

***Faculty of Science and Technology***

**Assignment Coversheet**

|  |  |
| --- | --- |
| **Student ID number &**  **Student Name** | U3239017 Sachleen Nahl |
| **Unit name** | Software Technology 1 |
| **Unit number** | 4483 |
| **Unit Tutor** | Pranav Gupta |
| **Assignment name** | ST1 Capstone Project – Semester 1 2023 |
| **Due date** | 12/05/2023 |
| **Date submitted** |  |

**You must keep a photocopy or electronic copy of your assignment.**

**Student declaration**

I certify that the attached assignment is my own work. Material drawn from other sources has been appropriately and fully acknowledged as to author/creator, source and other bibliographic details.

**Signature of student: SN Date: 10/06/2023**

Table of Contents

[Introduction 3](#_Toc143249246)

[Methodology 4](#_Toc143249247)

[Design and Development – Algorithm Design Stage 5](#_Toc143249248)

[Algorithm Implementation and Software Development Stage 18](#_Toc143249249)

[Conclusion 18](#_Toc143249250)

[References: 19](#_Toc143249251)

# Introduction

This report analyses the 2019 Women’s World Cup Matches dataset, created by FiveThirtyEight. Data science has been developing quickly, with its methodology and guiding ideas being used in a variety of fields. Sports analytics is one of these intriguing fields. This research explores the 2019 Women's World Cup match dataset and does a prediction analysis using FiveThirtyEight's Soccer Power Index (SPI) ratings and possibilities of advancement for each country. Based on the outcomes of each team's most recent matches, the SPI ratings provide an estimation of team strength. This prediction was made by FiveThirtyEight using 20,000 simulations of the competition, which provided a large and reliable dataset for our research. A thorough exploratory data analysis (EDA) was conducted for the report, which contains statistical summaries, data visualisations, and trends. After the EDA, predictive analysis utilising machine learning techniques is carried out to foretell the results of games between two teams. The report's prediction model is implemented using Streamlit, a well-liked open-source app framework that data scientists use to build interactive apps.

# Methodology

1. Design and Development

The three steps of the approach utilised to create the software platform are as follows:

1. Algorithm design and development for decision assistance based on exploratory data analysis and predictive analytics, in order to determine the most effective method for resolving a practical issue.
2. Development of a desktop Tkinter software tool based on the highest performing algorithm.
3. The tool's deployment as a web- or cloud-based platform utility.
4. Implementation
5. Deployment

# Design and Development – Algorithm Design Stage

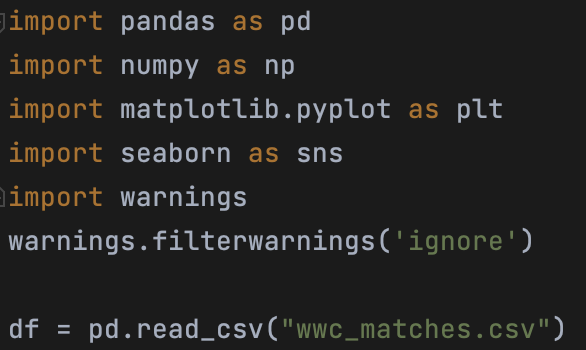


Dataset Description

"2019 Women's World Cup Predictions" is the name of the dataset, which was produced by FiveThirtyEight. It includes information on the odds of each side winning each match in the 2019 FIFA Women's World Cup as well as the odds of a draw. The dataset contains elements like the forecast's timestamp, the two teams involved in a match, and the tournament's stage. A model that considered elements including team strength, recent performance, and the venue of the match produced the anticipated victory probability. The dataset contains the estimated victory probabilities for both teams as well as the tie probability. It comprises 64 rows (18 columns) and 64 rows (one for each match).

Exploratory Data Analysis

The interpretation of the data, basic exploratory data analysis, and data visualisation were all part of the initial phase of the software development activity. PyCharm was selected as the experimental environment because it has virtual resources and hardware that can be used without the need for extra physical hardware and can be run straight from a web browser. The scripts that ran directly on PyCharm were written in the Python programming language. The following Python script must be run in order to import the necessary Python libraries for exploratory data analysis and gather the necessary datasets.



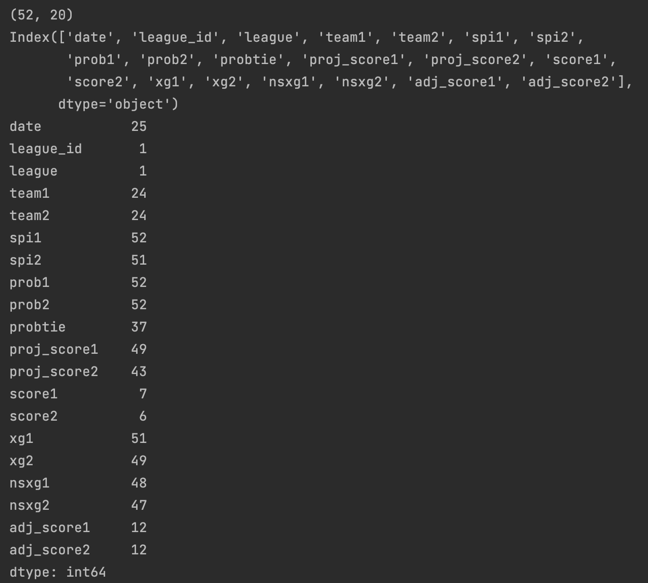
First and Last 5 rows

A picture containing text, screenshot, font, black

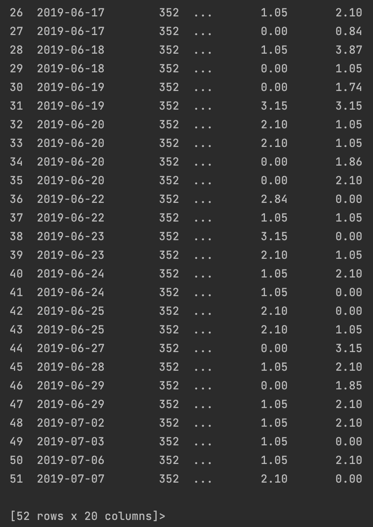
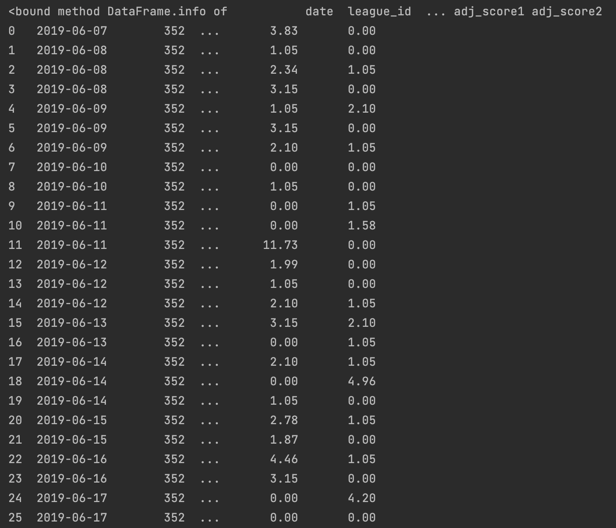
Description automatically generatedA picture containing text, screenshot, font, black

Description automatically generated

Data Shape – Rows, Columns, Attributes (Names and Values)



DataFrame Information



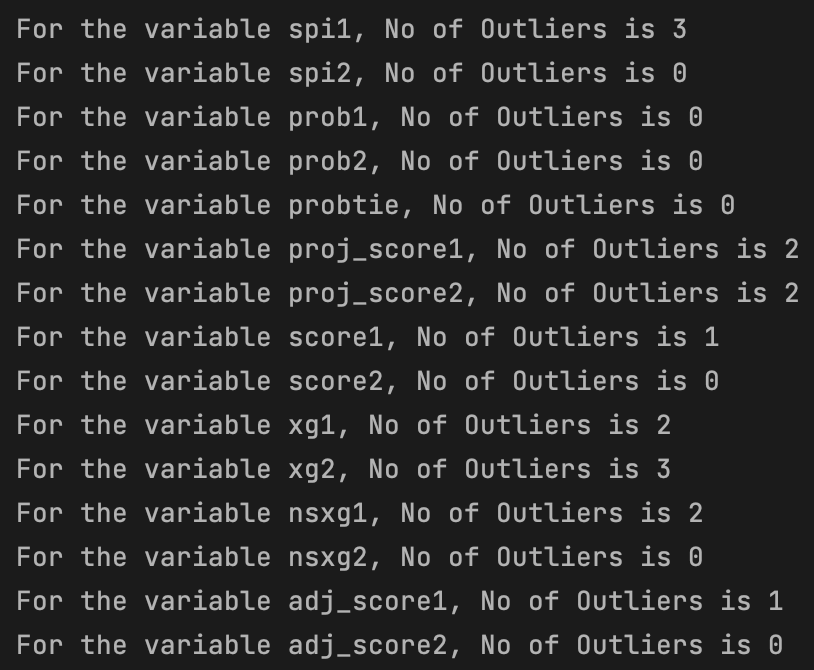
Data Visualisation

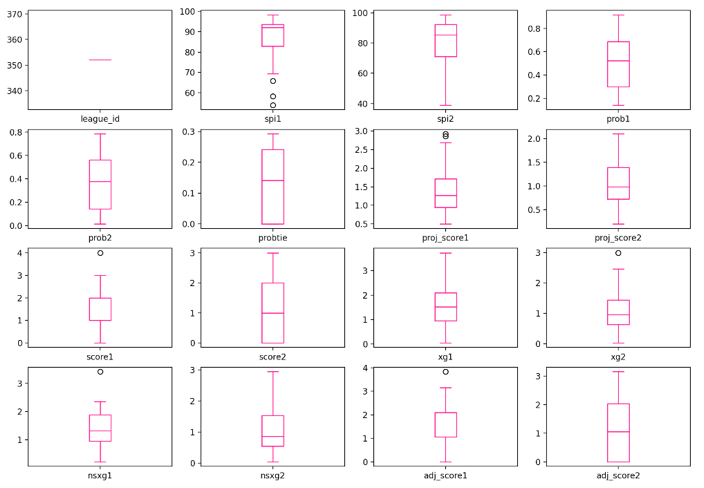
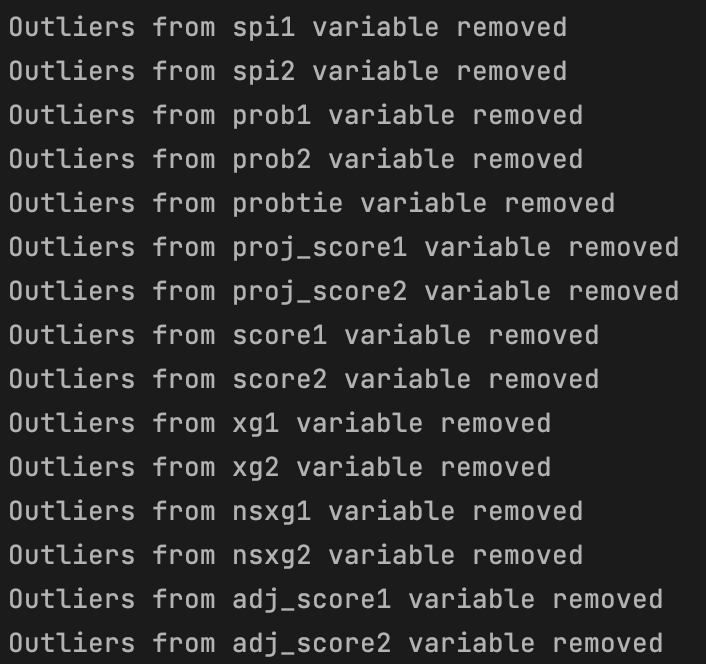
A picture containing text, diagram, plan, plot

Description automatically generated

Outlier Detection, Definition, Removal + Confirmation, DataFrame Shape following removal

A picture containing text, diagram, plan, schematic

Description automatically generated



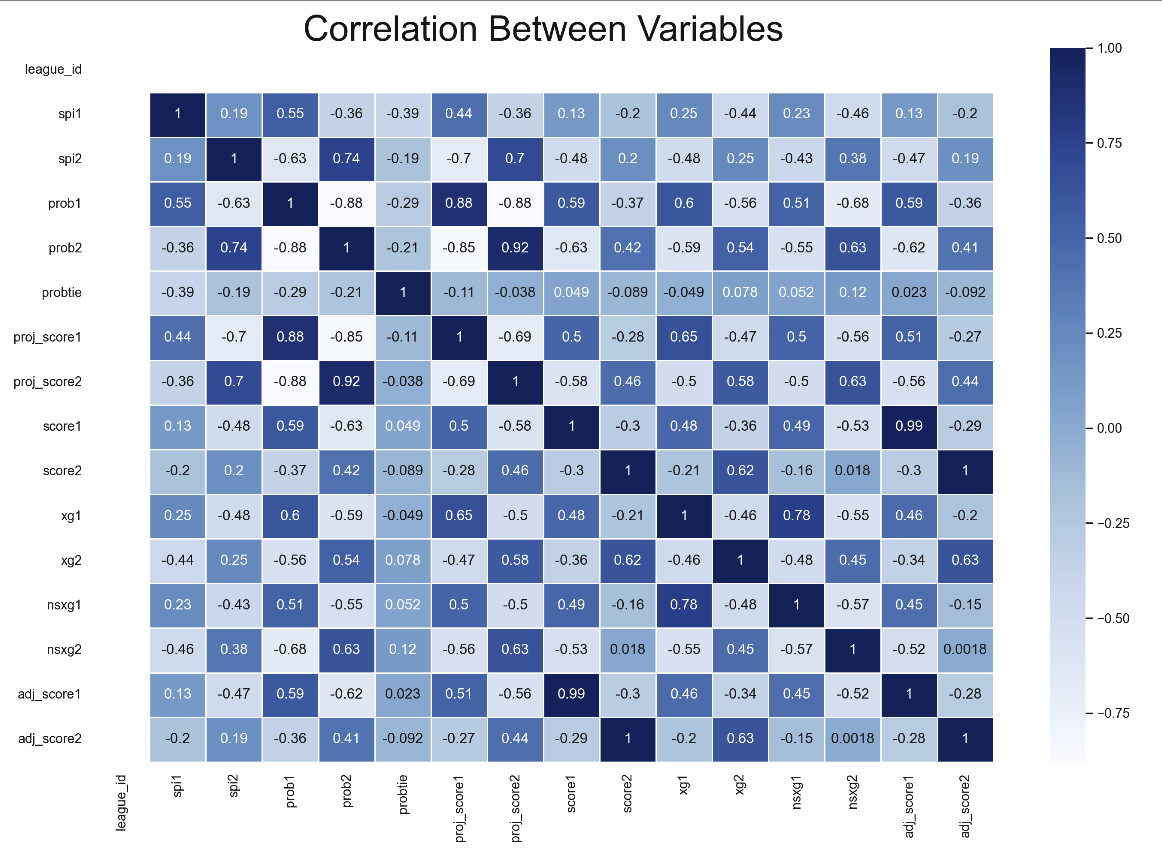


Target Value Distribution

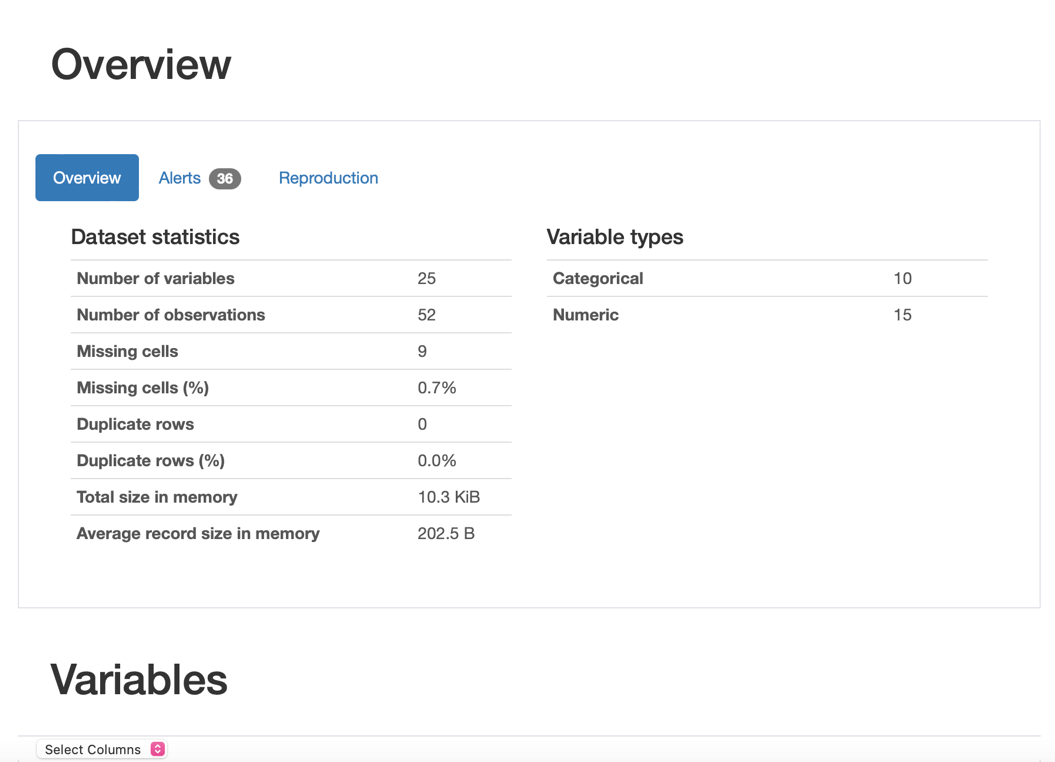
A picture containing text, screenshot, plot, number

Description automatically generated

Value Correlation



Full Profiler Report

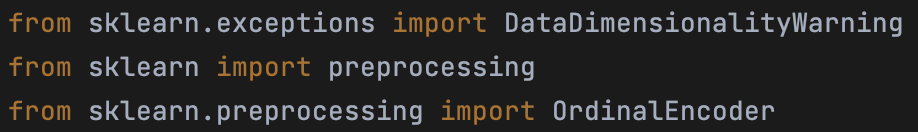


Predictive Data Analytics

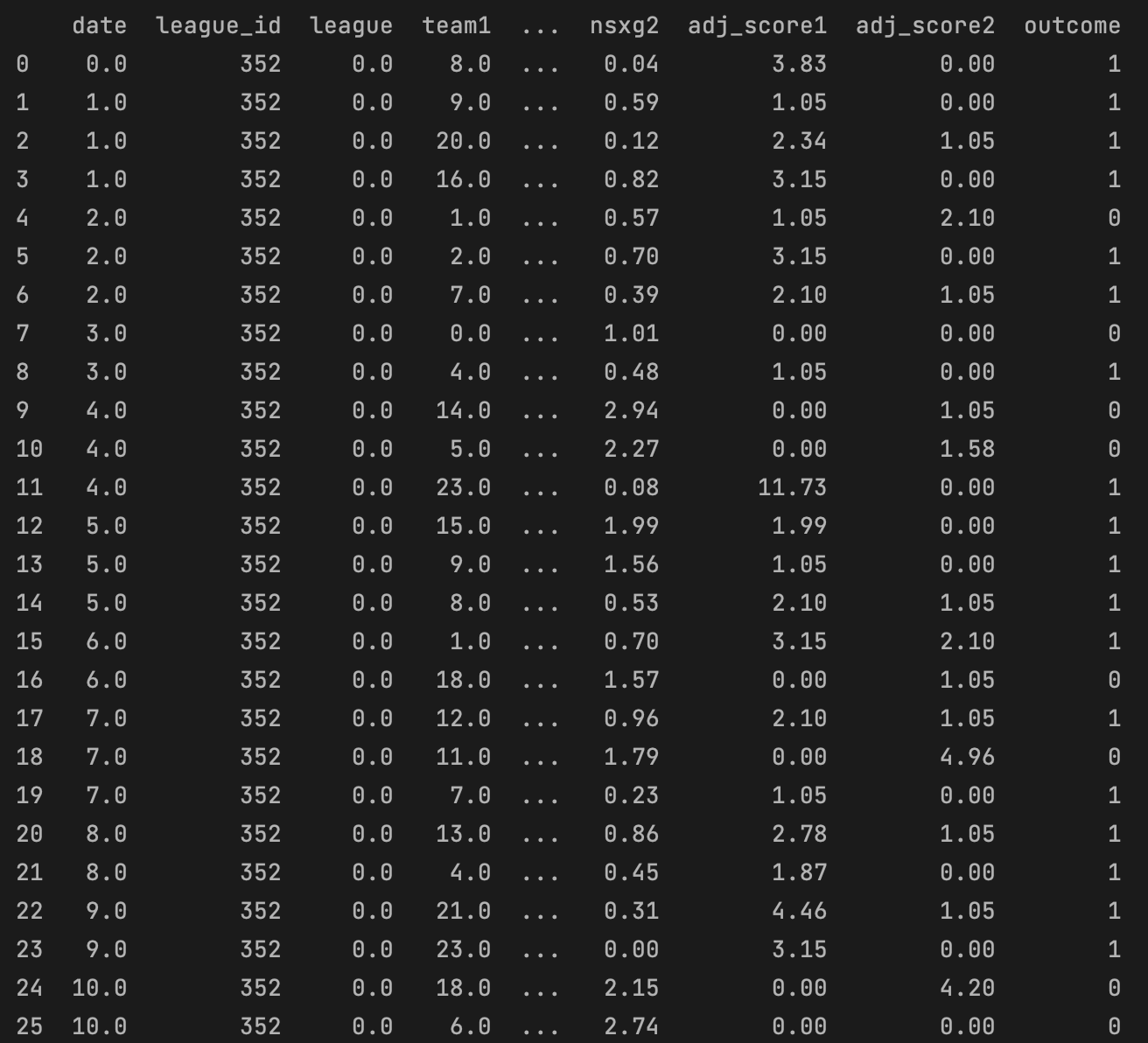
There are various processing stages needed for predictive analytics. These include pre-processing, classifier comparison to find the best machine learning classifier, and performance evaluation with various objective metrics. These were obtained using the Python scikit-learn package, including accuracy, classification report, confusion matrix, ROC-AUC curve, and prediction report. Next, each of these actions is explained.

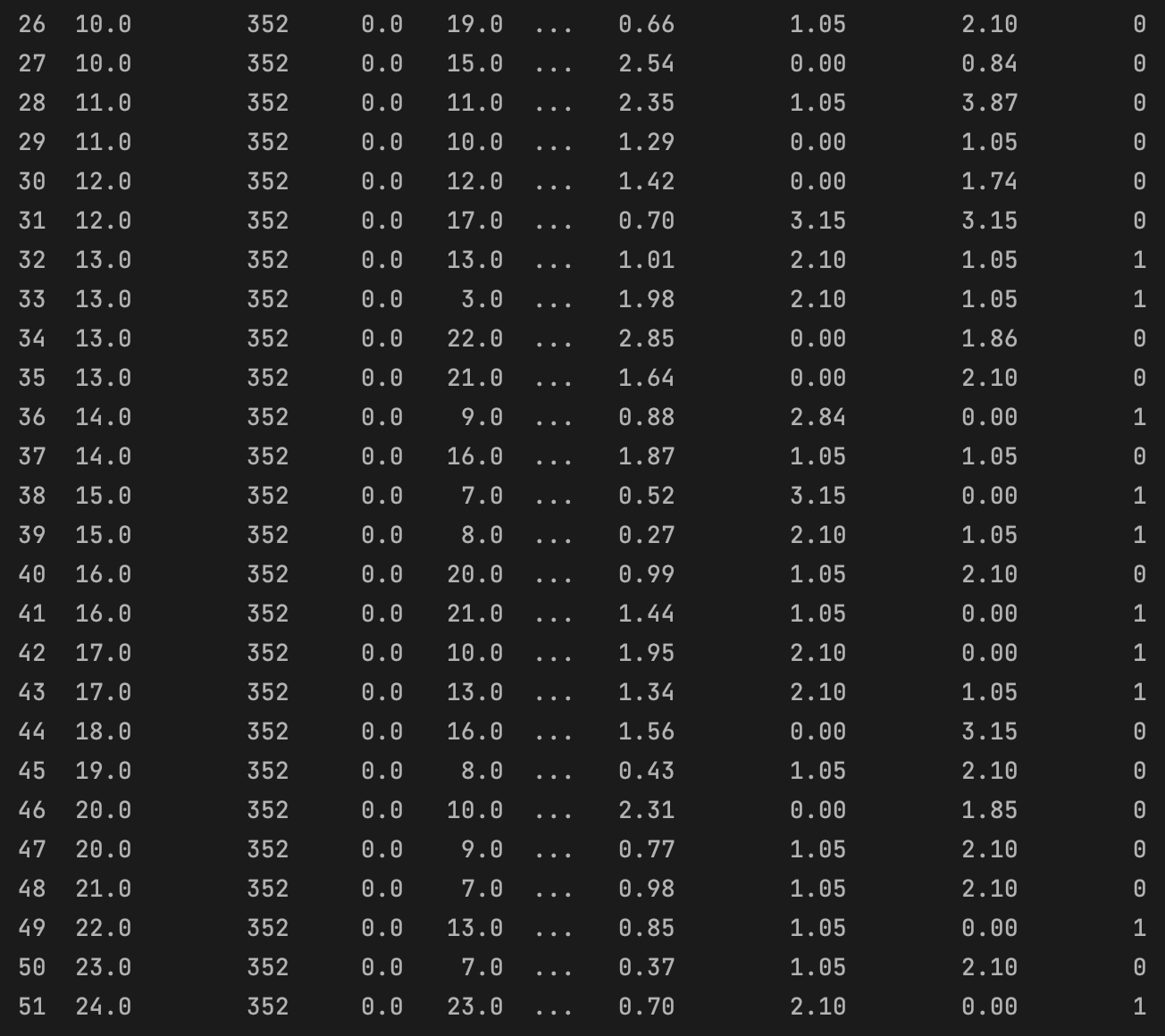
• Data pre-processing is necessary since the dataset combines continuous and categorical characteristics and variables. This requires attribute transformation, standardisation, and normalisation. To modify attribute values, we applied the OrdinalEncoder() tool from scikit-learn.

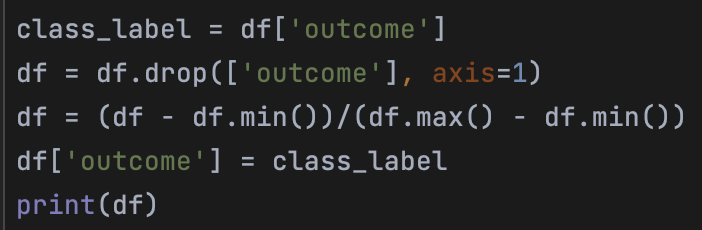
• The target was removed from the dataframe, normalised, and then reattached to the dataframe in order to normalise the independent values of the dataframe.

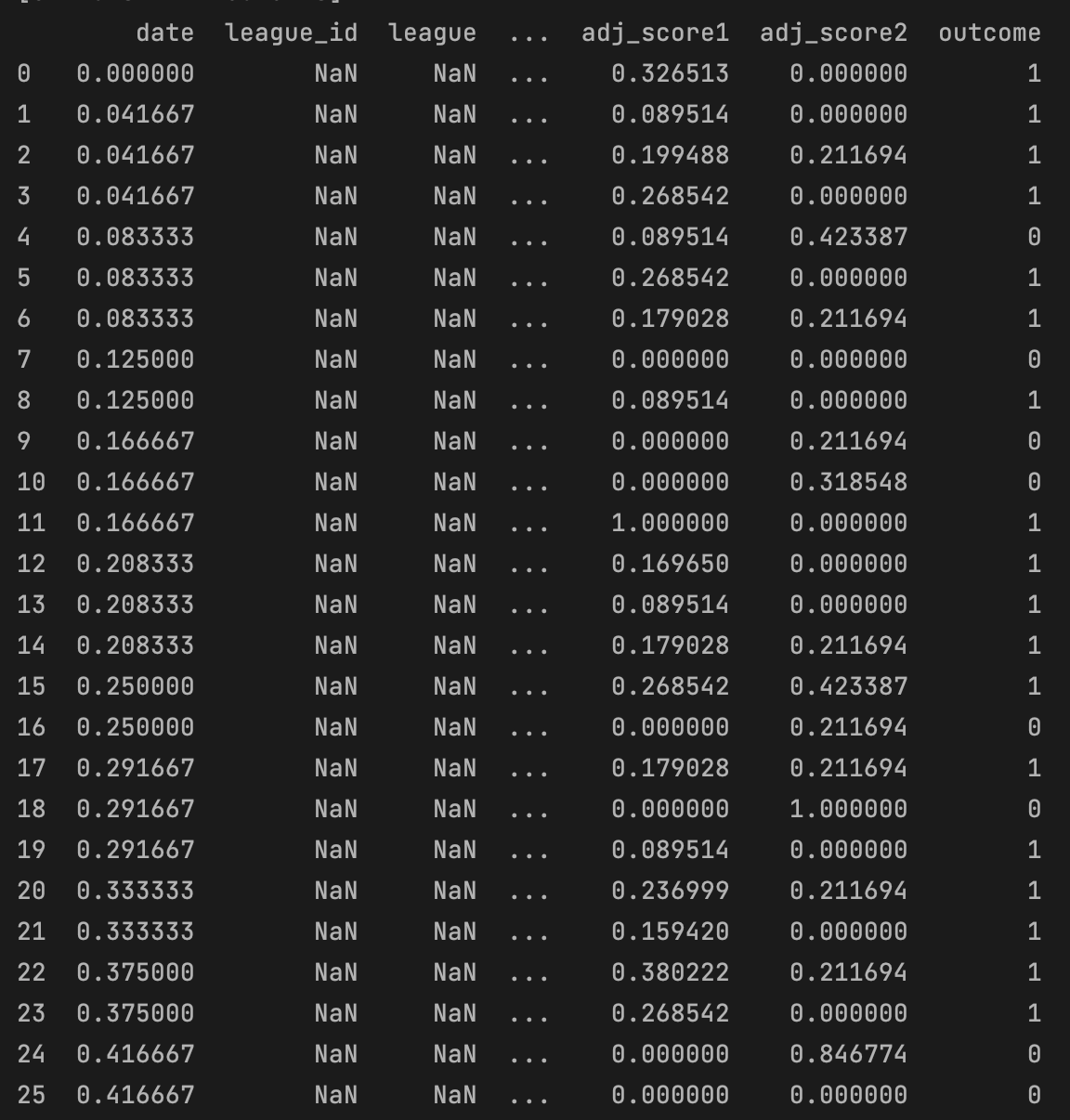
A picture containing text, screenshot, font

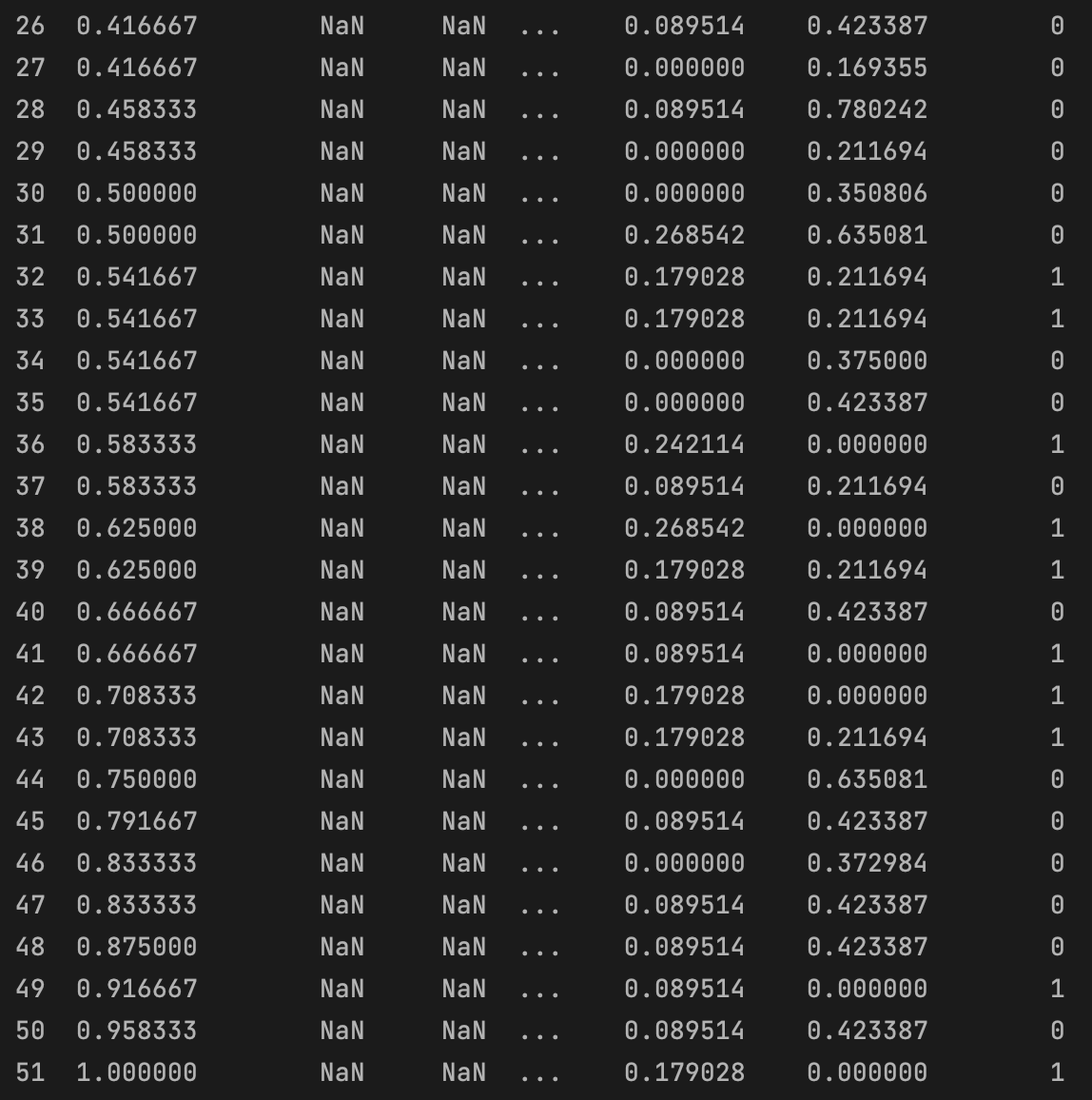
Description automatically generated

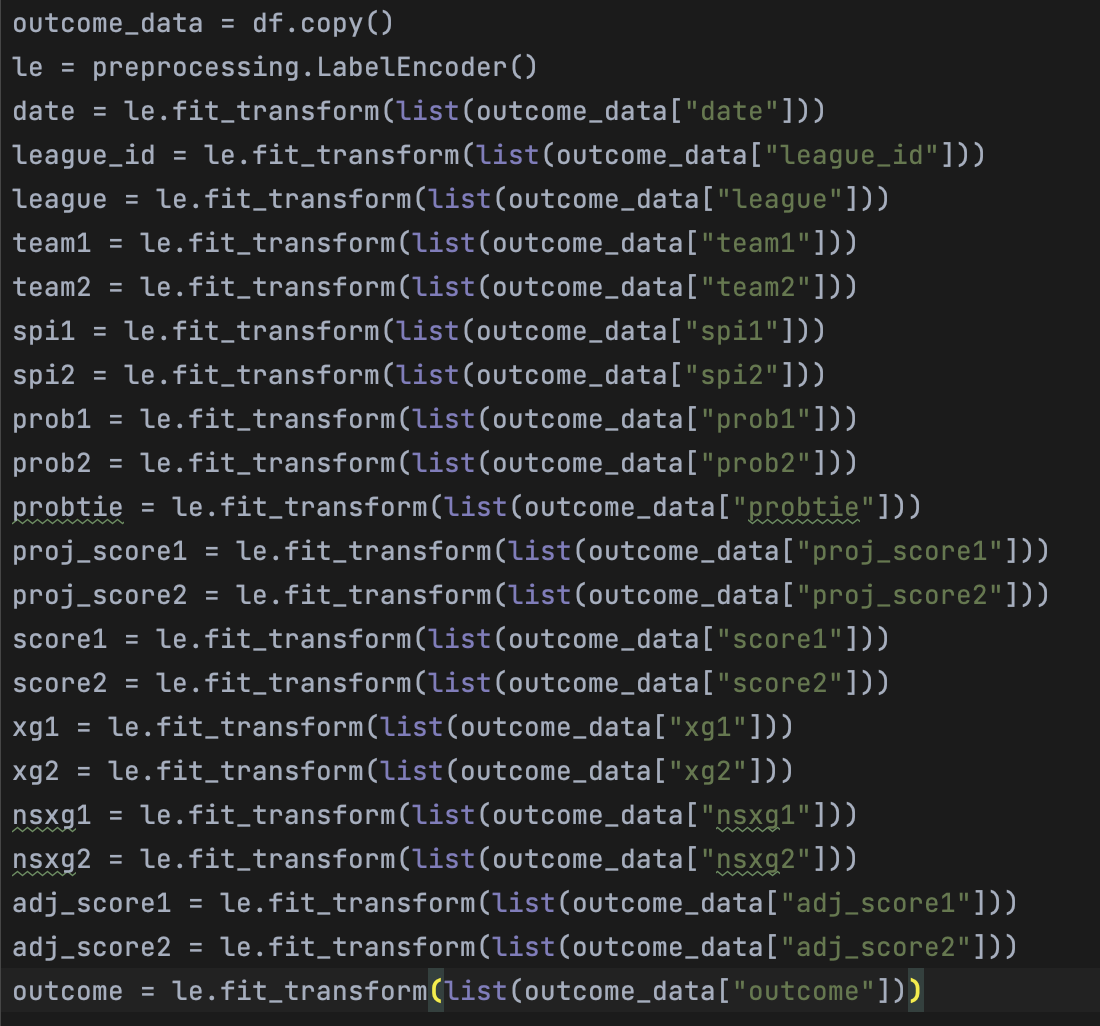












Preparation and Development of Model

Convert the dataframe into training and validation/test subsets by selecting a random sample of 80% of the data and identifying it as the train subset in order to build up and develop the model. Thus, 20% of the data remain for validation and testing.

o Remove all of the rows that make up the training set from the dataframe to create the validation/test set.

o Use the last column of train (the target class) to create y\_train.

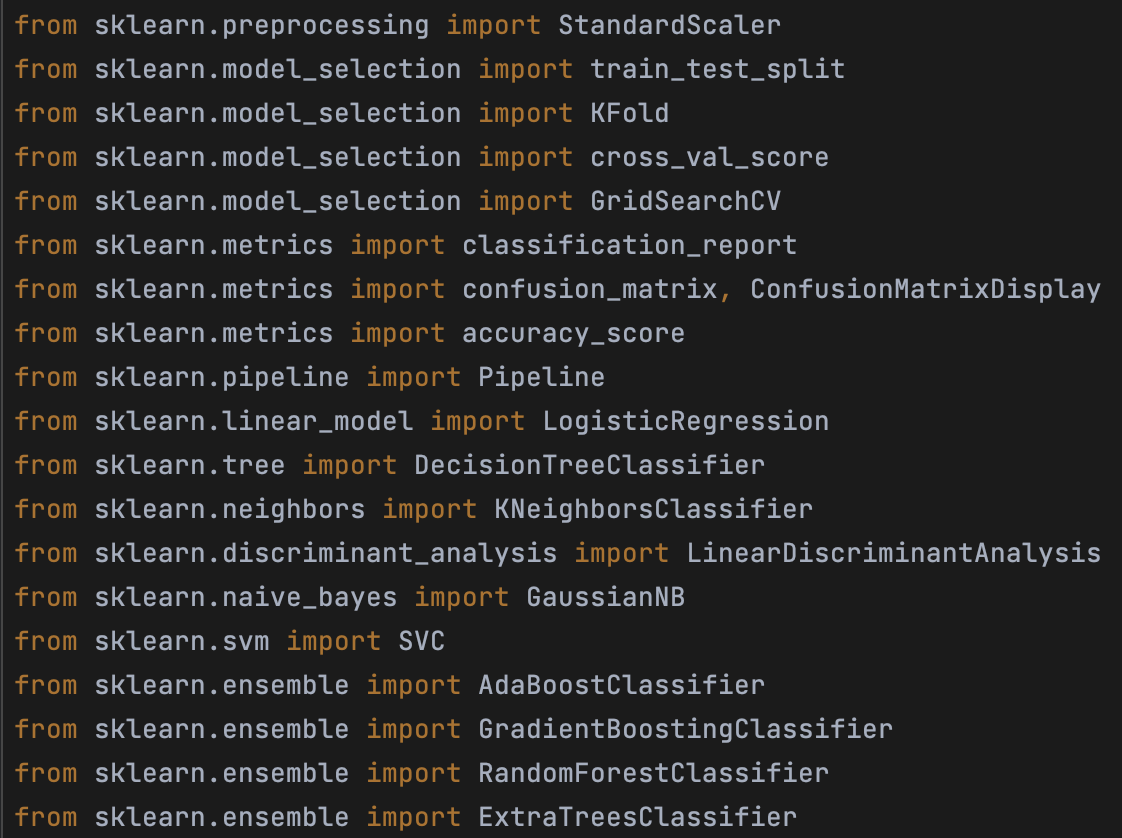
o Use all the columns in train to create x\_train, excluding the final one.

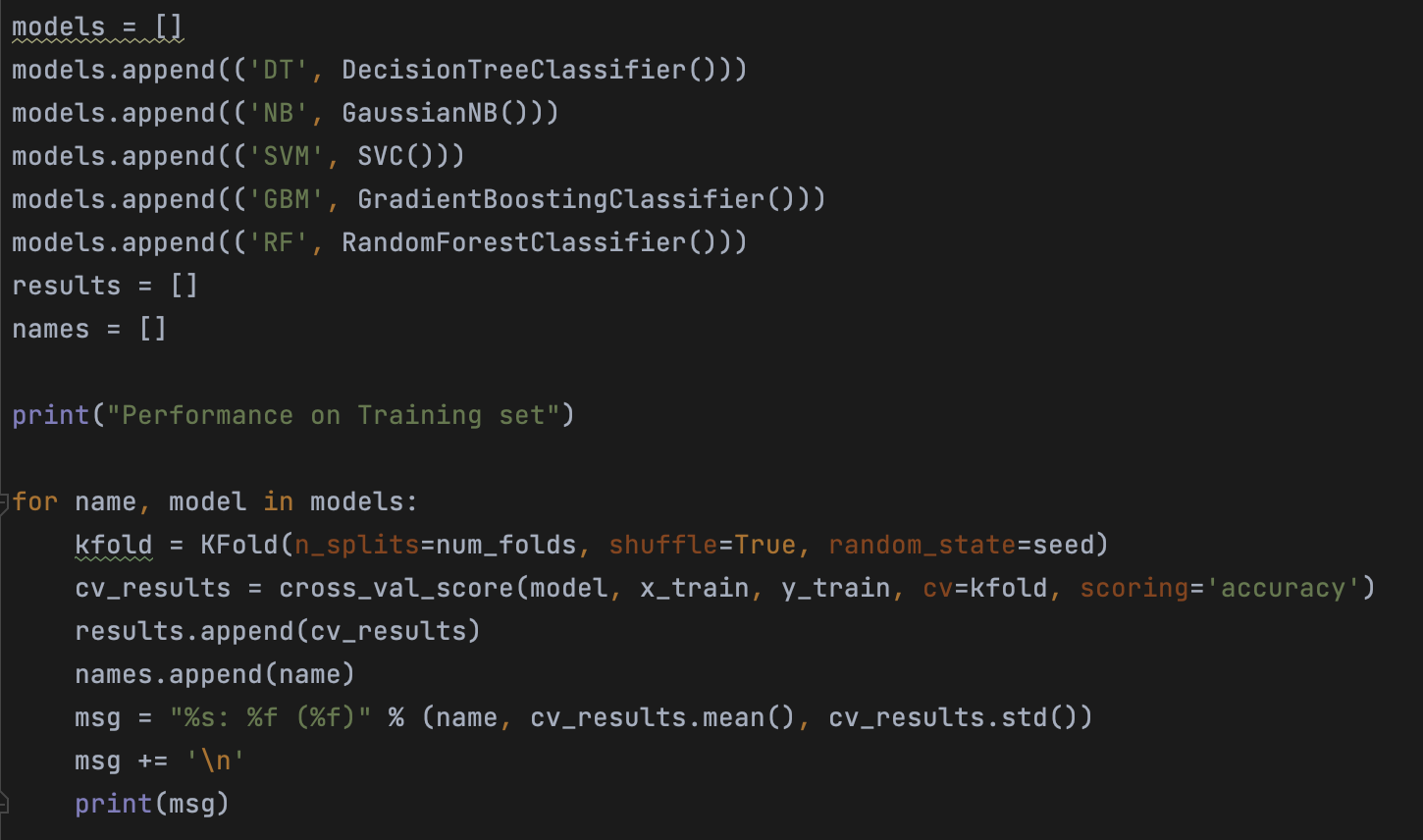
o The same process that was used to construct y\_train and x\_train may be utilised to create the validation set of y\_val and x\_val or (y\_test and x\_test).

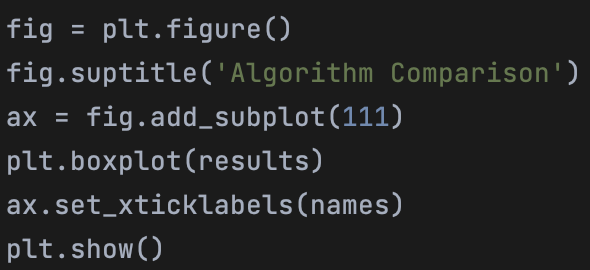
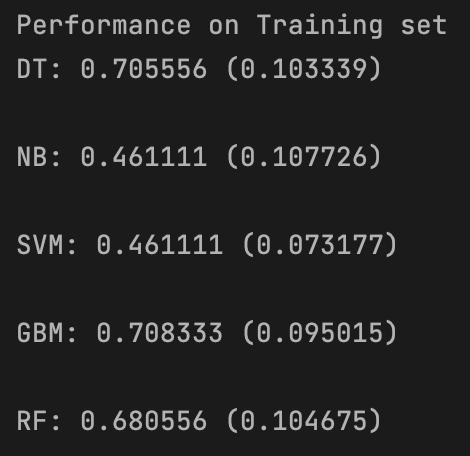
A black screen with white text

Description automatically generated with low confidence



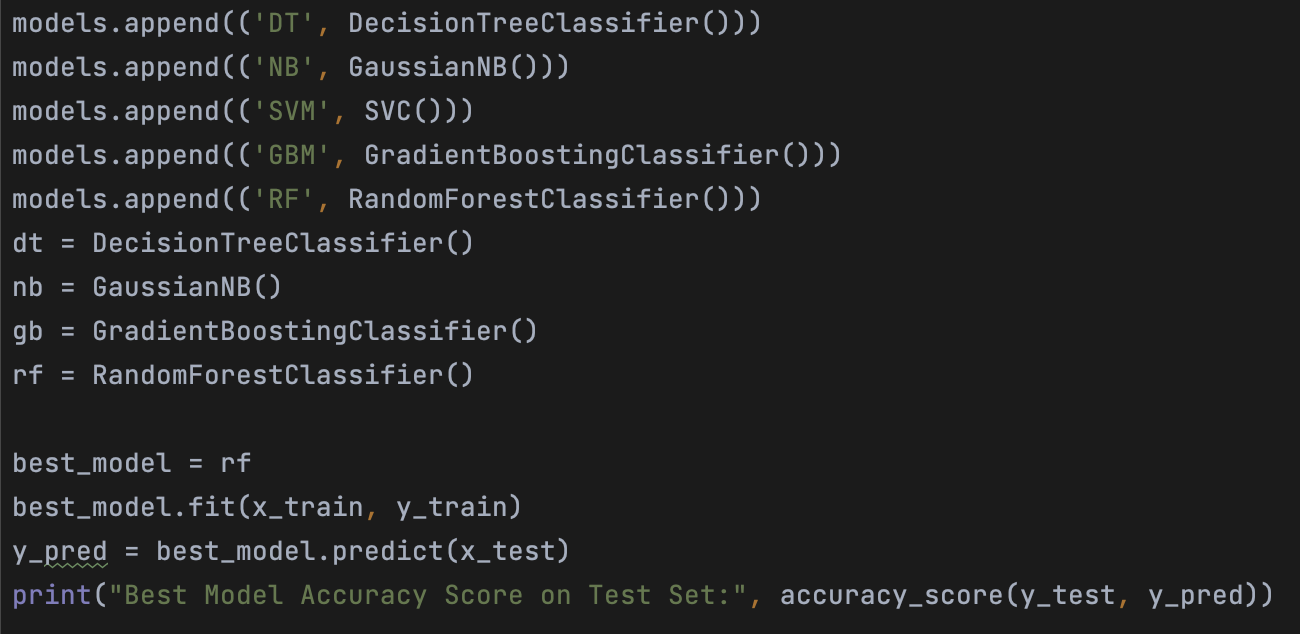






A picture containing diagram, text, rectangle, plan

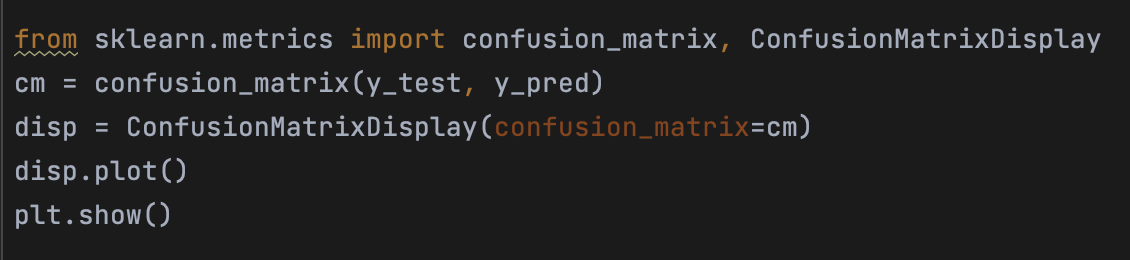
Description automatically generated

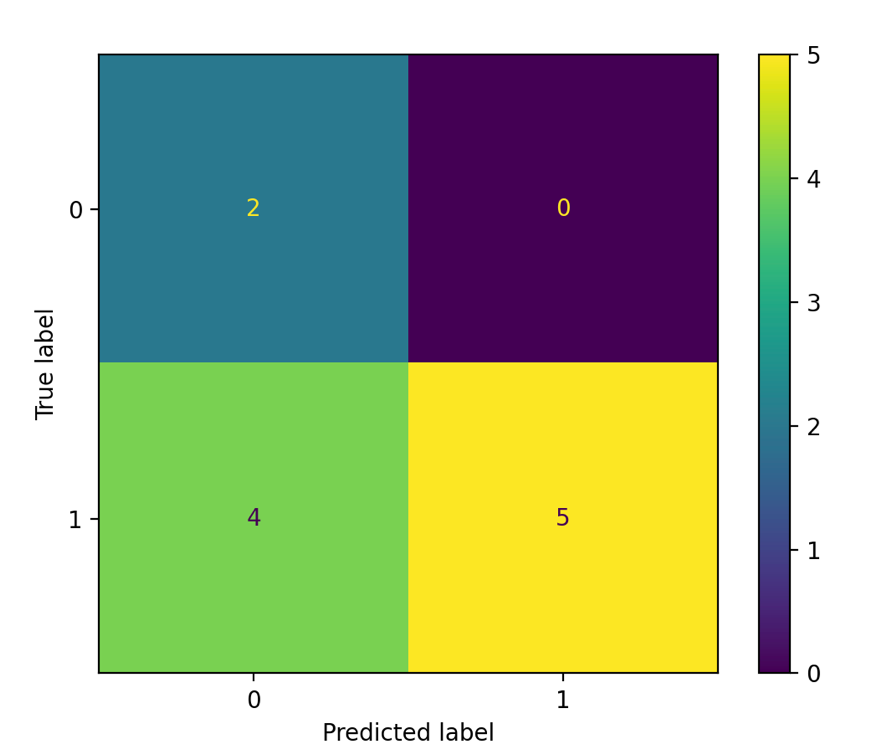


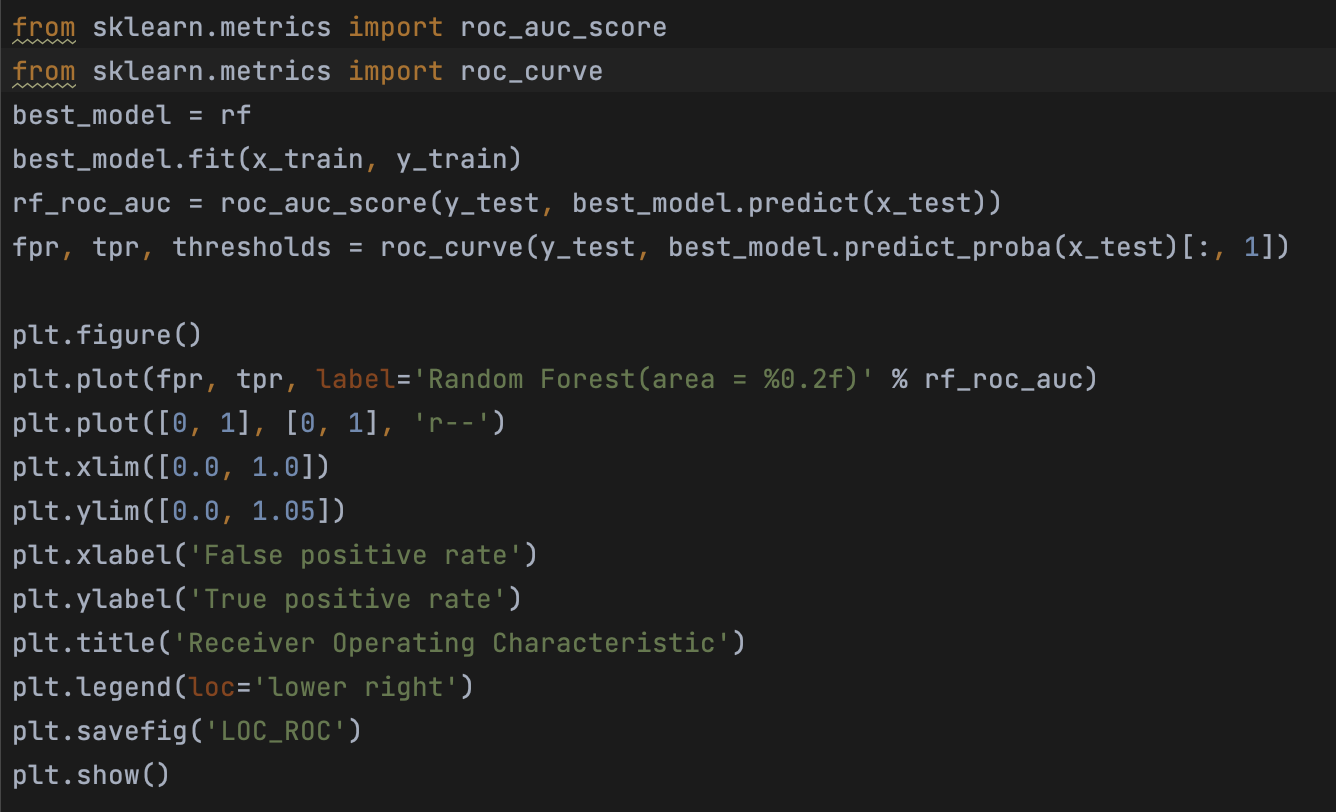


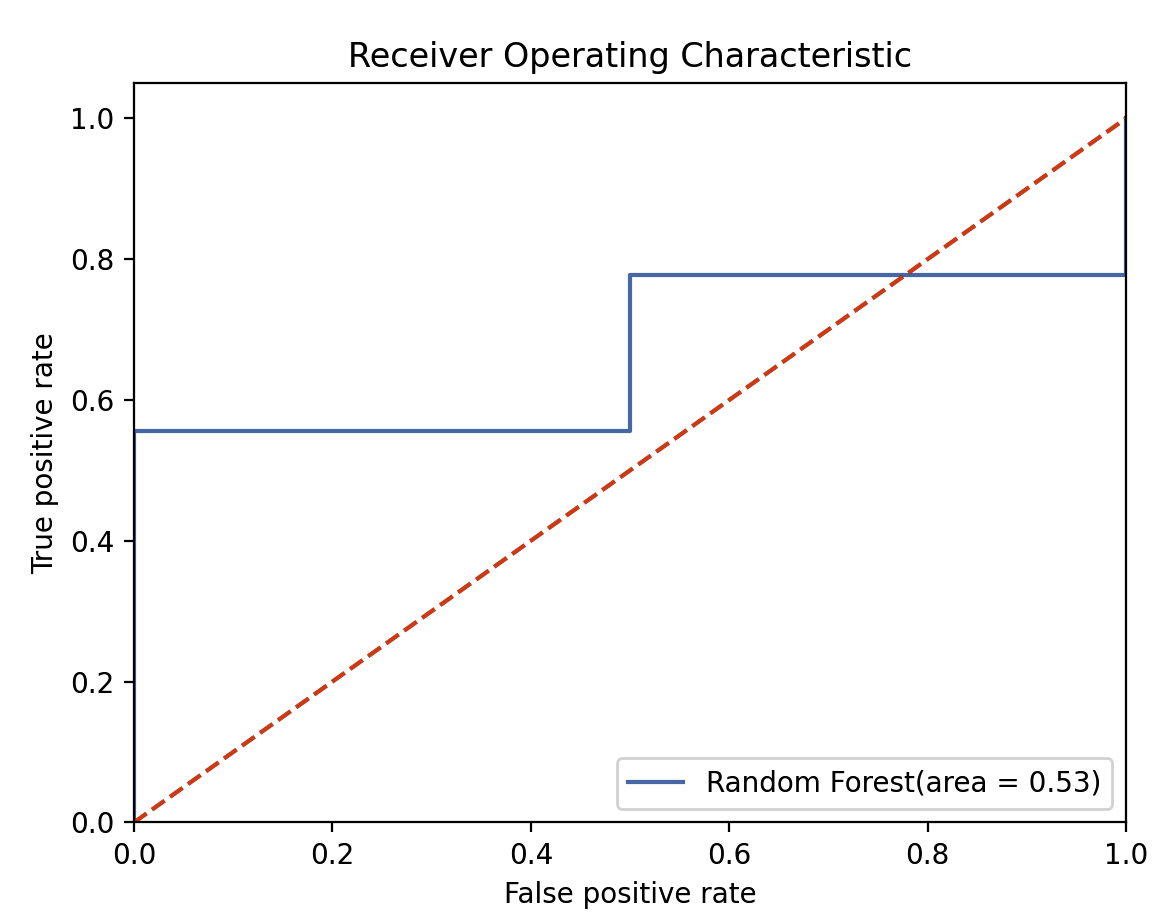
A screenshot of a computer screen

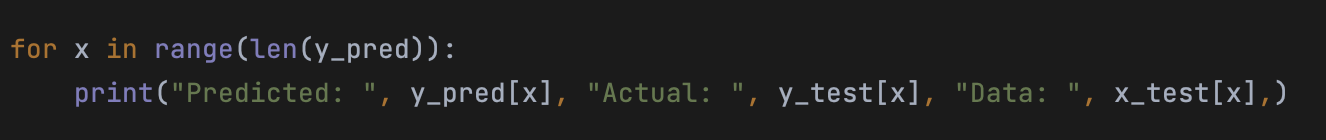
Description automatically generated with low confidence

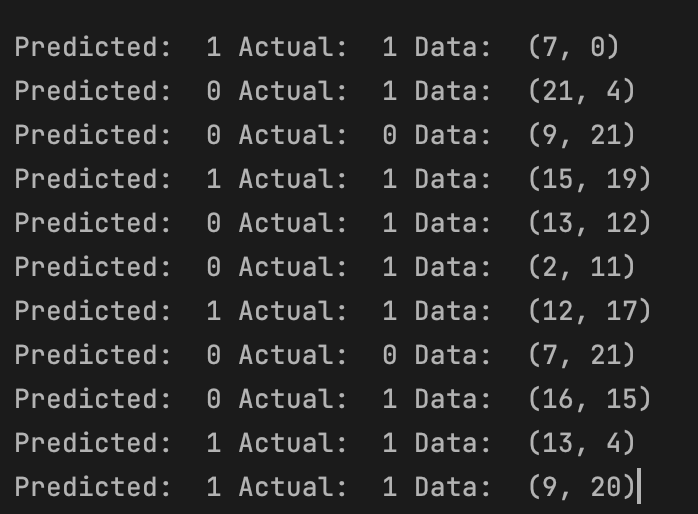












# Algorithm Implementation and Software Development Stage

1. Development of a desktop Tkinter software tool based on the highest performing algorithm.

2. The tool's deployment as a web-enabled platform utility.

The implementation of the algorithm as a desktop web application using the Streamlit package happens once the highest performing algorithm and machine learning model for heart disease prediction have been discovered from stage 1.

The implementation for this task is available at this link:

# Conclusion

In conclusion, the 2019 Women's World Cup match dataset analysis utilising EDA and predictive analytics has produced considerable insights into the patterns and elements influencing a team's performance. Our prediction model, which is based on team match performances and SPI ratings, has showed promise in predicting match results. This model can help academics, analysts, and sports fans better analyse and anticipate future Women's World Cup competitions when it is implemented on Streamlit. This study also highlights the significant contribution of data science to sports analytics, providing new opportunities for strategic planning and data-driven decision-making in the athletic industry. The model may be improved in the future by adding more datasets, including more factors, or using more sophisticated prediction algorithms to increase accuracy.

# References:

1. Boice, J. *et al.* (2019) *2019 Women’s World Cup predictions*, *FiveThirtyEight*. Available at: https://projects.fivethirtyeight.com/2019-womens-world-cup-predictions/ (Accessed: 16 August 2023).
2. The Scipy Developers (2016) *Scipy.stats.pearsonr*, *scipy.stats.pearsonr - SciPy v1.11.1 Manual*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html (Accessed: 16 August 2023).
3. The Matplotlib development team (2012) *Matplotlib.axes.Axes.text*, *matplotlib.axes.Axes.text - Matplotlib 3.7.2 documentation*. Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.axes.Axes.text.html (Accessed: 16 August 2023).
4. The Scipy Community (2016) *Scipy.stats.pearsonr*, *scipy.stats.pearsonr - SciPy v1.11.1 Manual*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html (Accessed: 16 August 2023).
5. NumPy Developers (2022) *Numpy.arange*, *numpy.arange - NumPy v1.25 Manual*. Available at: https://numpy.org/doc/stable/reference/generated/numpy.arange.html (Accessed: 16 August 2023).
6. The Matplotlib development team (2012a) *Matplotlib.axes.axes.set\_xticks*, *matplotlib.axes.Axes.set\_xticks - Matplotlib 3.7.2 documentation*. Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.axes.Axes.set\_xticks.html (Accessed: 16 August 2023).
7. The Matplotlib development team (2023) *Matplotlib.axes.axes.tick\_params*, *matplotlib.axes.Axes.tick\_params - Matplotlib 3.7.2 documentation*. Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.axes.Axes.tick\_params.html (Accessed: 16 August 2023).
8. The Matplotlib development team (2023a) *Matplotlib.axes.axes.legend*, *matplotlib.axes.Axes.legend - Matplotlib 3.7.2 documentation*. Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.axes.Axes.legend.html (Accessed: 16 August 2023).
9. P. (2023) *Box Plots using Plotly*, *Box plots in Python*. Available at: https://plotly.com/python/box-plots/ (Accessed: 16 August 2023).
10. Waskom, M. (2022) *Visualizing distributions of data#*, *Visualizing distributions of data - seaborn 0.12.2 documentation*. Available at: https://seaborn.pydata.org/tutorial/distributions.html (Accessed: 18 August 2023).