**Sports Prediction Application**

Final Project Report



Robert Anthony Zeelie

R1612D2076092

UNICAF University

BSc in Computer Science UU-COM-4006 Project II

Mr Anderson Chikazingwa

05/12/2022

# Acknowledgements:

I have taken efforts during this project and it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them however small or large their part was in the successful completion of this project. A special thanks to my mother, father, and sister who continued to cheer me on and provide a great source of inspiration to keep me motivated throughout my studies no matter the circumstances. Finally, I would also like to pay special appreciation towards all the guidance and support received from my lectures at UNICAF University. I take this opportunity to give my gratitude, without them I would not have been unable to complete this project and become who I am today.

# Abstract:

For this project, I proposed creating an application that leverages machine learning to predict outcomes for sporting events in Formula One. I began the project by deciding on the project management methodology I would use to manage the project as well as setting project goals and deliverables. Next, I started doing a large amount of research to better understand the numerous features of Formula 1 that influence the outcome of the sporting event. I then decided on the different technologies that I would use to implement this application from the database to the user interface. With these two steps complete, I started developing code that would fetch all the necessary data to train my machine-learning model and store it within my database.

After analysing my data I noticed some trends in the dataset which would be very effective for my application and began writing code that would run the dataset through multiple machine learning models to identify the most successful model when predicting the 2022 season. I decided to implement two models, namely Logistic Regression and XGBoost, to predict results for the application and started working on the source code for the user interface of the application. Once this was completed I updated my project documentation and released it for user testing, from which I received some insightful feedback to help improve the project. The project proved to be successful; however, there are areas which can be improved.

Table of Contents

[Acknowledgements: 2](#_Toc122267078)

[Abstract: 3](#_Toc122267079)

[Introduction: 6](#_Toc122267080)

[Executive Summary: 8](#_Toc122267081)

[Report Introduction: 8](#_Toc122267082)

[Report Executive Summary: 8](#_Toc122267083)

[Project Literature Reviews: 8](#_Toc122267084)

[Project Methodologies: 9](#_Toc122267085)

[Project Results Presentation: 9](#_Toc122267086)

[Discussion of Project Results: 9](#_Toc122267087)

[Report Conclusion: 9](#_Toc122267088)

[Report References: 10](#_Toc122267089)

[Report Appendices: 10](#_Toc122267090)

[Literature Review: 11](#_Toc122267091)

[Methodology: 13](#_Toc122267092)

[Methods and Tools: 13](#_Toc122267093)

[Participate Contribution: 15](#_Toc122267094)

[Project Management: 15](#_Toc122267095)

[Presentation of results: 17](#_Toc122267096)

[Numerical-Based Data: 17](#_Toc122267097)

[Project Illustration: 20](#_Toc122267098)

[Discussion of results: 22](#_Toc122267099)

[Conclusion: 24](#_Toc122267100)

[References 26](#_Toc122267101)

[Appendices: 28](#_Toc122267102)

[Figure 1: 28](#_Toc122267103)

[Figure 2: 29](#_Toc122267104)

[Figure 3: 30](#_Toc122267105)

[Figure 4: 30](#_Toc122267106)

[Figure 5: 31](#_Toc122267107)

[Figure 6: 31](#_Toc122267108)

[Figure 7: 32](#_Toc122267109)

[Figure 8: 33](#_Toc122267110)

[Figure 9: 34](#_Toc122267111)

[Figure 10: 35](#_Toc122267112)

[Figure 11: 35](#_Toc122267113)

[Figure 12: 36](#_Toc122267114)

[Figure 13: 36](#_Toc122267115)

[Code Snippet 1: 37](#_Toc122267116)

[Code Snippet 2: 38](#_Toc122267117)

[Code Snippet 3: 39](#_Toc122267118)

[Code Snippet 4: 40](#_Toc122267119)

# Introduction:

During the mid-late 1950s, breakthroughs within the sphere of artificial intelligence and machine learning stalled and expectations went unmet due to several factors such as running costs or computing power. The pursuit later picked up again in the 1980s which would then be boosted by the World Wide Web and resulted in an explosion of data and data sharing, on which machine learning relies. “An entity learns if, through its processing of information, the range of its potential behaviours is changed.” (Huber, 1991, p. 88). I am fascinated by machine learning and artificial intelligence, so I decided to join this growing trend by creating a machine-learning application that would be able to analyse and predict outcomes in sporting events. However, the only thing that trumps my enthusiasm for machine learning is my passion for Formula 1; thus, I decided to have the best of both worlds by developing an application that would leverage machine learning to predict results in Formula 1 racing.

Formula 1 is the pinnacle of motorsports and is a sporting spectacle to behold, but behind the scenes, there are hours upon hours of data collection and analysis to continuously evolve as well as improve both the vehicles and drivers. There are millions of simulations and tests being run on zettabytes of data collected from over 300 sensors on a Formula 1 car alone which is why the introduction of machine learning impacts the sport tremendously assisting humans with the understanding of the data. “In fact, humans have a certain limitation when processing a large set of information. However, Artificial Intelligence techniques can overcome this issue. Furthermore, sports have a great amount of data to consider, thus, it is a great example of an AI problem” (G.Fialho, A.Manhães, & J.P.Teixeira, 2019, p. 131). This brings me to the problem statement of this project; with the growing competitiveness in the world of motorsports, and hundreds of teams battling for the crown of the fastest on the grid, it becomes difficult for team managers to objectively analyse their competitors, drivers, as well as their performance in an effective manner.

The goals of the Formula 1 Predict Application will be to assist teams (constructors) by predicting both constructors and drivers chances at the championship, qualification position, and race results. The application will take into account how they are affected by external factors such as performance metrics, weather, country, nationality, and circuit of the diver and team. In addition, the application will suggest the best drivers for their cars and vice versa. That being said, a lot of research would have to be done before the project can begin using various techniques and resources that will help better understand all the features and factors that can contribute to the outcome of these events. This project will help me bring together the knowledge and skills developed throughout my degree programme in a bid to improve my weaknesses and also understand my strengths. I hope to develop technical expertise in computer science topics and gather experience in software development that will be essential to furthering my career. In addition, this project should give me practical experience in applying methodologies, skills, and tools learnt as well as improving my understanding of project management and the milestones associated with a software project.

# Executive Summary:

## Report Introduction:

The introduction is the opportunity for the report to introduce the project topic to the reader and describe the research problem that the project attempts to solve. The introduction section includes a brief background into the subject area of the project that will help readers who are unfamiliar with the chosen topic. Once the reader is given adequate insight into the knowledge area of the project, they will be better equipped to understand my overall project objectives, goals, learning objectives and the reasoning behind any final analysis, solution, resolution or conclusion. In addition, it provides the reader with the problem statement, information on the project's research methods, planning methodologies and tools used, goals and objectives the project should achieve, and the learning outcomes on completion of the project.

## Report Executive Summary:

An executive summary is a systematic summary of what will be included in this report. This section will expand on each section of the report's table of contents and will be concise but comprehensive to keep the reader informed about the contents of the report. It is more of an outline of the report and can be differentiated from the abstract by being a summary of the structure of the actual report rather than a summary of the content of the project itself.

## Project Literature Reviews:

The literature review section of this report is a comprehensive summary of previous research done on the topic area this report is aimed at. The literature review enumerates, describes, summarizes, and objectively evaluates acquired literature on theoretical aspects of the subject area using primary and secondary resources. “Put briefly, a literature review summarizes and evaluates a body of writings about a specific topic. The need to conduct such reviews is by no means limited to graduate students; scholarly researchers generally carry out literature reviews throughout their research careers.” (Knopf, 2006).

## Project Methodologies:

The methodologies section will provide the reader with some insight into the various tools and methodologies used to carry out the project research as well as the reasoning behind the use of these tools. It will describe each tool and methodology in sufficient detail to inform readers who may not be familiar with the particular approach. In addition, it outlines participant involvement, the level of contribution to the project from any participants, and project management methodologies.

## Project Results Presentation:

This section plays a significant role in the overall outcome of the report and is used to illustrate the raw results using graphs, charts and tables. These display numerical-based data as well as visualisations that display non-numerical-based data such as screenshots and illustrations of output from the application. All visualisation and images are adequately described and make use of appropriate colours, labels and descriptions.

## Discussion of Project Results:

The discussion of the project results section is a continuation of the aforementioned section and is purely focused on critically analysing the results to identify wherever or not they address the problem statement, research goals, objectives and learning outcomes of the project. “The discussion reviews the findings and puts them into the context of the overall research. It brings together all the sections that came before it and allows a reader to see the connections between each part of the research paper.” (R.Dunton, 2021).

## Report Conclusion:

The conclusion section of the project report reflects on all the work that has been done to get to the final stage of the project in a tone that honestly examines both the positives and negatives. It briefly summarizes all the main points of the report, comments on the successful elements and suggests areas in which future improvements could be made.

## Report References:

The reference list is an important part of the project report as it proves to the reader that the research of the report has credibility by referring to credible sources of work. The reference list makes use of the APA referencing style for all citations included in the body of this report.

## Report Appendices:

This section is used to include additional information about the project that may not need to be included directly in the main body of the report. The information included in the appendices section is only complementary to the main report and can include user documentation, user guides, programming code snippets, and any other relevant information or illustrations. Anything that is included in this section such as tables, graphs, images, source code and figures have appropriate headings, descriptions, and captions where necessary.

# Literature Review:

This section will classify and evaluate the various research and accredited scholarly pieces of work already explored in the topic areas involved with machine learning, sports prediction, and Formal 1. A prime example of data-driven performance optimization in sports is written about in a book called Moneyball: The Art of Winning an Unfair Game (M.Lewis, 2004), which is based on the story of how a baseball manager used statistical data and analytics to build a competitive baseball team despite the small budget. This is the exact problem that this project is aimed at solving except, the sport is Formula 1 which is rich in quantifiable features, making it ideal for the application of machine learning. “The proposed frameworks based on machine learning are known to produce additional valuable insights in sports analytics when it comes to performance analysis and assessment, decision-making in strategy and winner prediction, being consistently used in high-level competitions such as Formula 1” (H.Sicoie, 2022).

The first piece of work that I researched was “2022 Formula One Sporting Regulations” (FIA, 2022) official rule book released by the governing body of motorsports, Fédération Internationale de l'Automobile known as FIA, to fairly manage the sport. This piece of work was essential to understanding the various rules and regulations that govern the sport as well as some insight into features of the sport that can influence the results. I then decided to research previous studies done on the use of artificial intelligence in sports to understand how the project should progress and identify milestones. C.Ditcher, M.OReilly, and E.Delahunt, (2021) stated that the benefits of machine learning in sports have already led to improvements in the devices used to acquire data, information extracted from the data acquired by the devices, processing capabilities of the data, the understanding of sports performance, and injury risk prediction.

In addition, the book Machine Learning in Sports: Identifying Potential Archers briefly highlights the association of different performance metrics and features that influence the archers' performance and the employment of machine learning algorithms in the identification of potential archers (R.M.Musa, Z.Taha, A.P.P.A.Majeed, & M.R.Abdullah, 2018). R.M.Musa, Z.Taha, A.P.P.A.Majeed, and M.R.Abdullah (2018) concluded that the utilisation of machine learning is non-trivial in the pursuit of a technique for the objective evaluation and reasonable classification of the performance classes of perspective archers. However, these machine learning predictions are not absolute and will not always come to fruition as outlined in the results by O.Hubáček, G.Šourek, and F.Železný (2019) when measuring the accuracy of their sports betting algorithm. “. The accuracy of the bookmakers’ model, predicting the team with smaller odds to win, levels over these seasons at 69 ± 2.5. Generally, in terms of accuracy, the bookmakers’ model is slightly superior to the neural models, which in turn beat the logistic regression baseline (accuracy of 68.7 with odds, and 67.26 without).” (O.Hubáček, G.Šourek, & F.Železný, 2019).

Due to the successful results of machine learning integration in sports, I decided to research previous studies into the employment of machine learning algorithms in Formula 1, if any. It was very difficult to find previous explorations into this topic area but I was able to find a couple of academic papers partially aimed at solving the problem statement of my project. H.Sicoie (2022) aimed his paper at investigating the prediction of race winners by proposing and analysing multiple machine learning algorithms. His paper aligns with the scope of this project and was very informative concerning the various machine learning algorithms however the project did not take into account external factors but was purely based on features directly linked to the driver. The paper by H.Sicoie (2022) also helped me find Ergast API, an open-source Application Programming Interface (API) designed for users to freely retrieve data using specific URL parameters from a web-based database. Full documentation regarding Ergast API can be found at <https://ergast.com/mrd/> and stores historical Formula 1 data from the 1950s in a database illustrated in Figure 1 of the appendices section.

Another recent and deeply insightful paper written by C.Garvin, D.Julian, B.Lee, and A.Kosikowski (2022) aimed at successfully predicting the finishing position of each driver in a particular race given data such as their starting position, relevant team, and circuit information. The paper details the processes and steps taken to solve their problem statement as well as helped guide me to the various data modification and cleaning steps that would be necessary for the machine learning algorithms to understand and interpret the data. In addition, there was a very helpful Medium article by W. George (2021) that provided me with guidance on how to approach the problem statement of this project and interpret the results of the machine learning algorithm. This article clearly explains why this problem statement cannot be approached using a classification problem, but instead the results should be ordered by the regressor prediction results.

# Methodology:

Before I could begin the research necessary for this project, I first needed a clearly defined the topic and problem statement that will focus my research efforts in a specific direction. The main topic area for this project is the use of machine learning in the prediction of Formula 1 outcomes and the problem statement is ‘With the growing competitiveness in the world of motorsports, and hundreds of teams battling for the crown of the fastest on the grid, it becomes difficult for team managers to analyse their competitors, drivers, as well as their performance. Formula 1 teams need to be able to analyse and predict their chances at the championship, races, and qualifying events as well as which drivers best suit their cars to effectively manage and build their team.’

## Methods and Tools:

With both the topic area and problem statement defined, I began deciding on the necessary tools, methods, and resources I would employ to conduct the research. There are two main approaches to research that can be used depending on the goals and objectives. The first of these approaches is qualitative research which involves the collection of data primarily through observation and exploration, while quantitative research involves the collection of data that can be measured numerically. “It can be defined as research that explains phenomena according to numerical data which are analysed by means of mathematically based methods, especially statistics.” (K.Yilmaz, 2013). I decided to use quantitative research due of the nature of this project and the results that would be produced.

Quantitative research typically involves gathering data that can be placed directly into a table, chart, or graph to make observations using statistical analysis. The first step however, is gathering the historical data sources of this particular motorsport, for which I wrote a python script that collects all the necessary data from the Ergast API and store it in my database. This database as well as the necessary python scripts used to fetch all the data can be found in the ‘DB’ folder as seen in Figure 2 of the appendices section. Consulting literature, machine learning applications within sports can be done using the previous seasons for training and the season of interest for testing. Once I had collected all the necessary information about constructors, constructor standings, drivers, driver standings, races, results, qualifying times, circuits, and more, with corresponding attribute values as depicted in Figure 1 of the appendix section, the next step was for me to enrich it with weather information about each particular race.

For each race collected from the API, a URL link from Wikipedia is appended as a column in the database, and by accessing the specific link using the Python library BeautifulSoup, I was able to extract information about qualification times as well as the weather during the race. Next, I performed data manipulation using python libraries, like Pandas and Numpy, to prepare the data and merge all the required table data into a single table that holds all the required data. To assure the relevancy of the data, this singular table which holds all the data only starts from the year 1983 due to the differences in qualification rules before 1983 and the lack of information on the official Formal 1 website. Further preliminary steps included the calculation of the driver's ages using the date of the birth column as well as the cumulative difference in qualifying times for each. This would facilitate an indicator for the machine learning algorithm to show how much faster the first car on the grid is compared to the other cars. All this data presentation can be found in a file ‘dataPreperation.py’ in the DB folder and the main data preparation function can be seen in Code Snippet 1 of the appendices section.

Finally, it was time to perform data analysis on my dataset and discuss the observations, all the python scripts that I used are found in the ‘DataAnalysisResults’ folder as seen in Figure 2 of the appendices section. The two types of observations include descriptive observations that attempts to describe information about the sample data, and inferential observations attempts to draw conclusions about the sample data. I will make use of both types of observations, beginning with descriptive observations. Firstly, I analysed which tracks have hosted the most races, as seen in Figure 3 of the appendices section, and decided to plot out the importance of qualification in pole position that leads to winning the races, as seen in Figure 4 of the appendices section. Figure 4 depicts a clear correlation between the two features; I therefore decided to plot the importance of racing in your home country, as seen in Figure 5, which presented a clear trend in winning races in your home country.

In addition, I decided to create a scatter plot to understand trends between various features of the dataset, and as seen in Figure 6 of the appendices section, there are several trends between various features which is exciting for my machine learning algorithms. Lastly, I decided to run my dataset through numerous machine-learning models to see the accuracy percentage of each algorithm for predicting the 2022 season. Figure 7 in the appendices section displays these results with the Linear Regression and XGBoost Regressor outperforming all the other algorithms and tying for first at 66.56% accuracy. This is the reason why I decided to implement both algorithms, the Linear Regression algorithm for qualifying predictions while the XGBoost Regressor is responsible for all other predictions. Code Snippet 3 of the appendices section shows the Linear Regression implementation while Code Snippet 4 shows the XGBoost Regressor implementation, both of which can be found in the ‘ML’ folder of Figure 2.

## Participate Contribution:

Due to the nature and complexity of this project, I required some input at different stages of the development life cycle and this subsection is to highlight the contribution of these participants. During the initial phase and proposal of this project, I received some valuable guidance from my lecturer, Mr Richard Ndonye Ngung, who advised me on the setting up of the project thesis goals understandably and clearly. During the second phase of this project, I was assisted by my lecturer Mr Anderson Chikazingwa who helped clarify the objectives of this report and project presentation. Lastly, during the testing phase of this application, some of my friends and family graciously volunteered to test the project and provide me with some feedback and improvements which can be found in the file ‘Improvments.txt’ of Figure 2.

## Project Management:

Project management is the process of guiding the workflow of a project to achieve all project goals within the given constraints. This is an important aspect of any project as it provides control, direction, and purpose enabling the best work for the project. A project management methodology defines a set of principles and practices that assist with the guidance and organisation of projects. Several project management methodologies can be applied to a project, each with its benefits and tailored to a specific project based on project nature, size, time, constraints, and industry. When it comes to choosing the correct project management methodology there are several factors to consider such as team size, cost, ability to take risks, timeline, and client collaboration. For this project, I decided to make use of the Scrum project management methodology to provide a structure for my project development life cycle.

Scrum methodology evolved from the agile project management methodology and splits work or goals into short cycles, known as sprints, which are usually about 1 to 2 weeks in length. At the end of each sprint, a sprint review is done to assess goals and progress as well as to plan the following sprint and make any necessary changes to the plan. I chose this methodology for many reasons including the promotion of incremental development, continuous improvement, ease of understanding and implementation, encouragement of planning, improved focus, adaptability to change, and production of results that can be measured throughout the project to effectively track the progress. “Scrum was first introduced in 1997 and has since become the most widely applied agile software development framework. At its core, Scrum splits development into iterations not longer than four weeks (called sprints). At the end of each sprint, a shippable product increment is delivered to the user.” (M.Hron & N.Obwegeser, 2018, p. 5446). Please see Figure 8 in the appendices section for the planned timeline of phase 1 for this project, and Figure 9 for the current phase of the project, with each week being a sprint.

# Presentation of results:

## Numerical-Based Data:

To give an overview of the core findings of this report, I will first make use of tables and graphs to display the numerical-based data followed by illustrations of the software output. When evaluating the rankings for each driver in a race in the 2022 testing season, it was not surprising that the accuracy was unsatisfactory, as it is intrinsically difficult to correctly determine the standings of the driver line-up for each race in the season. As a solution, I decided to alter the values of the dummy value constraints in Code Snippet 2, which drastically improved the results. Table 1 provides the scores for the chosen machine learning models with the fine-tuning of the model parameters used to make predictions. The predicted final standings by both the Linear Regression and XGBoost Regressor models for the driver championship of the 2022 season can be found in Table 2 and Table 3 respectively to facilitate more comprehensive visualisation.

Table 1:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | fit | Grow-policy | n-estimators | max-depth | gamma | n-jobs | Random-state | Tree-method | Test Score |
| Linear Regression | true | X | X | X | X | X | X | X | 0.66 |
| XGBoost | X | 1 | 5 | 26 | 5 | 5 | 7 | auto | 0.66 |

Table 1: Model performance scores with tuned parameters after initial dataset training and using the 2022 season for testing.

Additionally, I have constructed Table 4 which presents the effectiveness of predictions for a region of interest, namely the top 10 drivers from Tables 2 and 3. This table is motivated by the fact that at the end of each race only the top 10 drivers earn points, so ideally the fewer discrepancies within this region the better. To add more value to Table 4, I have added different margins of error. It can observed that as the margin of error increases, the prediction percentage dramatically increased, with the best prediction percentage even reaching 84%. Finally, Chart 1 of this section presents the most important and influential features of my dataset on the overall prediction outcome. This chart is very informative and plays a big role in understanding which features had the most affect on the outcome of the results.

Table 2:

|  |  |  |
| --- | --- | --- |
| **results** | **driver** | **predicted** |
| 1.0030061 | hamilton | 1 |
| 1.2445014 | max\_verstappen | 2 |
| 2.6338356 | leclerc | 3 |
| 2.9977677 | perez | 4 |
| 4.0010867 | russell | 5 |
| 5.1051207 | sainz | 6 |
| 6.583343 | norris | 7 |
| 8.038206 | ocon | 8 |
| 9.0238695 | alonso | 9 |
| 10.131652 | bottas | 10 |
| 11.150773 | vettel | 11 |
| 11.984608 | ricciardo | 12 |
| 12.059384 | hulkenberg | 13 |
| 13.131813 | kevin\_magnussen | 14 |
| 13.704631 | gasly | 15 |
| 15.251419 | stroll | 16 |
| 15.806149 | mick\_schumacher | 17 |
| 16.298725 | tsunoda | 18 |
| 17.619259 | zhou | 19 |
| 18.734026 | albon | 20 |
| 19.516235 | latifi | 21 |

Table 2: This table presents the final standings of the driver championship prediction results according to the Linear Regression model.

Table 3:

|  |  |  |
| --- | --- | --- |
| **results** | **driver** | **predicted** |
| 1.0473894 | hamilton | 1 |
| 1.1410867 | russell | 2 |
| 1.8995324 | max\_verstappen | 3 |
| 2.2568364 | leclerc | 4 |
| 4.0094362 | norris | 5 |
| 4.6703614 | perez | 6 |
| 4.9896352 | sainz | 7 |
| 6.7372157 | ocon | 8 |
| 6.8927217 | alonso | 9 |
| 8.9647345 | bottas | 10 |
| 9.9373346 | vettel | 11 |
| 11.994256 | ricciardo | 12 |
| 11.998734 | gasly | 13 |
| 12.866553 | hulkenberg | 14 |
| 13.063378 | albon | 15 |
| 13.988468 | kevin\_magnussen | 16 |
| 15.855337 | stroll | 17 |
| 15.997362 | mick\_schumacher | 18 |
| 16.637667 | tsunoda | 19 |
| 16.966388 | zhou | 20 |
| 22.988378 | latifi | 21 |

Table 3: This table presents the final standings of the driver championship prediction results according to the XGBoost Regressor model

Table 4:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ML Models** | **TOP 10 +-1** | | **TOP 10 +-2** | | **TOP 10 +- 3** | |
| **True** | **False** | **True** | **False** | **True** | **False** |
| **Linear Regression** | 0.48 | 0.52 | 0.60 | 0.40 | 0.74 | 0.26 |
| **XGBoost Regressor** | 0.49 | 0.51 | 0.60 | 0.40 | 0.73 | 0.27 |

Table 4: This table represents the correctly and incorrectly predicted results for both types of machine learning models. It has three main groups, the margin of error +-1, +-2, and +-3 which helps understand the discrepancies between the different prediction models.

Chart 1:

Chart 1: This chart presents the values of influence that each of these 15 features has on any prediction. It is the best example of which features greatly affect results and gives the data meaning as we can see how the machine learning models interpret the data.

## Project Illustration:

This subsection will present and illustrate the raw results of the project which cannot be represented by graphs and tables. Firstly, Chart 2 presents the workflow chart for this project with the intermediary steps summing up the project's entire process when predicting. This gives a broad overview of the steps taken to complete this project as well as the processes and learning phases of the project. Figure 10 in the appendices section provides a glimpse into the dataset after all the preparation steps have been completed. This is an example of the training data used to train the machine learning models and eventually make predictions. Next, Chart 3 presents the hierarchy and structure of the actual application as well as a brief description of the goals for each page that the project has. Lastly, Figure 11 in the appendices section illustrates how the output of predicting a race for a driver is displayed; Figure 12 illustrates how the output of predicting the finishing position of a constructor in the championship is displayed, and Figure 13 illustrates how the output of predicting the best two drivers for a constructor is displayed.

Chart 2:

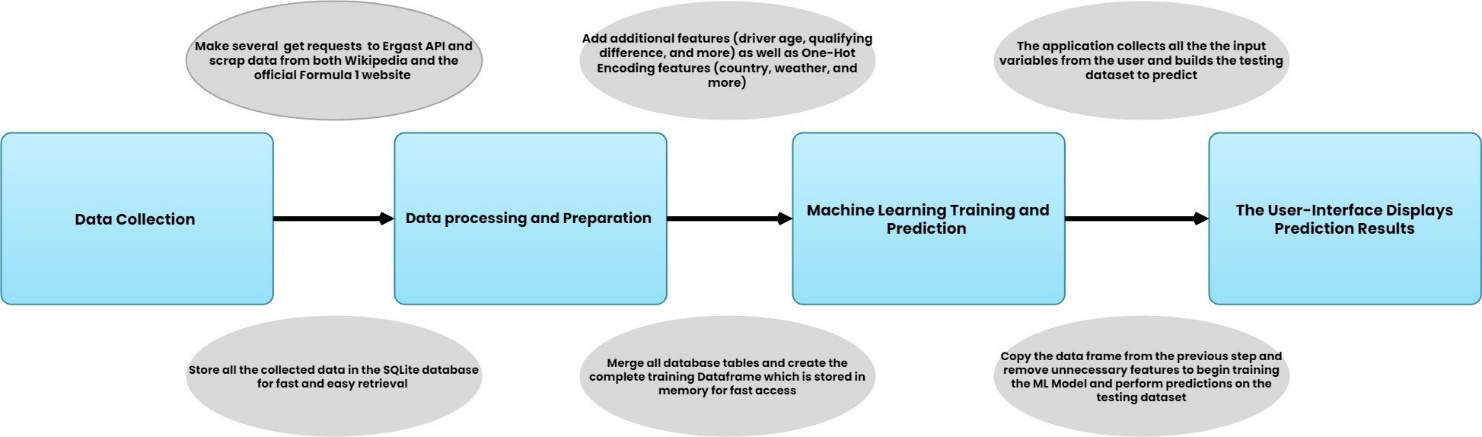


Chart 2: This chart presents the workflow of the application as well as the intermediary steps performed to make a prediction.

Chart 3:

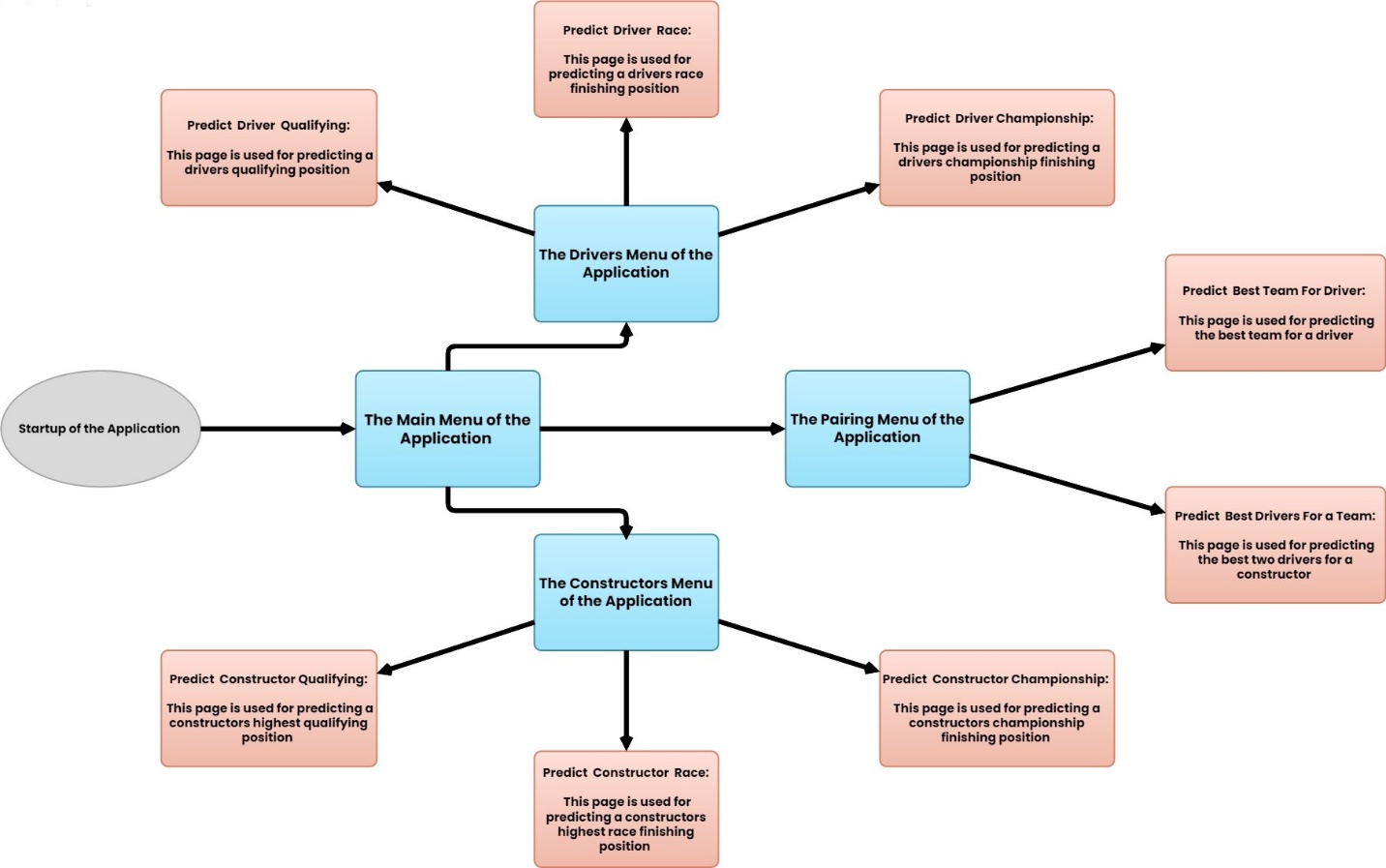


Chart 3: This chart presents the structure of the application as well as a summary of each page of the application.

# Discussion of results:

This section will aim to discuss and analyse the results obtained above and whether they address the problem statement, goals, objectives, and learning outcomes. However, first I will describe the challenges imposed by the methodology as well as threats to the validity of the dataset and issues faced throughout the project. Throughout this project, I faced multiple challenges when it came to the dataset that would be used for the machine learning training, starting with the initial data collection. My initial dataset was scraped from multiple Formula 1 historical data websites and took extremely long to fetch the data which is why the thought of a database was introduced. In my hurry to get a working dataset to begin training machine learning models, I ended up getting poor results due to low-quality feature sets. After a thorough review of several research papers, I found a reference to the Ergast API mentioned in a paper by H.Sicoie (2022) but, given the abundance of data available on the API, the next challenge was the decision on which data was the most relevant and valid.

Certain information had to be excluded from the dataset such as lap times, pit stops, racing incidents, and other features that would negatively impact the modelling in terms of time constraints. I would have saved significantly more time had I put more time into identifying the best features initially, rather than rebuilding my dataset multiple times. The next challenge I faced was during the processing of that dataset into train test splits and understanding how to feed the data to my two models, as I needed to train several variations of the models to make several different types of predictions. While this is not a dramatic setback, it did emphasise the versatility of the application and how different the goals of this project are from other academic works on this topic. Small setbacks similar to this occurred throughout the development life cycle causing greater delays than they should have. This would not have happened had I planned out each step more carefully or if I had more experience in these topics.

Despite these setbacks and challenges, the aforementioned results provide proof that this project was a success and has satisfied the requirements defined in the project profile. Table 1 from the results section provided the accuracy of prediction for my machine learning models which is a satisfactory percentage and can only improve with feature and model fine-tuning. Chart 3 of the results section gives an overview of the application structure which accomplishes all defined goals and even more by providing prediction capabilities for qualification, race results, and championship results for both drivers and constructors as well as providing the ability to suggest the best pairing for drivers and constructors. The machine learning models take into account the impact of external factors such as circuits, weather conditions, country, nationality, performance metrics, and more to make informed predictions. In addition, the application provides a very simple and user-friendly interface allowing users to navigate as well as understand the outputs with ease.

Ultimately, though, this project was a success and helped me gain invaluable experience in the field of machine learning as well as software development. It helped me bring together all the knowledge and skills developed throughout my degree programme and put it into actual practice rather than theoretical understanding. I realise that my weaknesses include the experience of problem-solving and various machine learning topics but, just with the work performed in this project, I have noticed some subtle improvements. I am hopeful that this application will assist Formula 1 team managers as well as others by making predictions on several aspects of the sport in a bid to gain a competitive advantage and improve planning and strategy.

# Conclusion:

This report began with an introduction to the topic area as well as a definition of the goals and problem statement that this project seeks to solve. With the focus on giving a Formal 1 team a competitive advantage, the approach to provide a scientific answer to the problem statement is supported by a thorough review of literature conducted within the field of machine learning and sports prediction. Tools and methodologies from similar academic pieces of work related to machine learning and sports were investigated and key findings were noted and adopted. The collection of the data through web-scraping and API requests was conducted directly from reliable data sources and the decisions relating to data selection were based on domain knowledge in this field as well as from analysis and understanding of the sport. Furthermore, all the data underwent processing and manipulation before being merged into the final training set.

After the development of the application and data analysis was completed, it was clear to see that the results of the project were a success, however, there was still room for improvement. There were several times in this project when I struggled with topics that occurred due to a lack of knowledge or experience in a specific topic. Thankfully these hurdles were cleared by reading through various pieces of documentation, but these struggles were time consuming and put pressure on my deadlines, resulting in the lack of time to begin working on the list of improvements for the project. Overall, the project can predict the results of qualifying, race and championship for all current drivers and constructors with a 66% accuracy rate as discussed in the results section. In addition, the application can suggest the best pairing between drivers and constructors taking into consideration the impact of external factors. This result satisfies the projects goals and problem statement; however, some improvements can be made such as increasing the prediction accuracy percentage by fine-tuning the machine learning models, including different features to the dataset to accommodate more external factors and display the predictions for all constructors and drivers at the same time.

The success of this project may spark further interest in the integration of machine learning into Formula 1 to improve the sport as well as gain competitive advantages over competitors. However, this does not only go for Formula 1 but, most data-driven sports where knowledge is key and continued advancements are being made. I was able to successfully complete the project with the intended functionality within the specified time frame despite the setbacks and, more importantly, I was able to learn some vital problem-solving and project management skills. This project has made a tremendous impact towards my academic development and future career prospects as it provided me with a platform to bring together all the knowledge and skills I have gained thus far and put them to practice to better understand my abilities and grow as a developer. I believe that this project forms a foundation for future efforts in this topic area and can be valuable to future endeavours aimed at machine learning predictions within the world of sports.

# References

C.Ditcher, M.OReilly, & E.Delahunt. (2021, March 20). *Machine learning in sports science: challenges and opportunities*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/14763141.2021.1910334

C.Garvin, D.Julian, B.Lee, & A.Kosikowski. (2022, May 20). *F1 Race Prediction*. Retrieved from https://andrewkosikowski.com/Formula1/F1RacePredictionFinalReport.pdf

FIA. (2022). *2022 FORMULA ONE SPORTING REGULATION*. Retrieved from https://www.fia.com/sites/default/files/formula\_1\_-\_sporting\_regulations\_-\_2022\_-\_iss\_4\_-\_2022-02-18.pdf

G.Fialho, A.Manhães, & J.P.Teixeira. (2019). *Predicting Sports Results with Artificial Intelligence – A Proposal Framework for Soccer Games*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050919322033

H.Sicoie. (2022, January 14). *MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR*. Retrieved from http://arno.uvt.nl/show.cgi?fid=157635

Huber, G. P. (1991). *Organizational Learning: The Contributing Processes and the Literatures. Organization Science*. Retrieved from http://www.jstor.org/stable/2634941

K.Yilmaz. (2013). *Comparison of Quantitative and Qualitative Research Traditions: epistemological, theoretical, and methodological differences*. Retrieved from https://d1wqtxts1xzle7.cloudfront.net/54217916/Comparison\_of\_Quan\_and\_Quali-libre.pdf?1503479076=&response-content-disposition=inline%3B+filename%3DComparison\_of\_Quantitative\_and\_Qualitati.pdf&Expires=1670769098&Signature=ViaoVbi1FxV-VAFZmoVz~jJ~HPUH4AMAb

Knopf, J. (2006, February 13). *Doing a Literature Review*. Retrieved from https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/abs/doing-a-literature-review/00B62000B6760AB78E1BD27E32A94C9F

M.Hron, & N.Obwegeser. (2018). *Scrum in practice: an overview of Scrum adaptations*. Retrieved from https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/c7dad016-2792-4ec5-96a5-1941a7d43c11/content

M.Lewis. (2004, March 17). *Moneyball: The Art of Winning an Unfair Game*. Retrieved from https://nanopdf.com/download/moneyball-the-art-of-winning-an-unfair-game\_pdf

O.Hubáček, G.Šourek, & F.Železný. (2019). *Exploiting sports-betting market using machine learning*. Retrieved from http://ida.felk.cvut.cz/zelezny/pubs/ijf.2019.pdf

R.Dunton. (2021). *Discussion Section for Research Papers*. Retrieved from https://www.sjsu.edu/writingcenter/docs/handouts/Discussion%20Section%20for%20Research%20Papers.pdf

R.M.Musa, Z.Taha, A.P.P.A.Majeed, & M.R.Abdullah. (2018, September 19). *Machine Learning in Sports: Identifying Potential Archers*. Retrieved from https://link.springer.com/book/10.1007/978-981-13-2592-2

W.George. (2021, August 6). *Formula 1 Championship Predictor*. Retrieved from https://medium.com/@willgeorge93/formula-1-championship-predictor-a-machine-learning-solution-a86efcb9298#9688

# Appendices:

## Figure 1:

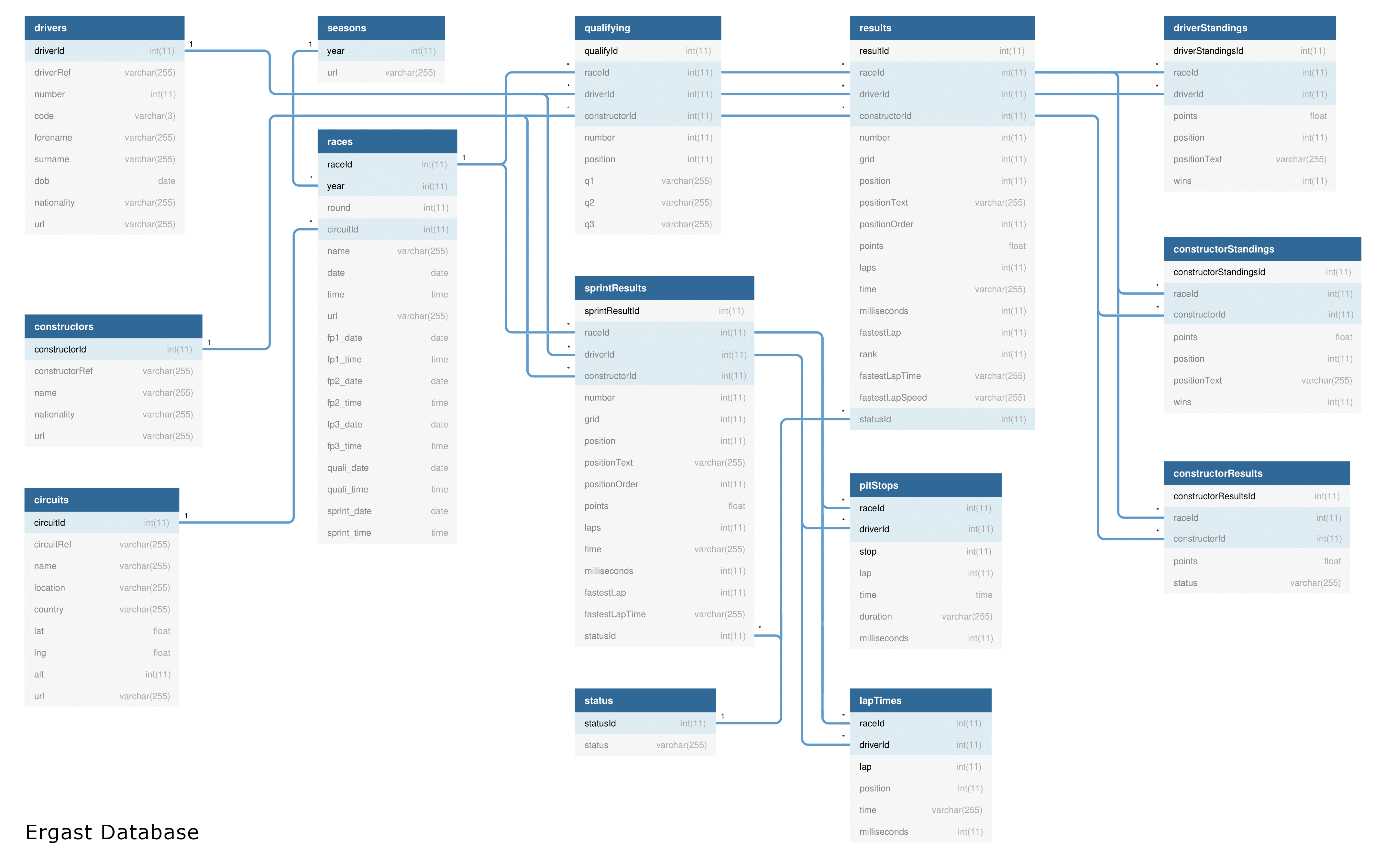


Figure 1: This image depicts the structure of the database used by the Ergast API. For further information about the database please have a look at https://ergast.com/docs/f1db\_user\_guide.txt

## Figure 2:

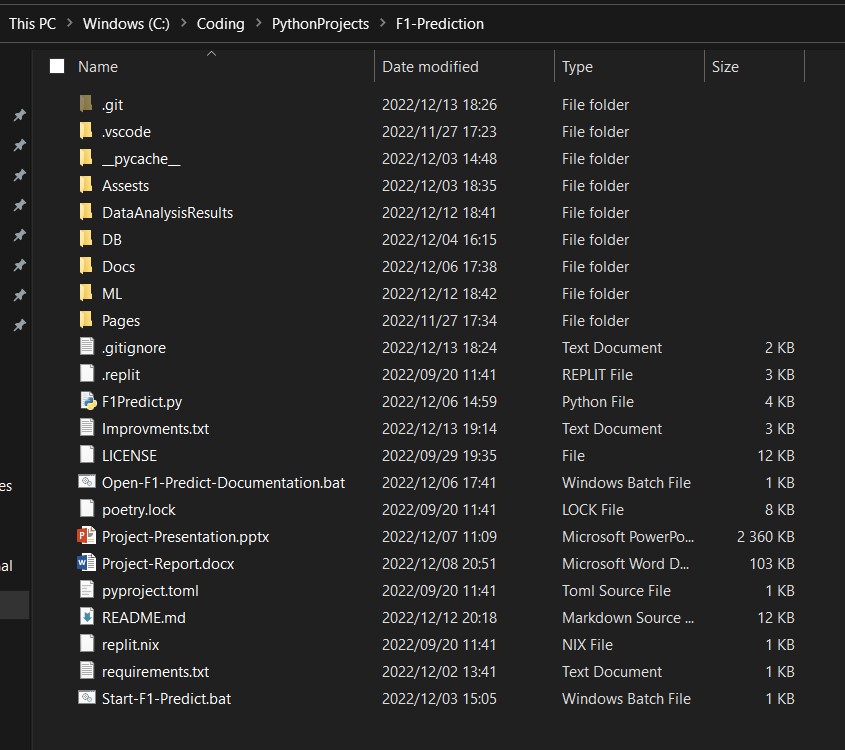


Figure 2: This image is the most recent example of the project's file/directory structure. In this image, you can find all of the contents that make up this project.

## Figure 3:

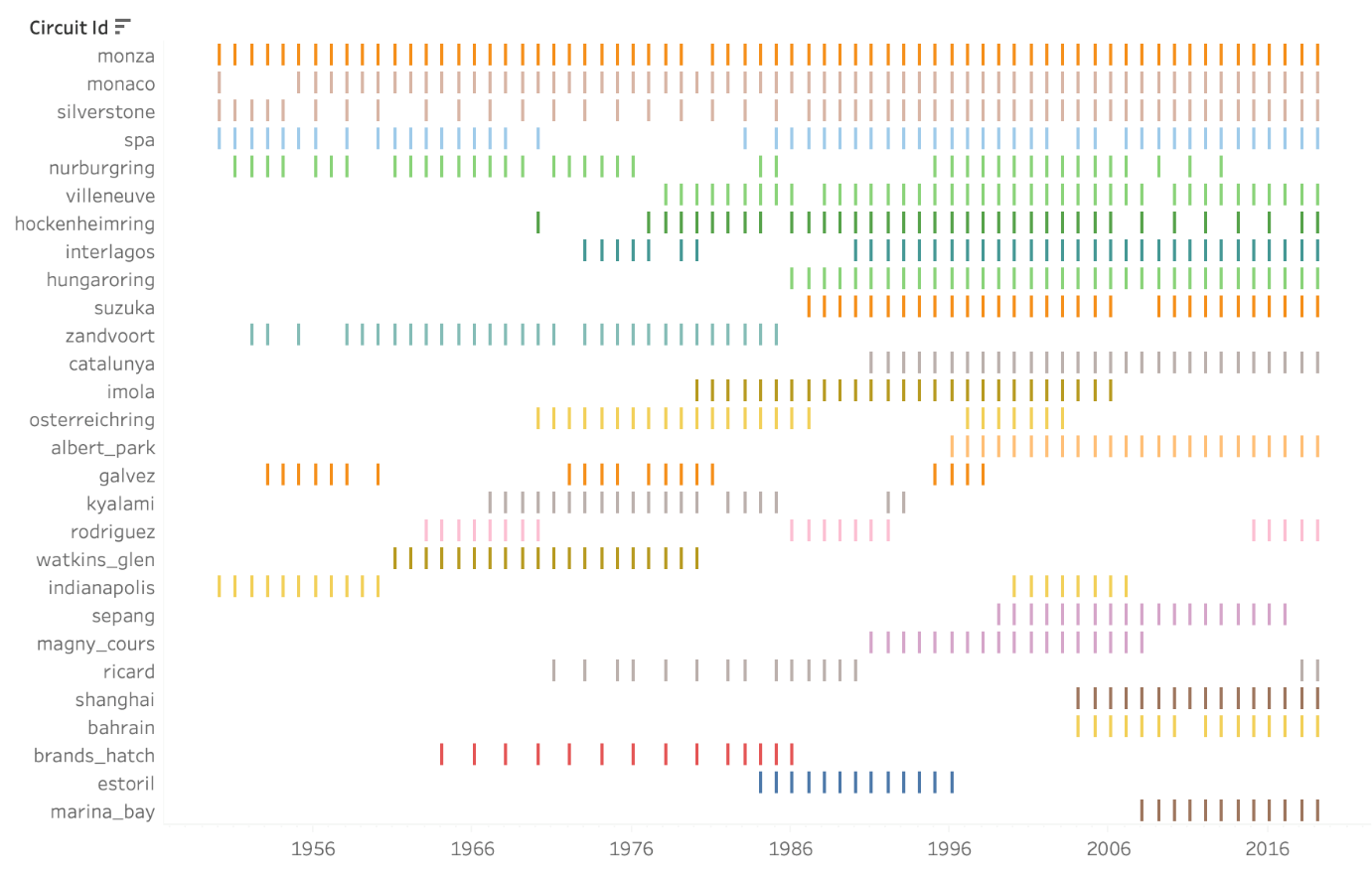


Figure 3: This image shows the most popular circuits over the years in Formula 1.

## Figure 4:

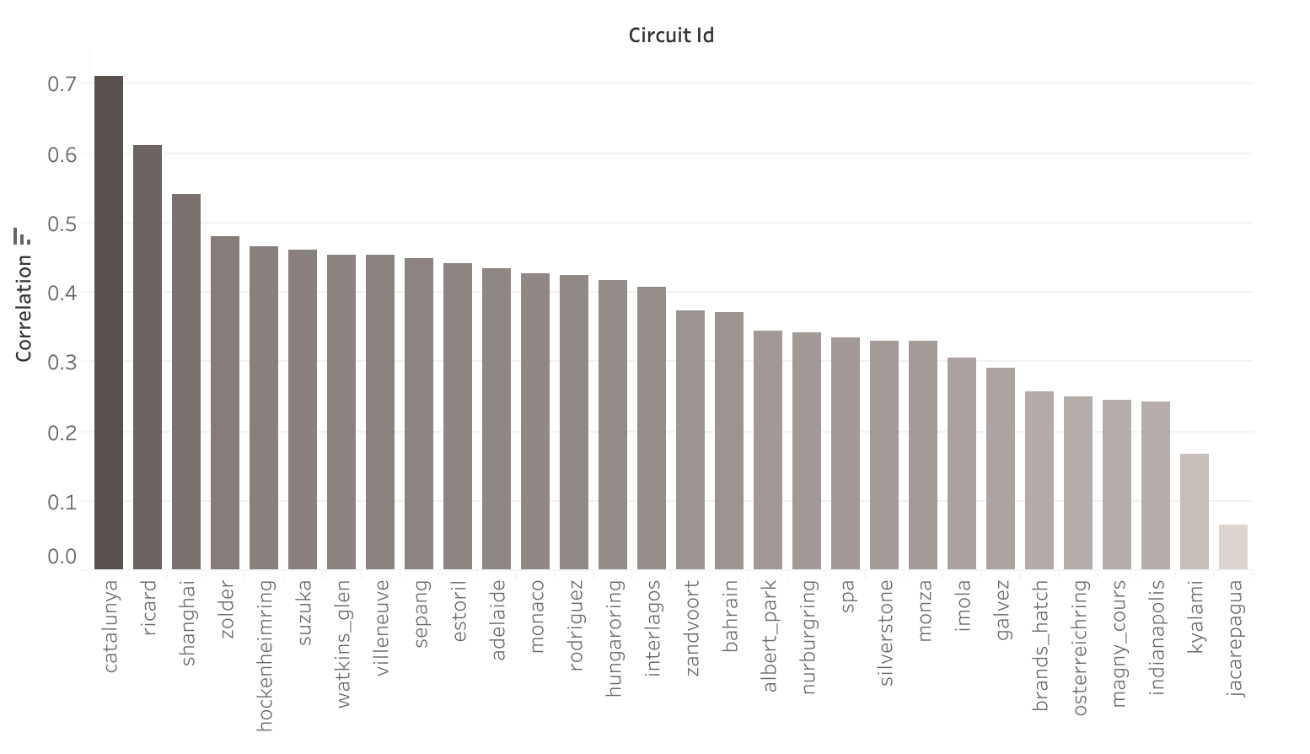


Figure 4: This image shows the correlation between starting in the pole position and winning the race.

## Figure 5:

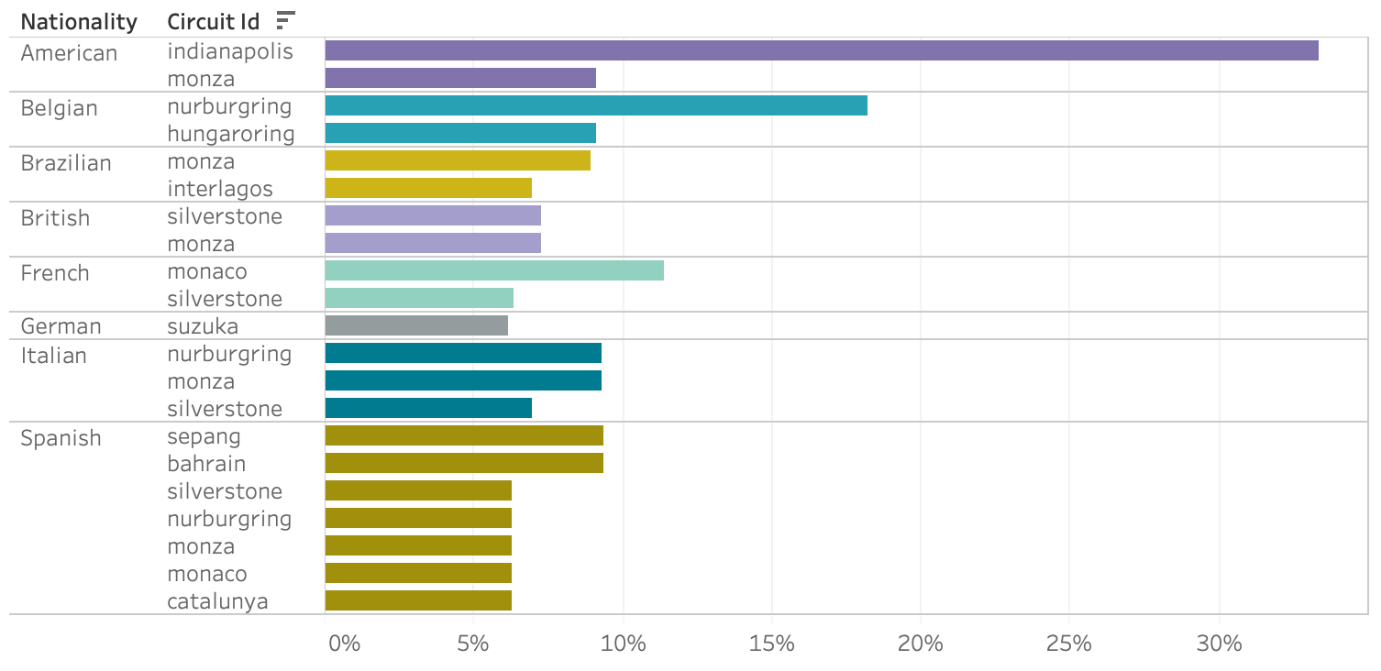


Figure 5: This image shows the correlation between the driver’s nationalities and the country of the races they won.

## Figure 6:

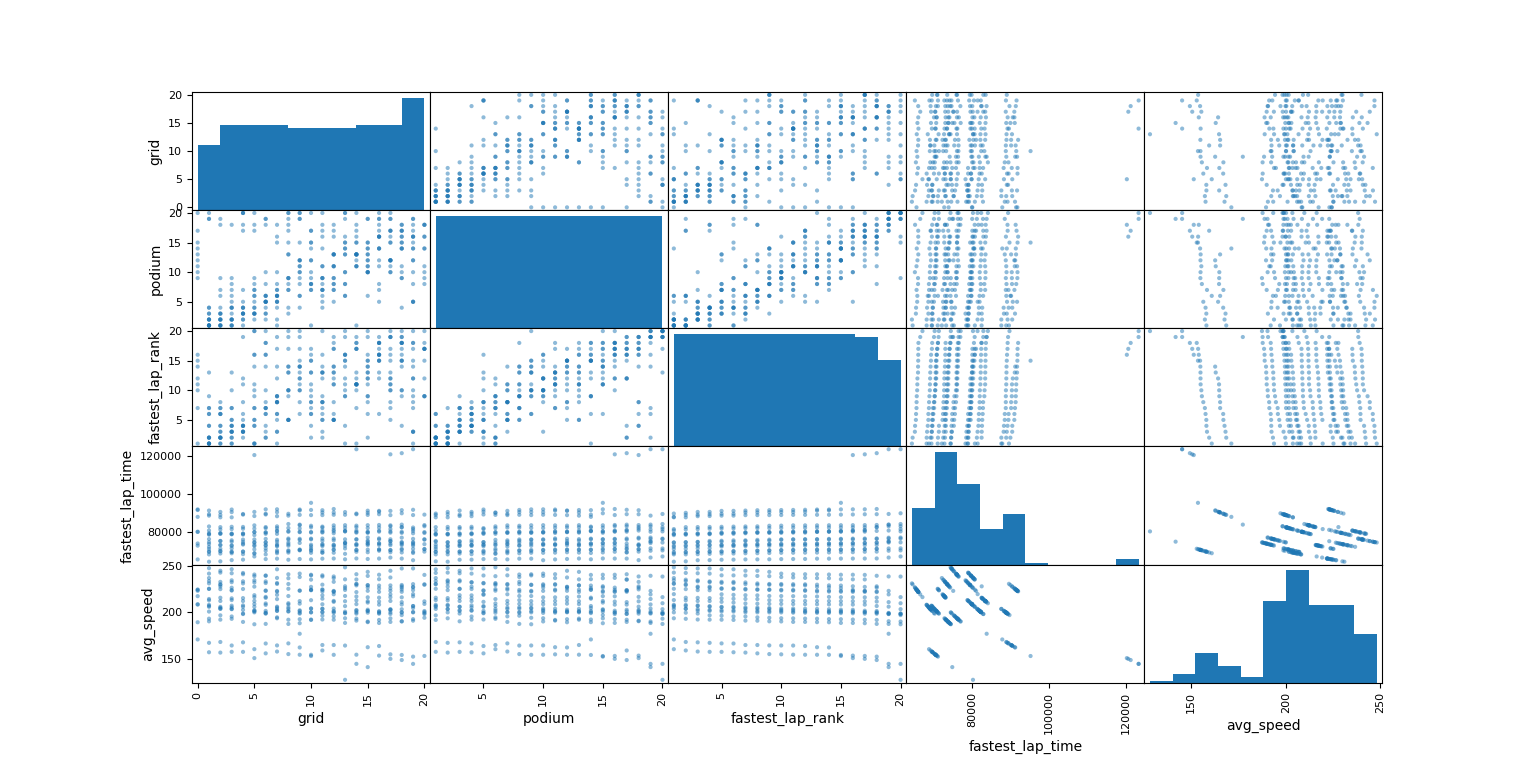


Figure 6: This scatter plot is very effective at depicting the various trends and correlations formed between different features of the dataset. This image is very encouraging as it provides strong evidence that the machine learning algorithm will be able to make informed predictions.

## Figure 7:

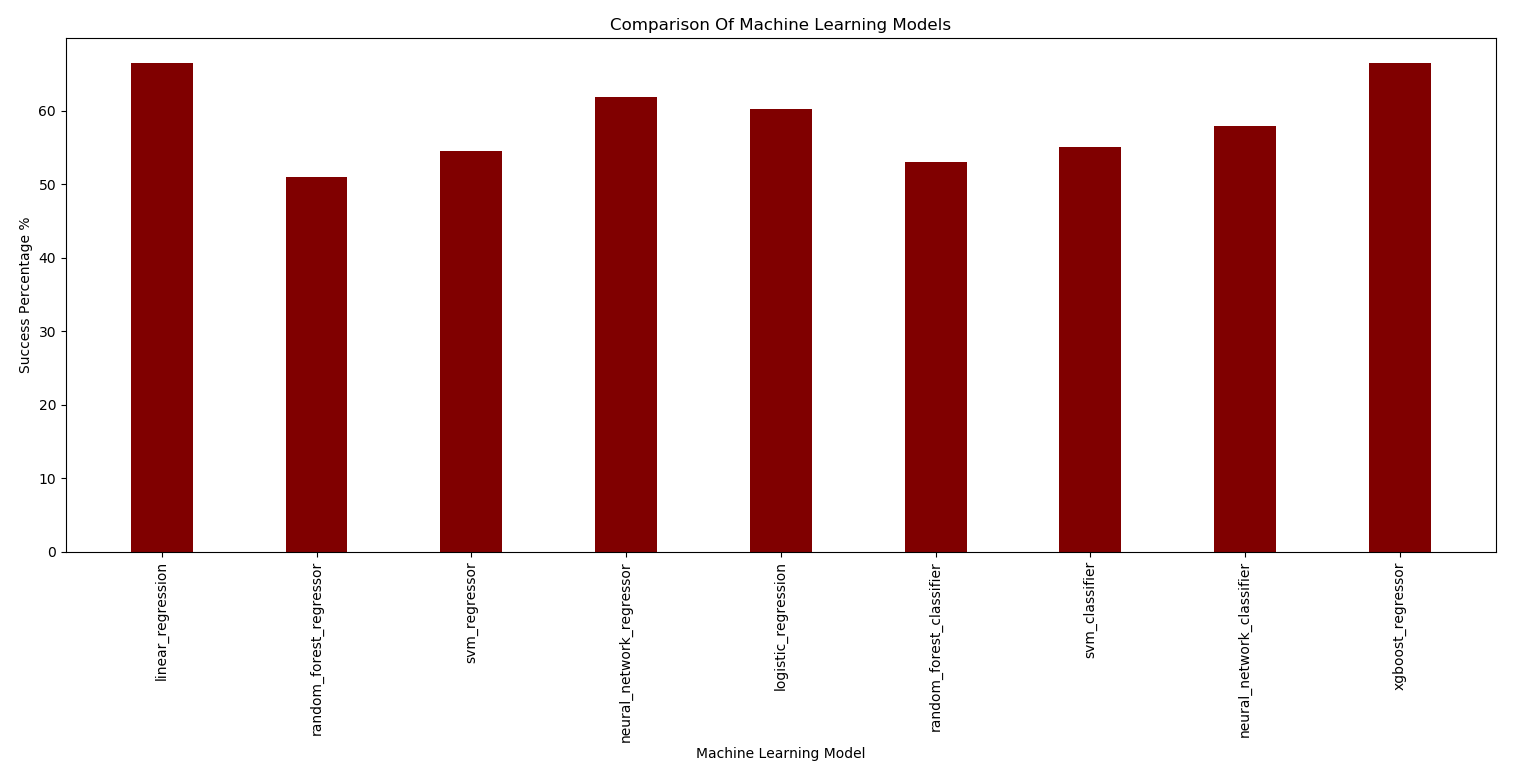


Figure 7: This image displays the accuracy percentages of each machine learning model when predicting the results for each race in the 2022 season. As depicted by the image, both Linear Regression and XGBoost Regressor have tied for first place with both predicting 66.56% of the races accurately.

## Figure 8:

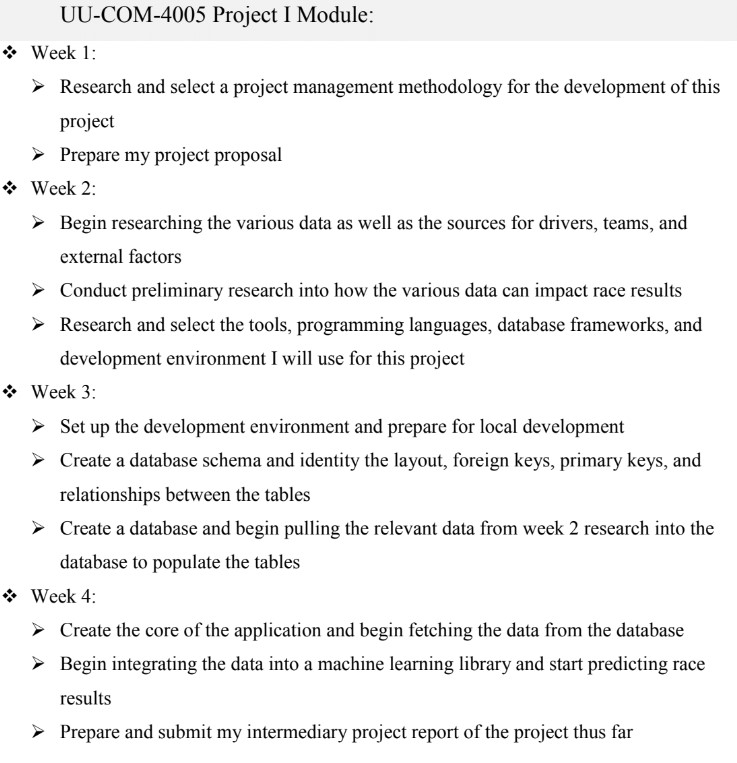


Figure 8: This image depicts the Scrum Weekly Sprint planning for the first phase of this project as well as the goals for each sprint.

## Figure 9:

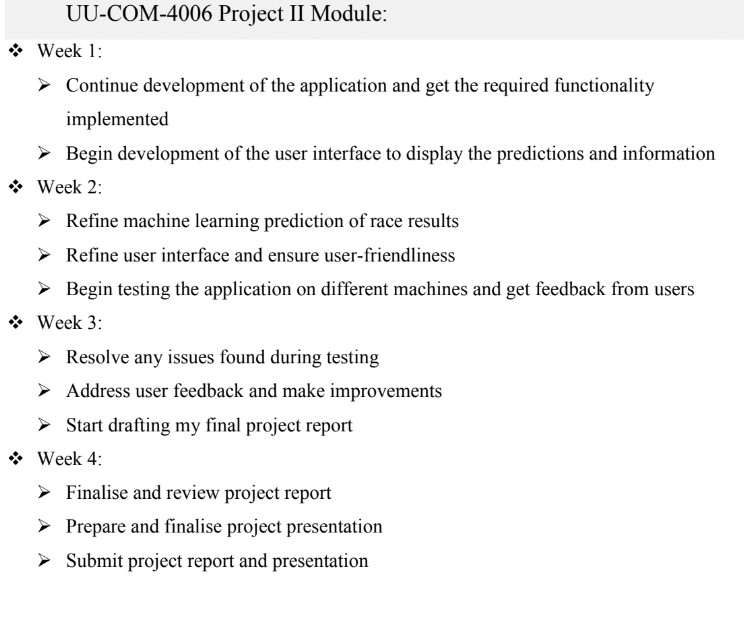


Figure 9: This image depicts the Scrum Weekly Sprint planning for the second (current) phase of this project as well as the goals for each sprint.

## Figure 10:

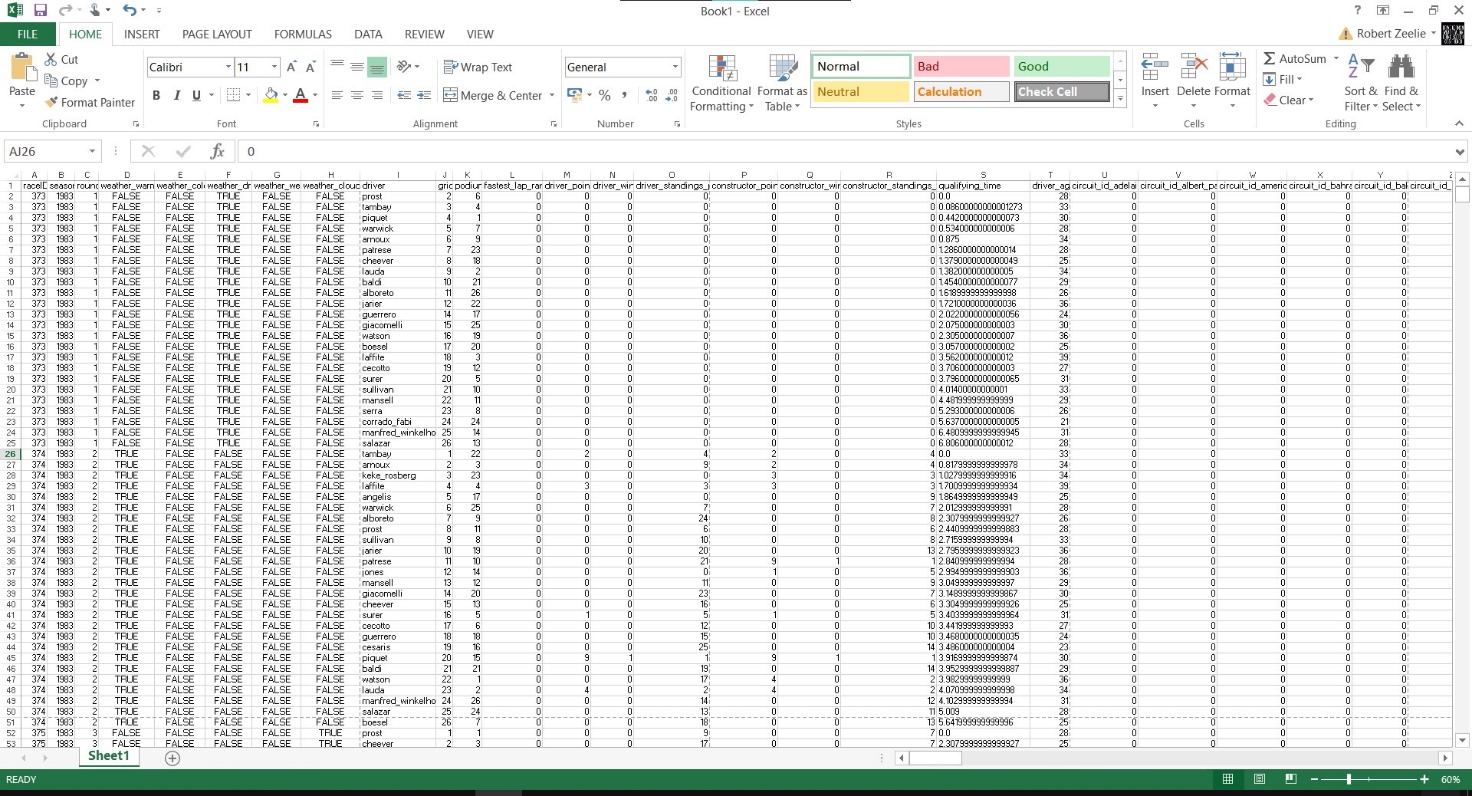


Figure 10: This image is a small partial glimpse into the dataset used to train the machine learning models. This image only captures the first few columns and rows of the entire dataset made of over 15 000 rows and over 300 feature columns but the entire training dataset can be found in the DataAnalysisResults folder at the root of the project.

## Figure 11:

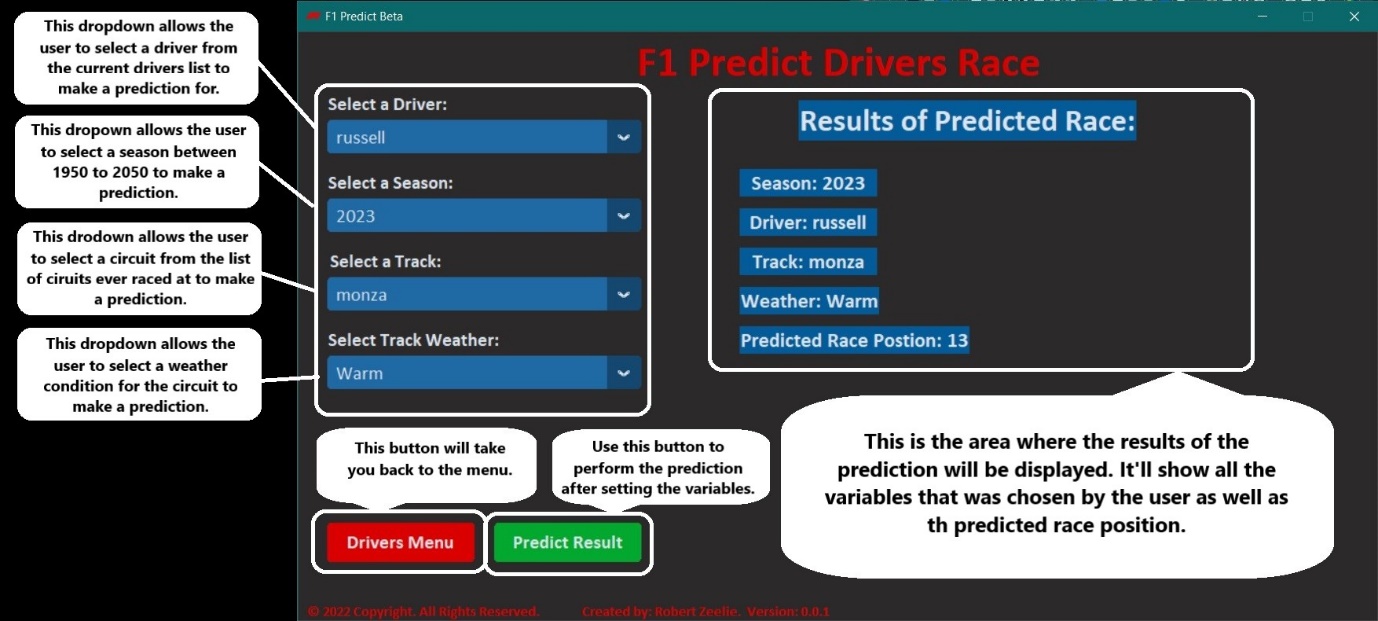


Figure 11: This image illustrates the various components of the ‘Predict Drivers Race’ page. It gives a brief overview of the page and what the different buttons do.

## Figure 12:

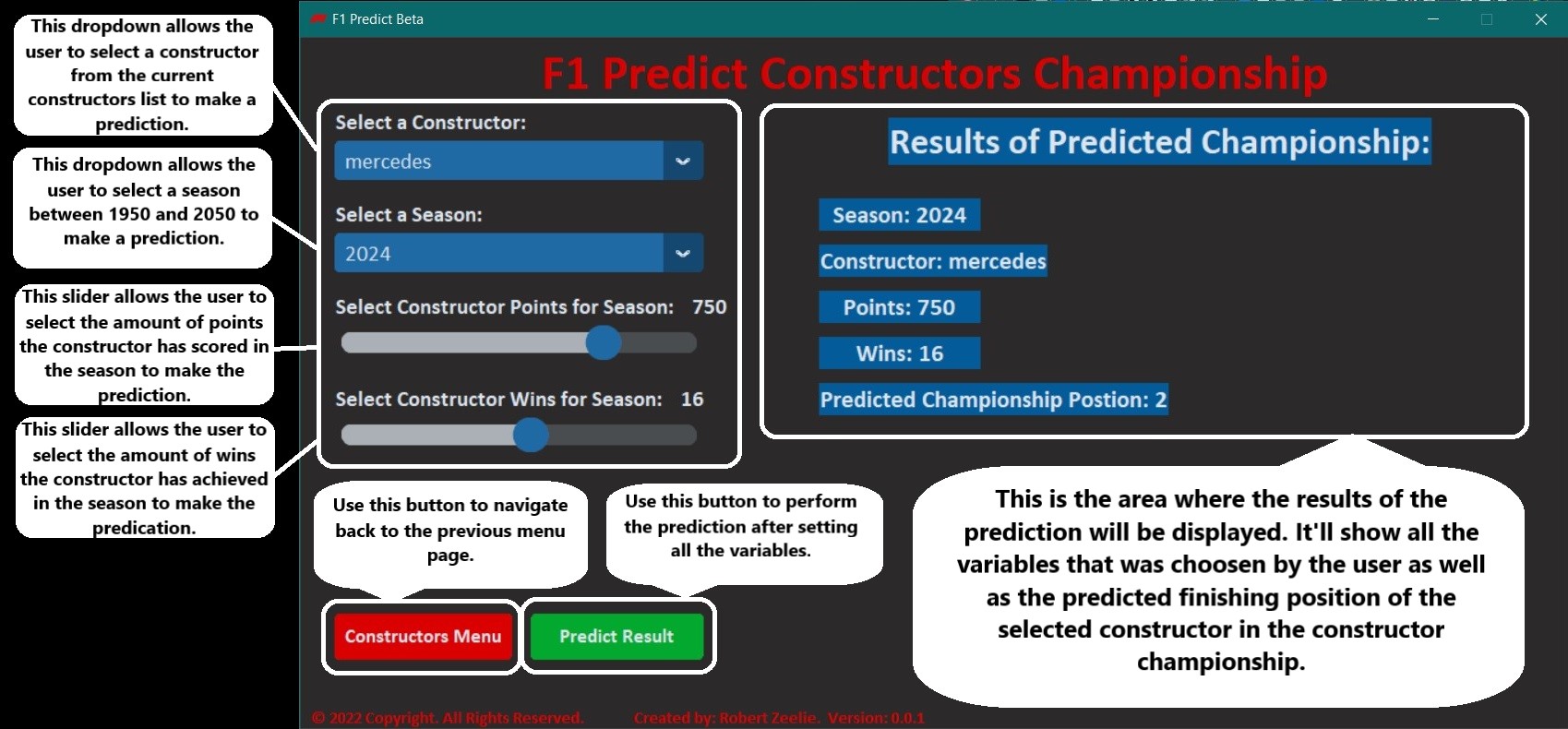


Figure 12: This image illustrates the various components of the ‘Predict Constructors Championship’ page. It gives a brief overview of the page and what the different buttons and sliders do.

## Figure 13:

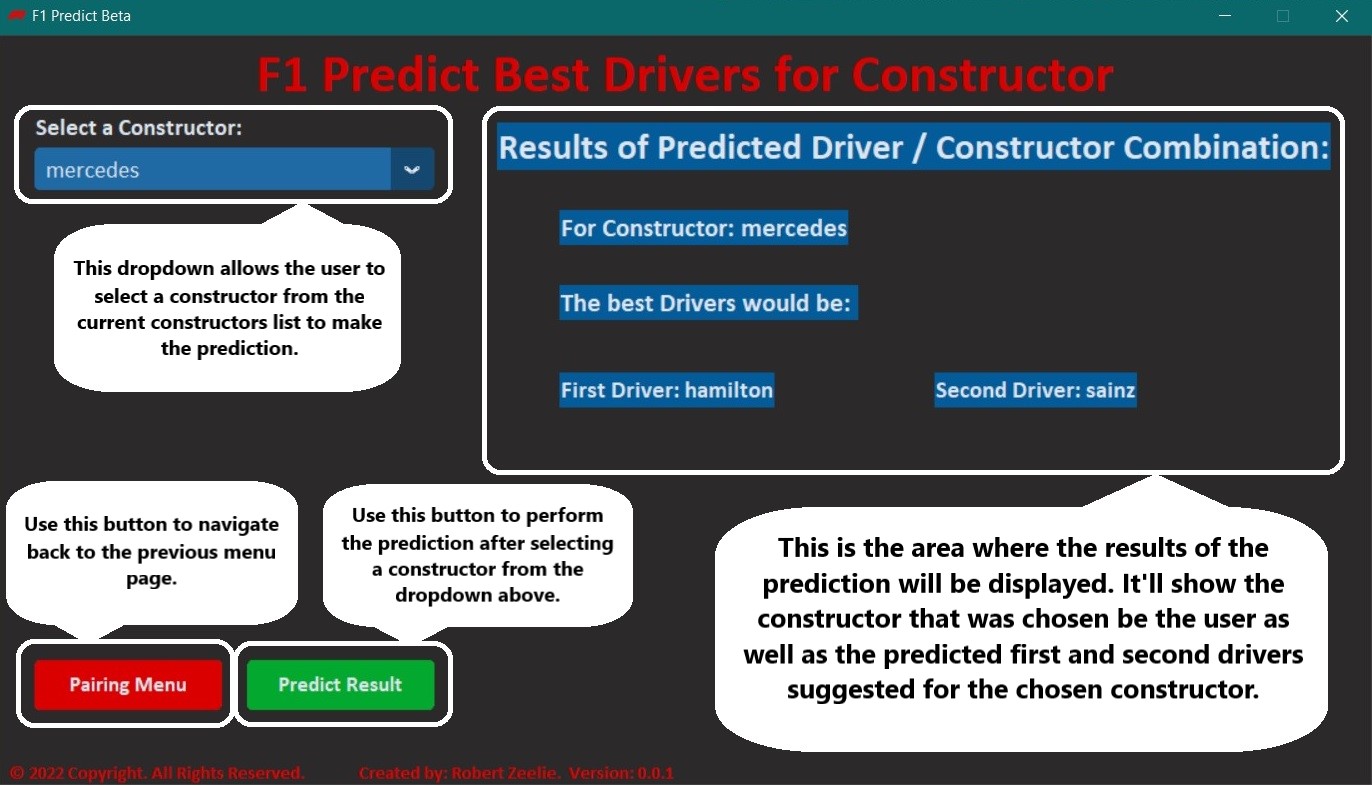


Figure 13: This image illustrates the various components of the ‘Predict Best Drivers for Constructor’ page. It gives a brief overview of the page and what the different buttons do.

## Code Snippet 1:

def dataPreperation():

    # Set global configurations

    Database.connectToDatabase()

    np.set\_printoptions(precision=5)

    # Get all the necessary data

    finalDataframe = getData()

    # Calculate driver ages

    finalDataframe = calculateDriverAge(finalDataframe)

    # Fill or remove nulls from dataset

    finalDataframe = removeOrFillNulls(finalDataframe)

    # Convert weather numbers to booleans

    finalDataframe = convertToBooleans(finalDataframe)

    # Calculate qualifying time differences

    finalDataframe = calculateQualifyingTimeDifference(finalDataframe)

    # Get dummy values and drop non-significante variables

    finalDataframe = getDummyValues(finalDataframe)

    # Close db connections

    Database.disconnectFromDatabase()

    # Return final dataframe

    return finalDataframe

Code Snippet 1: This code snippet is the main function used for data preparation and manipulation. It is called on application start-up in a separate thread and first collects all data from the database, calculates driver age, cleans nulls from data, converts integer values to Boolean for memory usage, calculates qualifying time differences from the leader, and finally gets dummy variables / One-Hot Encoding which turns categorical data into a binary vector representation.

## Code Snippet 2:

def getDummyValues(finalDataframe):

    finalDataframe['driverFeature'] = finalDataframe['driver']

    finalDataframe['constructorFeature'] = finalDataframe['constructor']

    dummyDataframe = pd.get\_dummies(finalDataframe, columns = ['driverFeature', 'circuit\_id', 'country', 'nationality', 'constructorFeature', 'constructor\_nationality'] )

    for column in dummyDataframe.columns:

        if 'driverFeature' in column and dummyDataframe[column].sum() < 25:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        if 'nationality' in column and dummyDataframe[column].sum() < 120:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        elif 'constructorFeature' in column and dummyDataframe[column].sum() < 120:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        elif 'circuit\_id' in column and dummyDataframe[column].sum() < 60:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        elif 'country' in column and dummyDataframe[column].sum() < 100:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        elif 'constructor\_nationality' in column and dummyDataframe[column].sum() < 120:

            dummyDataframe.drop(column, axis = 1, inplace = True)

        else:

            pass

    return dummyDataframe

Code Snippet 2: This is the getDummyValues function which is in charge of One-hot encoding the categorical data into a binary vector representation. Once the variables have been One-Hot Encoded the function proceeds to remove any irrelevant information such as circuit\_id’s that occur in the table less than 60 times.

## Code Snippet 3:

def performLinearRegressionDriver(X\_train, y\_train, X\_test, y\_test):

    model = LinearRegression(fit\_intercept = True)

    model.fit(X\_train, y\_train)

    prediction\_df = pd.DataFrame(model.predict(X\_test), columns = ['results'])

    prediction\_df['driver'] = y\_test.reset\_index(drop = True)

    prediction\_df.sort\_values('results', ascending = True, inplace = True)

    prediction\_df.reset\_index(inplace = True, drop = True)

    prediction\_df['predicted'] = prediction\_df.index

    prediction\_df['predicted'] = prediction\_df.predicted.map(lambda x: x + 1 if x >= 0 else -1)

    return prediction\_df

def performLinearRegressionConstructor(X\_train, y\_train, X\_test, y\_test):

    model = LinearRegression(fit\_intercept = True)

    model.fit(X\_train, y\_train)

    prediction\_df = pd.DataFrame(model.predict(X\_test), columns = ['results'])

    prediction\_df['constructor'] = y\_test.reset\_index(drop = True)

    prediction\_df.sort\_values('results', ascending = True, inplace = True)

    prediction\_df.reset\_index(inplace = True, drop = True)

    prediction\_df['predicted'] = prediction\_df.index

    prediction\_df['predicted'] = prediction\_df.predicted.map(lambda x: x + 1 if x >= 0 else -1)

    return prediction\_df

Code Snippet 3: This Code Snippet depicts the two Linear Regression methods; the first ‘performLinearRegressionDriver’ is used to predict driver results while ‘performLinearRegressionConstructor’ is used to predict constructor results. Both functions take X-axis training data, Y-axis training data, X-axis testing data, and Y-axis testing data. The X-axis contains all of the features that will be used to predict while the Y-axis contains the feature the model should predict. The testing data contains the event data that the function should predict and finally the functions return a new dataframe that contains the predicted results.

## Code Snippet 4:

def performXGBoostDriver(X\_train, y\_train, X\_test, y\_test):

    model = XGBRegressor(random\_state=7)

    model.fit(X\_train,y\_train)

    prediction\_df = pd.DataFrame(model.predict(X\_test), columns = ['results'])

    prediction\_df['driver'] = y\_test.reset\_index(drop = True)

    prediction\_df.sort\_values('results', ascending = True, inplace = True)

    prediction\_df.reset\_index(inplace = True, drop = True)

    prediction\_df['predicted'] = prediction\_df.index

    prediction\_df['predicted'] = prediction\_df.predicted.map(lambda x: x + 1 if x >= 0 else -1)

    return prediction\_df

def performXGBoostConstructor(X\_train, y\_train, X\_test, y\_test):

    model = XGBRegressor(random\_state=7)

    model.fit(X\_train,y\_train)

    prediction\_df = pd.DataFrame(model.predict(X\_test), columns = ['results'])

    prediction\_df['constructor'] = y\_test.reset\_index(drop = True)

    prediction\_df.sort\_values('results', ascending = True, inplace = True)

    prediction\_df.reset\_index(inplace = True, drop = True)

    prediction\_df['predicted'] = prediction\_df.index

    prediction\_df['predicted'] = prediction\_df.predicted.map(lambda x: x + 1 if x >= 0 else -1)

    return prediction\_df

Code Snippet 4: This Code Snippet depicts the two XGBoost Regressor methods; the first ‘performXGBoostDriver’ is used to predict driver results while ‘performXGBoostConstructor’ is used to predict constructor results. Both functions take X-axis training data, Y-axis training data, X-axis testing data, and Y-axis testing data. The X-axis contains all of the features that will be used to predict while the Y-axis contains the feature the model should predict. The testing data contains the event data that the function should predict and finally the functions return a new dataframe that contains the predicted results.