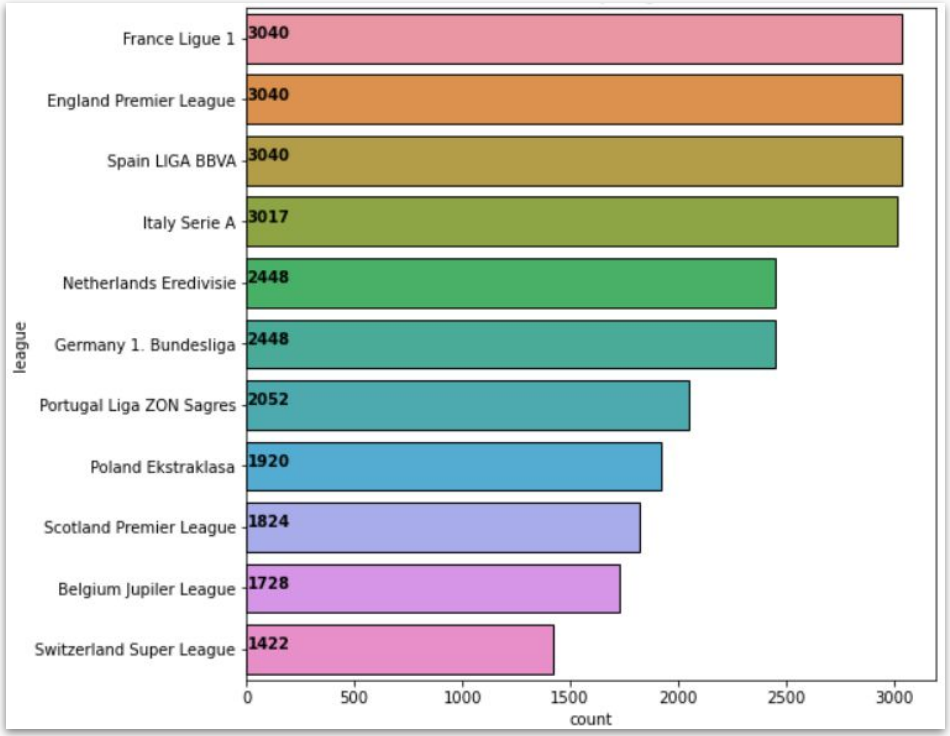


Fig 1 - **Número de Jogos por Liga**

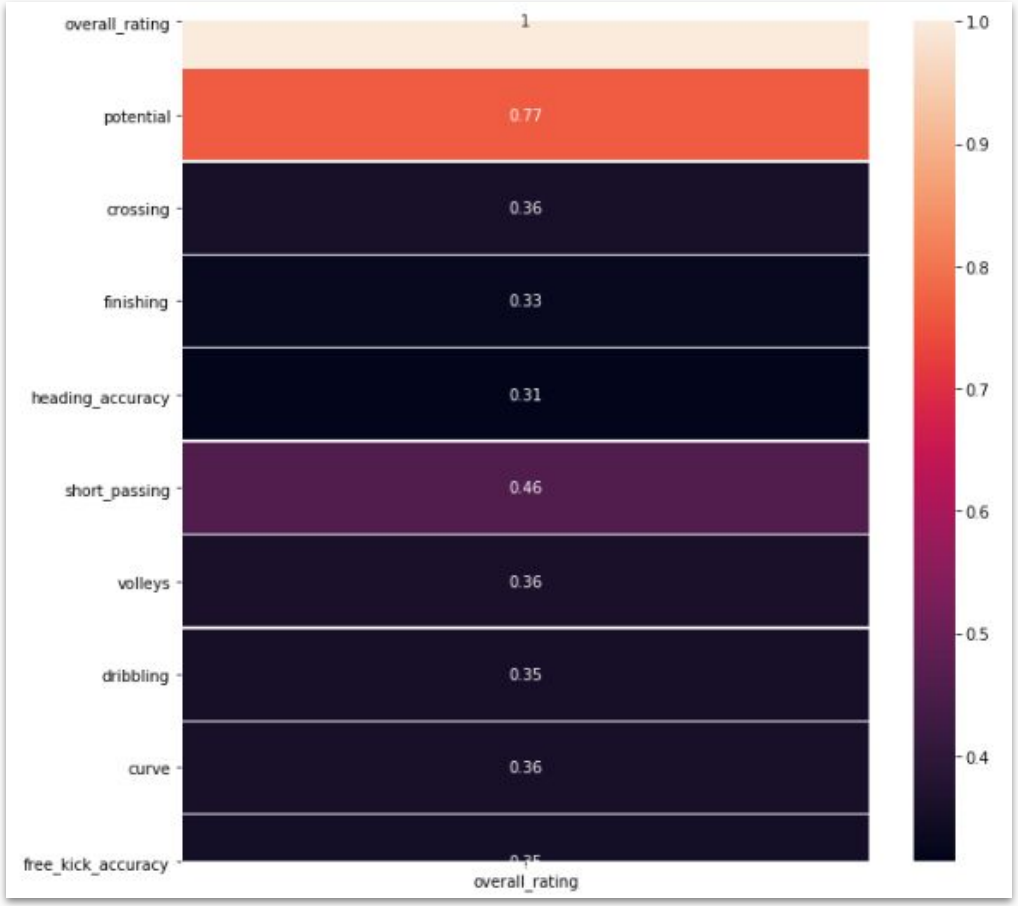
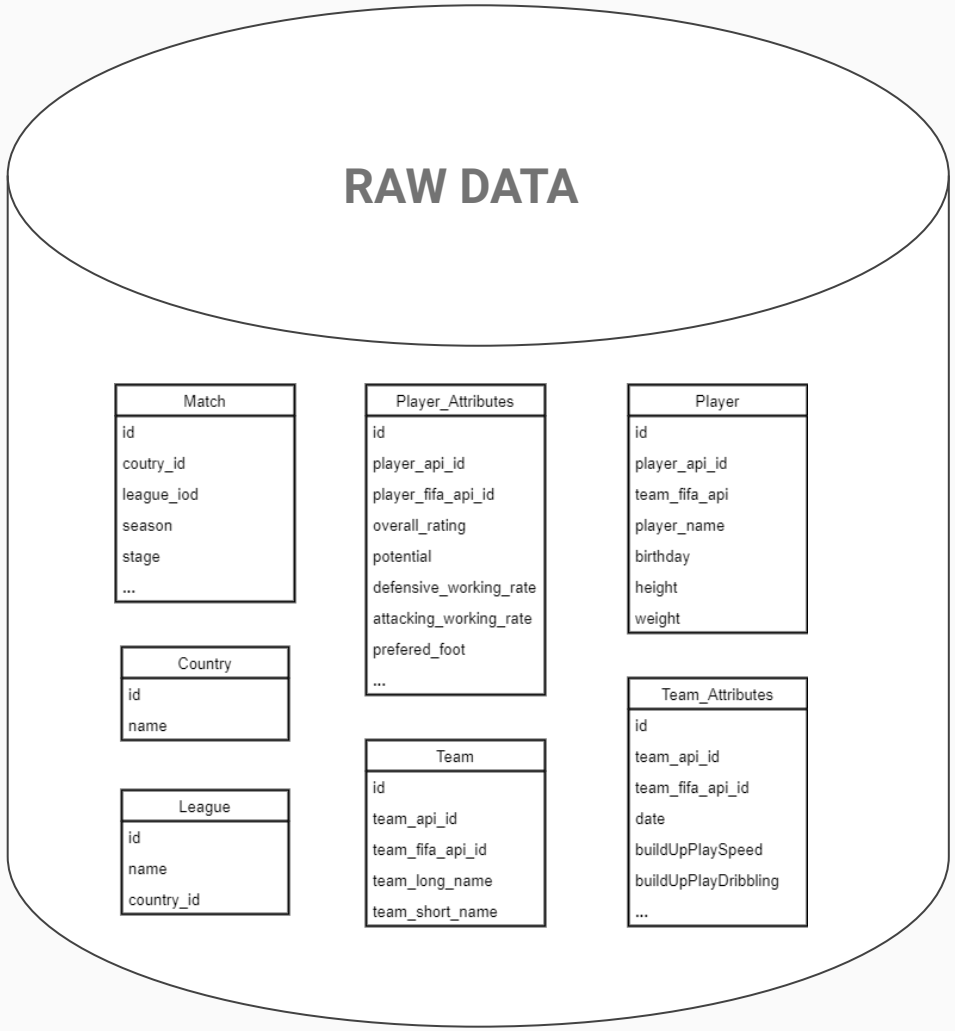


Fig 2 - **Peso dos Atributos no Overall dos Jogadores**

Pre-processing: Data Aggregation



Match
home_team_goals_difference
away_team_goals_difference
games_won_home_team
games_won_away_team
games_against_home
games_against_away
home_player_1_overall_rating
home_player_2_overall_rating
...
away_player_1_overall_rating
away_player_2_overall_rating
...
odds_win
odds_draw
odds_lose
label

Pre-processing: Dimensionality Reduction

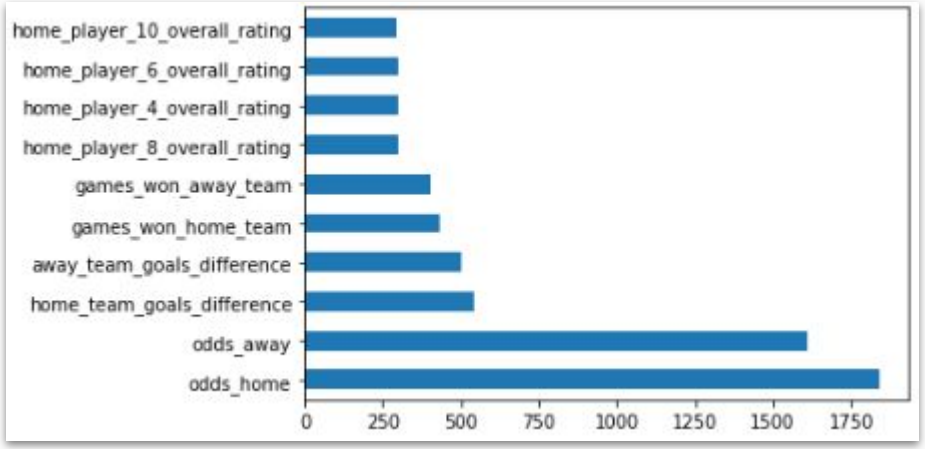


Fig 3 - **Univariate Selection**

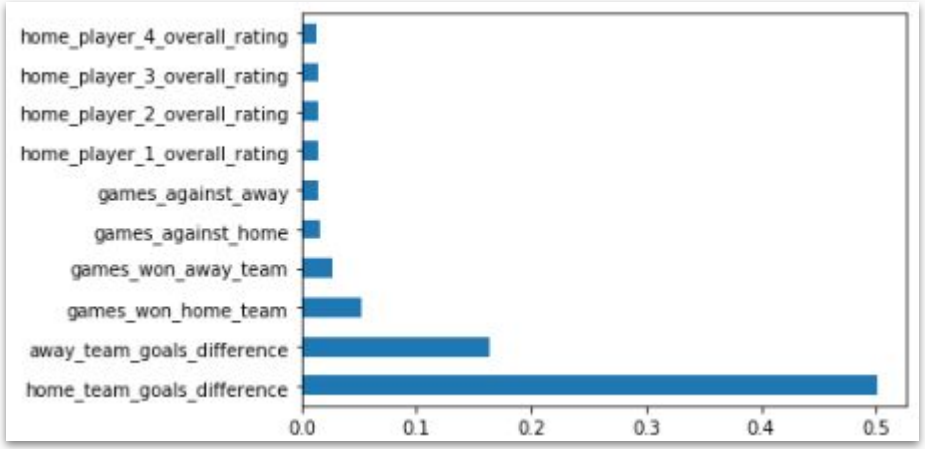


Fig 4 - **Principal Component Analysis**

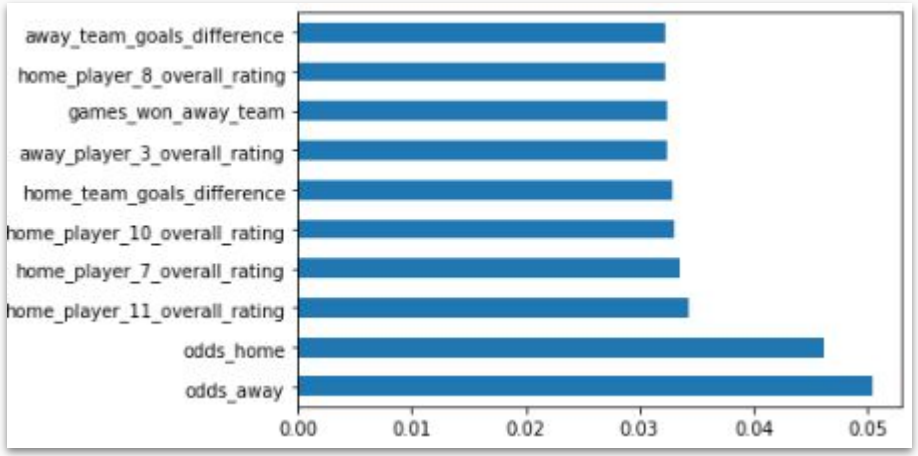


Fig 5 - **Feature Weighting**

Juntando os overalls de cada jogador

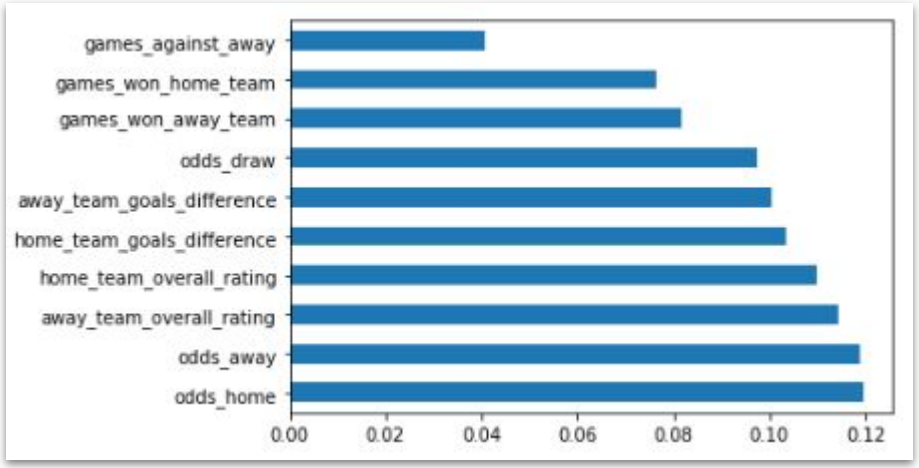
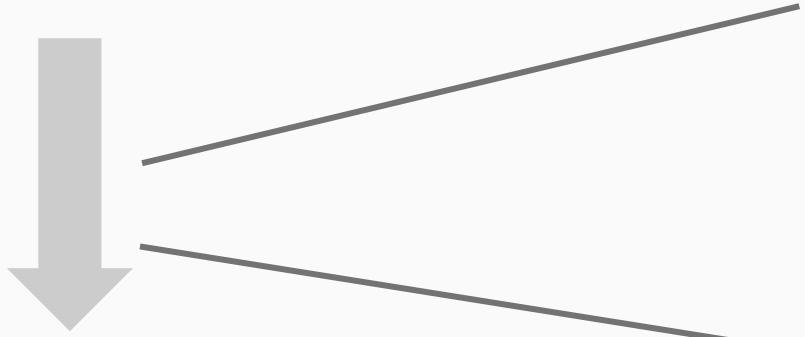


Fig 6 - **Feature Weighting with Team's Overall**

Pre-processing: Data Normalization

home_player_1_overall_rating	home_player_2_overall_rating	home_player_3_overall_rating
58.0	57.0	67.0
64.0	64.0	63.0
67.0	72.0	69.0
58.0	57.0	67.0
61.0	66.0	61.0

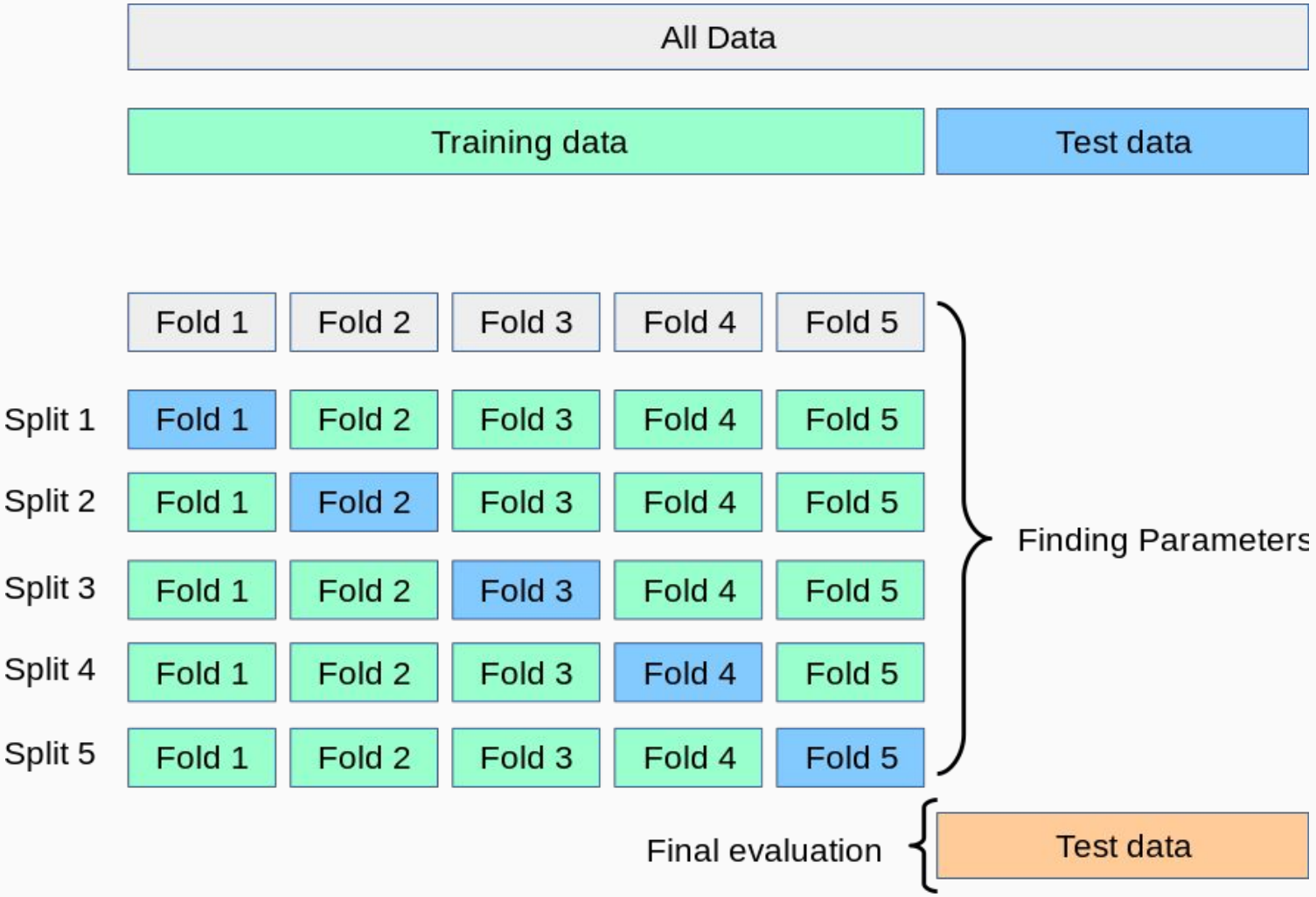


```
def normalize(df):
    result = df.copy()
    for feature_name in df.columns:
        max_value = df[feature_name].max()
        min_value = df[feature_name].min()
        result[feature_name] =
        (df[feature_name] - min_value) /
            (max_value - min_value)
    return result
```

home_player_1_overall_rating	home_player_2_overall_rating	home_player_3_overall_rating
0.3125	0.319149	0.500000
0.4375	0.468085	0.409091
0.5000	0.638298	0.545455
0.3125	0.319149	0.500000
0.3750	0.510638	0.363636

Data Mining

K-Fold Cross Validation:



Algoritmos Utilizados:

- Support Vector Machine
- Árvores de Decisão
- Redes Neuronais
- K-Nearest Neighbor
- Random Forest

Decision Tree

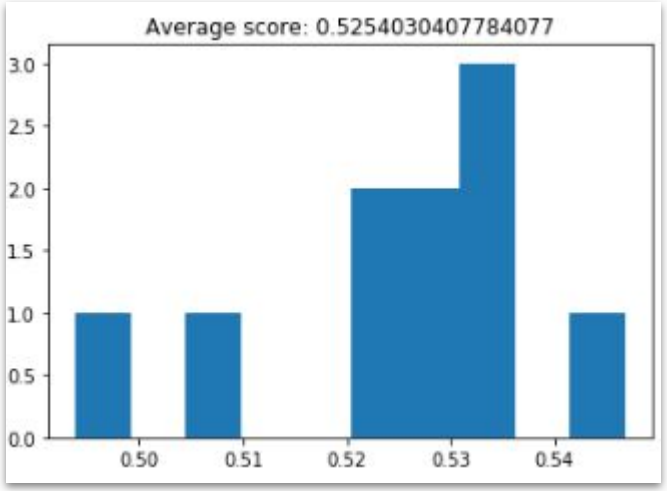


Fig 6 - **Decision Tree Accuracy**

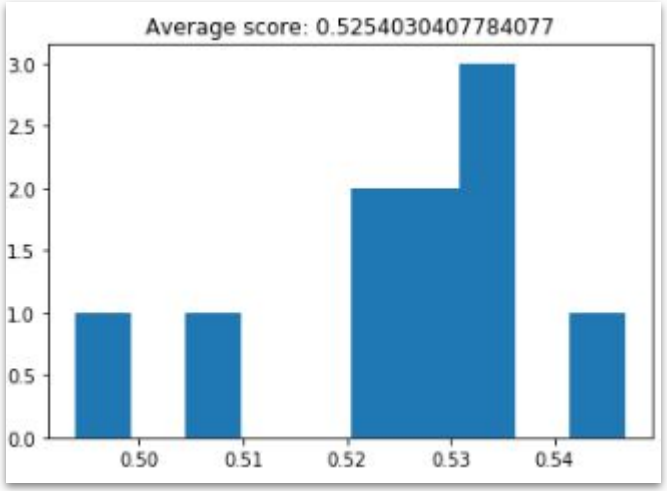


Fig 7 - **Decision Tree Accuracy** (“Minified dataset)

Decision Tree	
Accuracy	0.52541
Precision	0.36483
Recall	0.43538
F1-Score	0.37805
Loss	1.0333
Time (s)	31.0

Tab 1 - **Decision Tree Metrics**

Random Forest

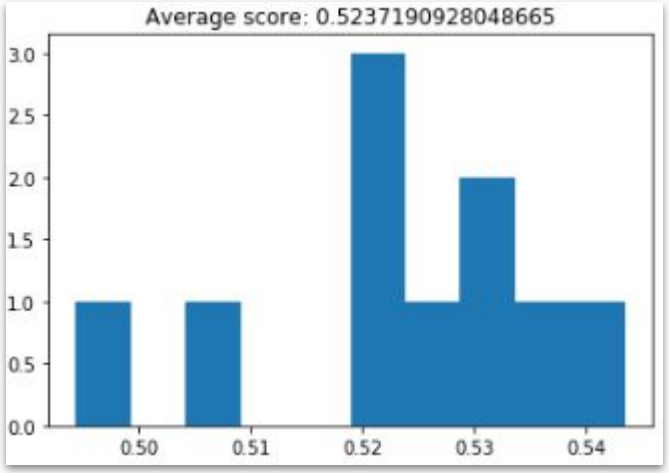


Fig 8 - **Random Forest Accuracy**

Random Forest	
Accuracy	0.52479
Precision	0.42584
Recall	0.43116
F1-Score	0.41177
Loss	1.05146
Time (s)	878.8

Tab 2 - **Random Forest Metrics**

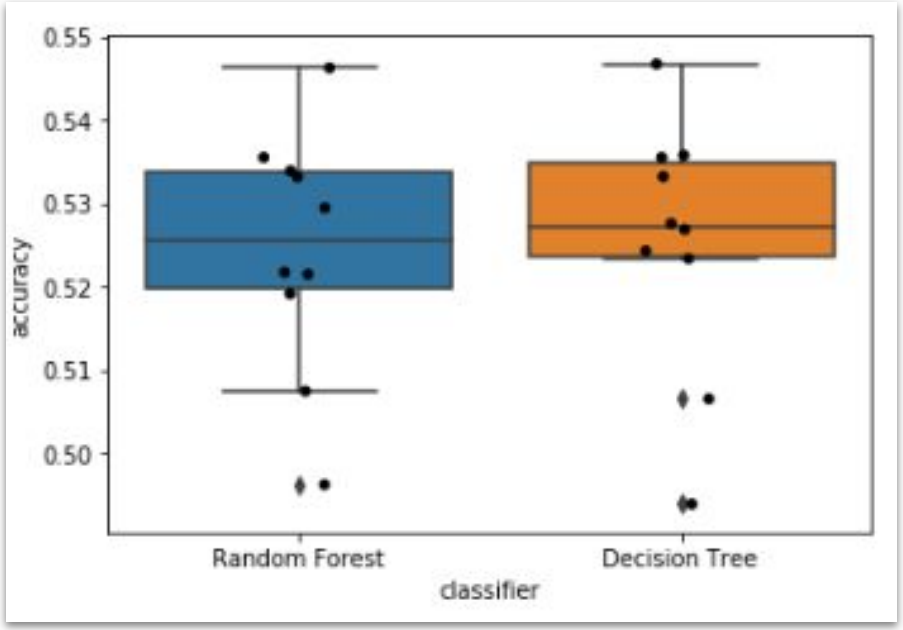


Fig 9 - **Random Forest vs Decision Tree**

K Nearest Neighbors and Support Vector Machine

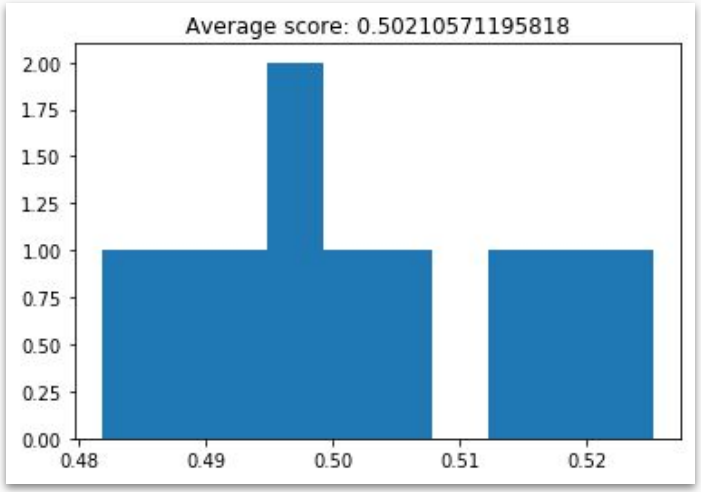


Fig 10 - **KNN Accuracy**

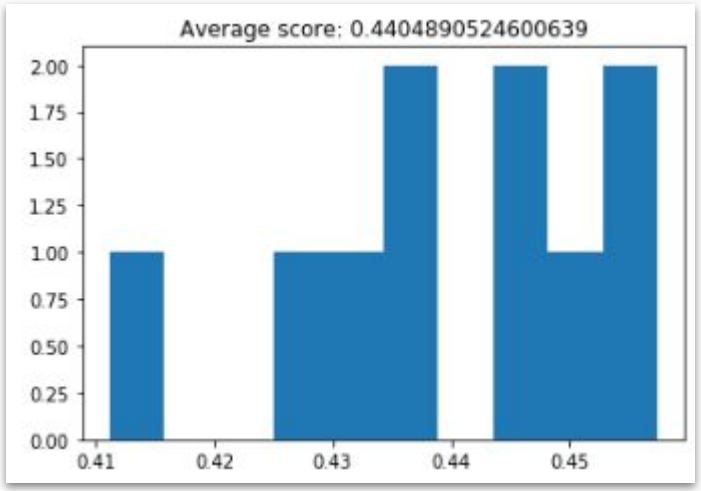


Fig 11 - **SVM Accuracy**

KNN	
Accuracy	0.50210
Precision	0.4181
Recall	0.4282
F1-Score	0.40796
Loss	1.0568
Time (s)	876.8

Tab 3 - **KNN Metrics**

SVM	
Accuracy	0.46505
Precision	0.39528
Recall	0.41329
F1-Score	0.38881
Loss	1.02189
Time (s)	2001.3

Tab 4 - **SVM Metrics**

Neural Network

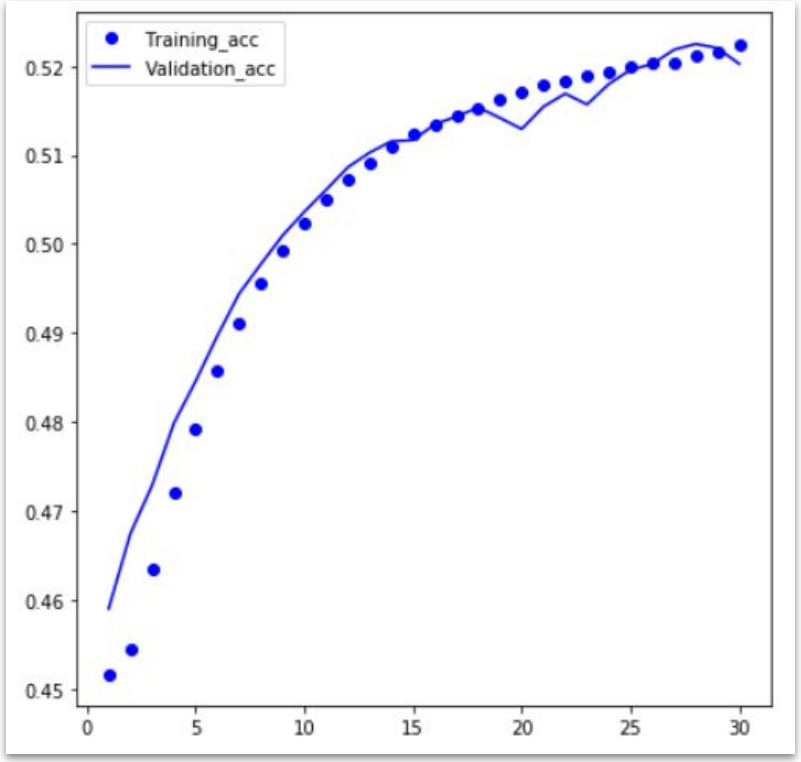


Fig 12 - **Neural Network Training**

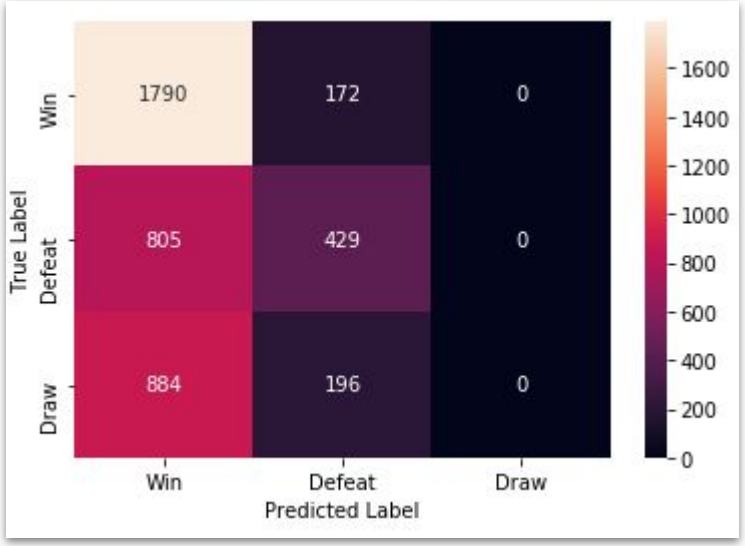


Fig 13 - **Neural Network Confusion Matrix**

Neural Network	
Accuracy	0.521983
Precision	0.59201
Recall	0.2678
F1-Score	0.2485
Loss	1.00876
Time (s)	31.3

Tab 5 - **Neural Network Metrics**

Classifiers Comparison

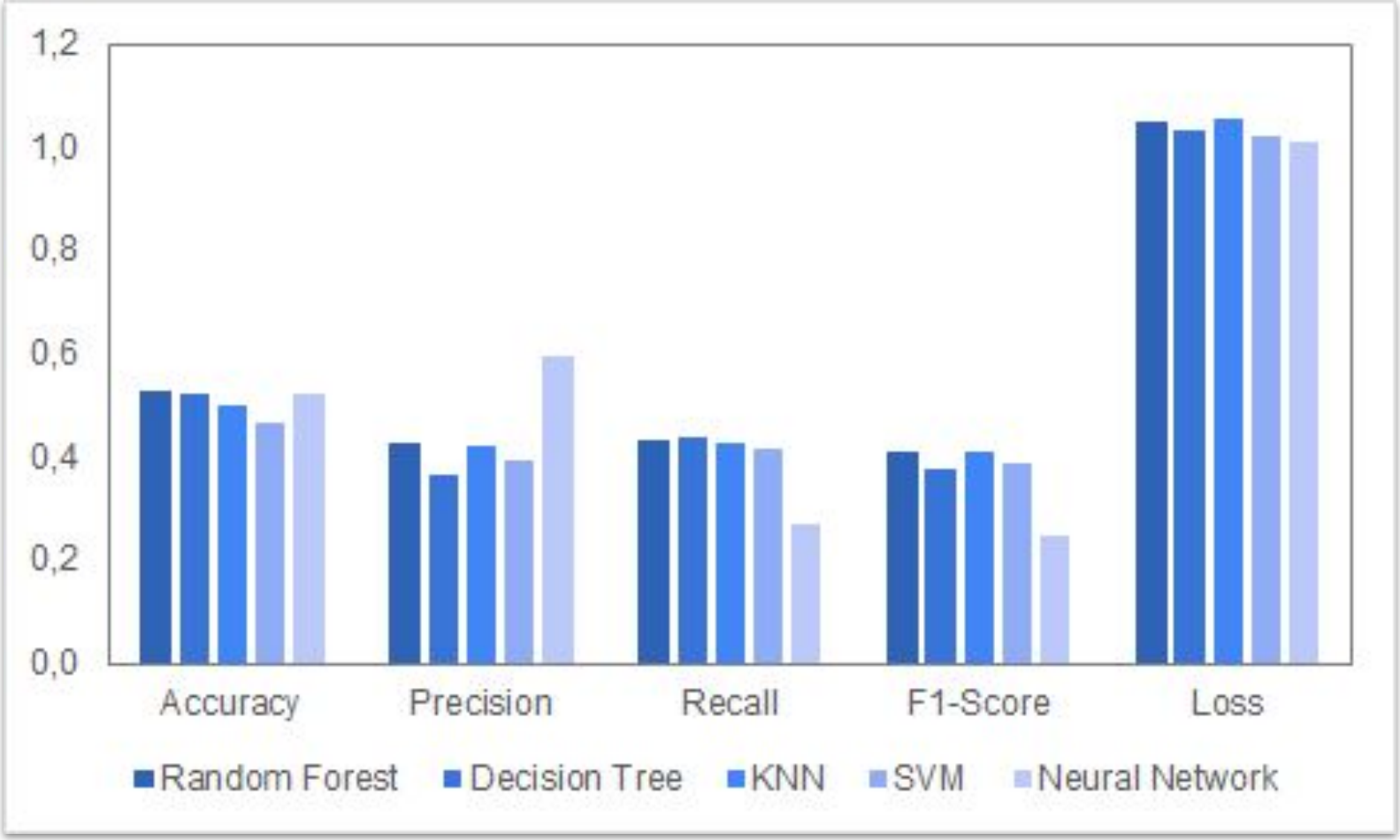


Fig 14 - **Classifier Metrics Comparison**

Odds Predictions

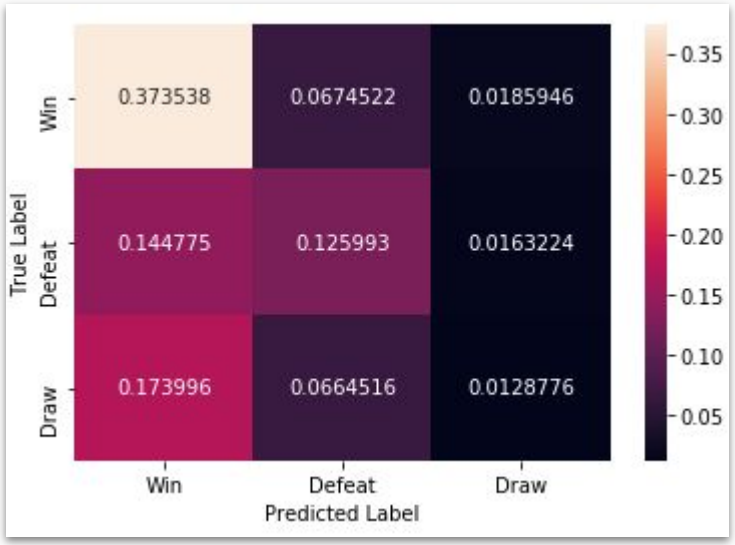


Fig 15 - **Odds Confusion Matrix**

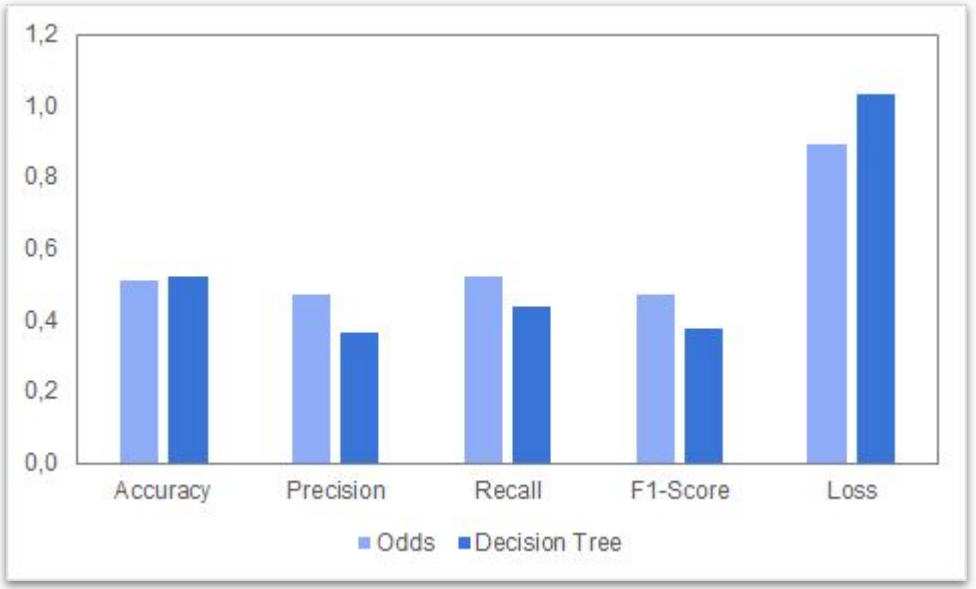


Fig 16 - **Odds vs Decision Tree**

Odds	
Accuracy	0.5124
Precision	0.4738
Recall	0.5223
F1-Score	0.4719
Loss	0.8912

Tab 6 - **Odds Predictions Metrics**

Conclusões

- O projecto foi desenvolvido com sucesso e de uma forma robusta.
- Comparando os diferentes algoritmos de aprendizagem supervisionada, o Decision Tree resultou na melhor performance.
- A replicabilidade e robustez dos resultados foi assegurada graças ao K-Cross Fold Validation.
- A abordagem realizada pode ser facilmente aplicada a outros tipos de problemas de classificação com sucesso, dado termos seguido cuidadosamente uma pipeline que tenta derrubar problemas que surgem na aprendizagem supervisionada.
- Atingiu-se uma *accuracy* acima dos 50% o que é relativamente alto, tendo em conta o número de fatores que afetam o resultado de um jogo de futebol.
- Compreensão e consolidação de vários conteúdos relacionados com aprendizagem supervisionada, bem como aquisição de conhecimentos sobre as ferramentas que o Python fornece ligadas à aprendizagem computacional.

Trabalho Futuro

- Testar a criação de features diferentes (p.e. diferença de golos na época em questão)
- Incluir outros datasets sem ser do jogo FIFA.
- Investigar mais técnicas e algoritmos de aprendizagem supervisionada.

Referências

- **A machine learning framework for sport result prediction** - https://www.sciencedirect.com/science/article/pii/S2210832717301485
- **Feature Selection Techniques** - https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
- **Classification with Keras** - https://www.pluralsight.com/guides/classification-keras
- **Working with Categorical Data** - https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63
- **Machine Learning Notebook**

  https://github.com/rhiever/Data-Analysis-and-Machine-Learning-Projects/blob/master/example-data-science-notebook/Example%20Machine%20Learning%20Notebook.ipynb
- **Dataset Usado** - https://www.kaggle.com/hugomathien/soccer

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Prever Resultados de Futebol

Aprendizagem Supervisionada - **Classificação**

João Abelha - up201706412

João Varela - up201706072

Vítor Barbosa - up201703591

Inteligência Artificial