

PROJECT TITLE: SOFTWARE HSC MAJOR PROJECT REPORT

STUDENT NAME: Aahan Dharammali

SUBMISSION DATE: 4/7/25

COURSE: Software Engineering Stage 6

GITHUB URL: <https://github.com/Aahan-dev/HSC-Major-Project-F1-Race-Win-Predictor>

WEBSITE URL: <https://hsc-major-project-f1-race-win-predictor-aahan-dharammali.streamlit.app/>



Table of Contents:

- 1 Identifying and Defining
- 2 Research and Planning
- 3 System Design
- 4 Producing and Implementing
- 5 Testing and Evaluation
- 6 Client Feedback and Reflection
- 7 Appendices

1. Identifying and Defining

1.1

- The Formula 1 fans are affected by my software engineering solution as it develops a predictive model for Formula 1 race outcomes, addressing the unpredictability of the sport. With many variables affecting race results such as driver performance, weather conditions and track characteristics traditional forecasting methods often fall short.

1.2

- This project will serve the purpose of enhancing fan engagement as with this application fans can see the predicted outcomes and insights for their favourite teams and drivers. Hence this identifies and addresses an opportunity through allowing fans an opportunity to have insights into a race before it occurs.

1.3

- The stakeholders in this project are the F1 community and peers who have reviewed the app.

1.4

Functional Requirements:

- Allow user to pick race season
- The application or system must allow a user to select a race track and year
- The system must predict the most likely race winner using a machine learning model
- The system must display prediction results in an easy-to-read format
- The system must provide historical winner data for selected track and year

1.5

Non-Functional Requirements:

- Fast loading times
- Performance of the application should be acceptable
- Application allows many concurrent users to access web page

1.6

- The main constraints on this project were time and technical limits in terms of complexity. With a short time frame of 10 weeks for completion, the project had to be handled with efficiency and have a sufficient level of complexity that was expected of a machine learning model. The options for the model were to make a linear regression model, decision tree, random forest and neural network. A Linear regression model would have easily fit within the given timeframe, however lacked complexity, challenge and an ability to account for the many different variables found within Formula 1. Random forests and neural networks would have exceeded both the complexity required of this project and also require much longer to make. Hence the decision tree was the best choice given that it acknowledged both of the constraints effectively.

2. Research and Planning

2.1

- My chosen approach to this project was the agile development process because of its scalability in a time bound environments and its project management features such as the following:
- **Iterative development:** Break the project into smaller, manageable iterations (sprints), each focusing on delivering specific features or functionalities within a defined time.
- **Sprint Reviews:** At the end of each development sprint hold a review to re-evaluate position in the project timeline, what went well and what could be improved, etc.
- **Use Time-Boxed Sprints:** Organise work into fixed-duration sprints to make sure that the project meets deadlines.

2.2

- **Language:** Python 3.10
- **IDE Used:** Visual Studio Code
- **Libraries Used:** Streamlit,Pandas,Sklearn.tree
- **Framework Used:** Agile Framework

2.3

Stage	WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9	WK10
Planning										
Deciding on features										
UI design										
Adding functionality to UI										
Coding UI										
Debugging UI										
Deciding how to make the prediction algorithm										
Coding decision algorithm										
Testing										
Improving decision algorithm										
Theory (report)										

The GANTT chart was used for its simplicity as this was a time-bound project. Along with this, the GANTT chart made it very easy to check if the project was on track. The development was approached by using the agile development process as it was a short term project. This was used by implementing quick development sprints and prioritising time for testing, allowing for continuous improvement.

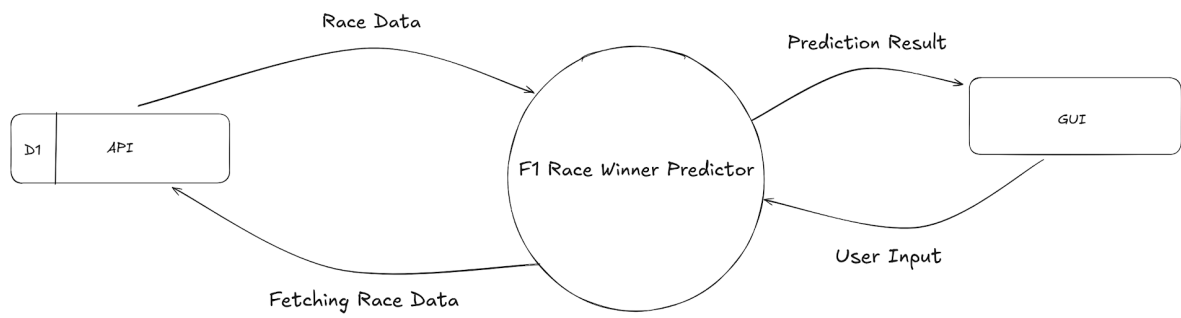
2.4

- To engage with my client or simulate feedback for my F1 win predictor project, I plan to conduct surveys to gather opinions from my classmates, perform user testing and hold feedback sessions to discuss ideas with peers.

3. System Design

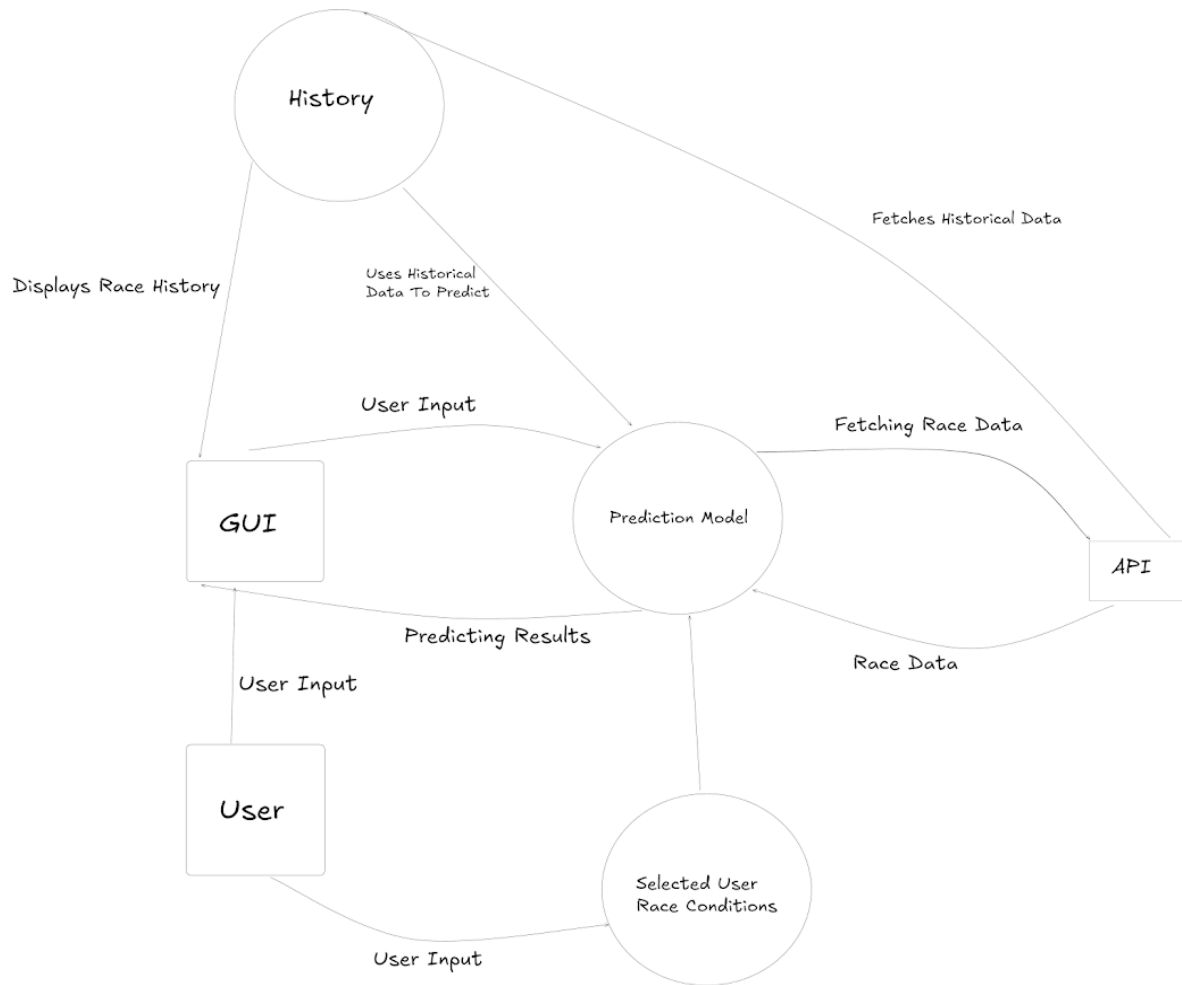
3.1

- Context Diagram (LVL 0 Data Flow Diagram)



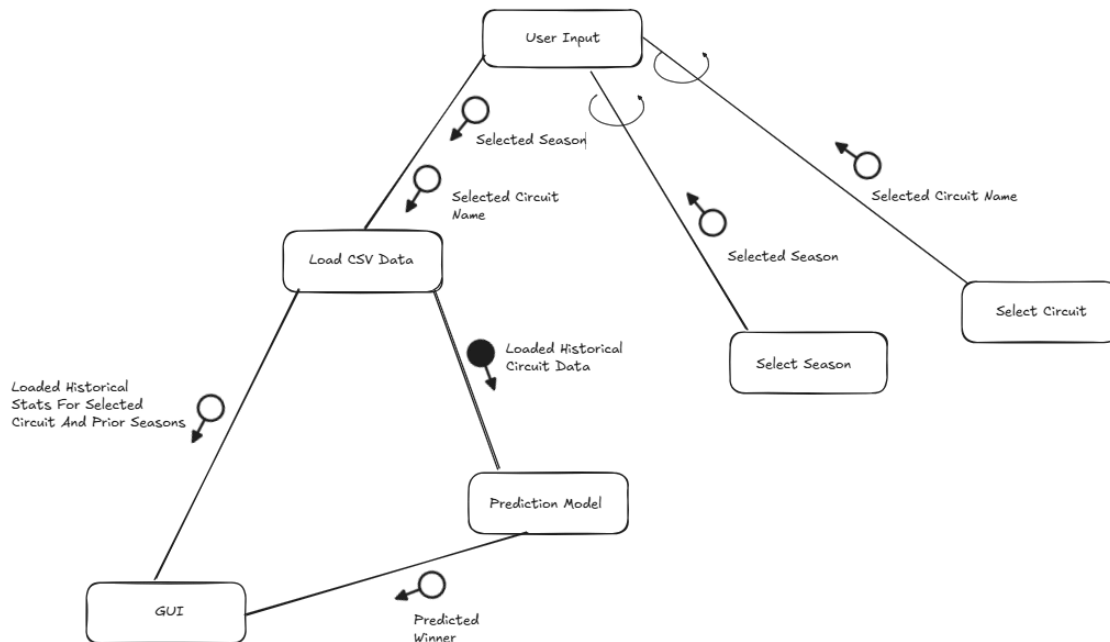
3.2

- LVL 1 Data Flow Diagram



3.2

- Structure Chart



3.3

- IPO Chart

Module	Inputs	Processes	Output
Main UI	<ul style="list-style-type: none"> - User selections: Race, Year, Qualifying checkbox 	<ul style="list-style-type: none"> - Display UI elements and get inputs - Trigger prediction function - Load and clean CSV data - Encode categorical features - Train ML model 	<ul style="list-style-type: none"> - Selected inputs - Show prediction and historical data
Predict Winner	<ul style="list-style-type: none"> - Race name (track_name) - Year - Include qualifying flag - CSV data file 	<ul style="list-style-type: none"> - Filter race data for prediction - Calculate probabilities - Pick top driver prediction 	<ul style="list-style-type: none"> - Predicted driver name - Winning probability (as string)

3.4

- Data Dictionary

Field Name	Data Type	Size/Format	Description	Example Value	Constraints/ Validation Rules
Season	Integer	4 digits	The year in which the Formula 1 race took place	2025	Integers only and must be a valid year
Round	Integer	2 digits at max	The race round number in the season	9	Must not be larger than the amount of races in a calendar year
Circuit	Text/String	Full name of the circuit	Name of the race circuit where the event takes place.	Albert Park Circuit	Must be in the listed races for that year
Driver	Text/String	Last name of Driver	Name of the Formula 1 driver competing in the race.	Hamilton	Must be in the 20 possible drivers to win
Grid	Integer	Integer between 1-20	The driver's starting position on the grid before the race.	12	Must be between 1-20
Constructor	Text/String	Team name	The name of the driver's team	Ferrari	Must be one of the 10 possible teams
Won	Integer	1 digit	Indicates whether the driver won the race.	1 = Yes, 0 = No, null = Future race.	Must either be 0,1 or null

Weather	Text/String	Possible weather values	Weather condition during the race	"Sunny", "Overcast", "Rainy"	Must be either sunny, overcast or rainy
constructor_strength	Integer	1 digit	A number assigned to each constructor based on category encoding.	2	Must be ≥ 0
weather_code	Integer	1 digit	A code for weather:	0 = Sunny, 1=Overcast, 2 = Rainy.	Must be between 0-2
win_prob	Float	4 digits at max	Predicted probability that a driver will win the race	24.5%	Can never be 100%

4. Producing And Implementing

4.1

- The first step as outlined in figure 2.3 was planning what the project was going to be, then deciding on features and designing what the UI will look like. Next came the development that was originally approached using Figma, which is a tool that allows for drag and drop elements into your UI. However I later opted for streamlit as it was much faster to make and had a free tier cloud app hosting capabilities. After the building and then debugging of the UI, I had to decide what prediction model would be used. The decision I came to was that a decision tree algorithm would effectively address both the difficulty and time constraints for this project as further elaborated in 1.6. Then the prediction model was programmed and tested using various methods, after this the last step was to improve the model where possible and polish the UI.

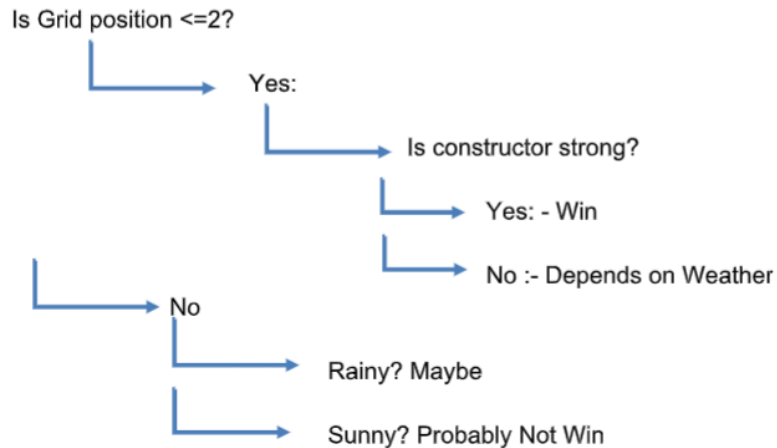
4.2

The core features of the program are:


- Race winner prediction feature**
Justification: This allows for personalised predictions based on specific contexts, enhancing user engagement. Different tracks and years may have varying historical data, affecting the accuracy of predictions.
- Historical stat fetcher**
Justification: Historical stats aid the user in giving information about past winners and team performances at certain circuits.

Example of how the decision tree works in my project


1. Training the model based on extracted data
2. The decision tree learns rule based on the past races
 - a. If Grid =1 and weather = Sunny and Constructor = Mercedes – Win likely
 - b. If Grid > 5 and weather = Rainy and Constructor = Haas – Not likely
3. When I add new data from qualifying results the model,
 - a. Looks at the grid, constructor strength, and weather
 - b. Follows the decision tree path to reach a prediction
 - c. Returns the result: “Win” or “No Win”, and a probability like 24%



4.3

 **Race Winner Predictor**

Powered by data. Inspired by speed.

 **Race Settings**

Select Race Location:

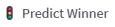
Albert Park Grand Prix Circuit


Select Race Year:

2020


☐ Include Qualifying Data

Selected Race: Albert Park Grand Prix Circuit

 Predict Winner

 **Prediction Panel**

Prediction will appear here once inputs are provided.

 **Historical Stats**

- 2025: Norris
- 2024: Sainz
- 2023: Verstappen

Weather Impact Insights:

Rainy races historically increase unpredictability by 32%.

The main menu allows users to select a race track, the race year and has a button to predict the winner.

5. Testing And Evaluation

5.1

The following methods of testing were used throughout the project:

- **Unit Testing:** Unit testing was used throughout the project to ensure that certain parts of the application were working as expected before integration into the system. For example the UI code was tested separately from the prediction model to ensure that both components were working as expected before integrating the model into the UI
- **User Testing:** To accomplish user testing friends and family were asked to test the application for any outstanding bugs or issues.
- **Regression Testing:** Regression testing was utilised when features needed to be added to the model and the UI. For example when adding the historical stat fetcher, I needed to ensure that the prior prediction feature wasn't affected.

5.2 Test Cases

Test ID	Description	Expected Result	Actual Result	Pass/Fail
CSVSTST01	Testing if application will properly take take info from CSV file	Success Message	UTF error message (because some drivers and tracks have accents in their names)	Fail
GPTST01	Testing if the application can correctly predict the Canada race	Successful prediction	Successful prediction	Pass
GPTST02	Testing if the application can correctly predict the Spain race	Successful Prediction	Successful prediction	Pass
GPTST03	Testing if the application can correctly predict the Austria race	Successful prediction	Successful prediction	Pass
UITST01	Testing if the UI works as expected	Error Message	Error with configuring the page and rendering elements correctly	Fail
UITST02	Testing if the "include qualifying" checkbox works	Checkbox works	Checkbox doesn't work and includes qualifying data regardless	Fail

- This CSV file is generated using data from an API (ergast) from the period of 2020-2024 and stored in the local directory. The API doesn't provide data for the 2025 season. Hence the 2025 data has been manually entered in the CSV file.
- After fetching the data there were anomalies that were fixed by cleaning such as changing the tracks with accents into normal letters so the application can read the CSV file properly, as shown below

228	226	2020	12	Autódromo Internacional do Algarve	Perez	5	Racing Point	0	Sunny		
229	227	2020	12	Autódromo Internacional do Algarve	Ocon	11	Renault	0	Sunny		
230	228	2020	12	Autódromo Internacional do Algarve	Ricciardo	10	Renault	0	Sunny		
231	229	2020	12	Autódromo Internacional do Algarve	Vettel	15	Ferrari	0	Sunny		
232	230	2020	12	Autódromo Internacional do Algarve	Raikkonen	16	Alfa Romeo	0	Sunny		
233	231	2020	12	Autódromo Internacional do Algarve	Albon	6	Red Bull	0	Sunny		
234	232	2020	12	Autódromo Internacional do Algarve	Norris	8	McLaren	0	Sunny		
235	233	2020	12	Autódromo Internacional do Algarve	Russell	14	Williams	0	Sunny		
236	234	2020	12	Autódromo Internacional do Algarve	Giovinazzi	17	Alfa Romeo	0	Sunny		
237	235	2020	12	Autódromo Internacional do Algarve	Magnussen	19	Haas F1 Team	0	Sunny		
238	236	2020	12	Autódromo Internacional do Algarve	Grosjean	18	Haas F1 Team	0	Sunny		
239	237	2020	12	Autódromo Internacional do Algarve	Latifi	20	Williams	0	Sunny		
240	238	2020	12	Autódromo Internacional do Algarve	Kvyat	13	AlphaTauri	0	Sunny		
241	239	2020	12	Autódromo Internacional do Algarve	Stroll	12	Racing Point	0	Sunny		
242	240	2020	13	Autodromo Enzo e Dino Ferrari	Hamilton	2	Mercedes	1	Sunny		
243	241	2020	13	Autodromo Enzo e Dino Ferrari							
244	242	2020	13	Autodromo Enzo e Dino Ferrari							
245	243	2020	13	Autodromo Enzo e Dino Ferrari							
246	244	2020	13	Autodromo Enzo e Dino Ferrari							
247	245	2020	13	Autodromo Enzo e Dino Ferrari							
248	246	2020	13	Autodromo Enzo e Dino Ferrari							
249	247	2020	13	Autodromo Enzo e Dino Ferrari							
250	248	2020	13	Autodromo Enzo e Dino Ferrari							
251	249	2020	13	Autodromo Enzo e Dino Ferrari							
252	250	2020	13	Autodromo Enzo e Dino Ferrari							

Find and Replace

Find Replace

Find what: Autódromo Internacional do Algarve

Replace with: Autodromo Internacional do Algarve


Options >>

Replace All Replace Find All Find Next Close

- Changing driver names with accents into normal letters and changing incorrect names to proper ones so the application can read the CSV file. Names like “Perez” and “Hulkenberg” had to be changed as the API had some corrupt values for certain people's names and certain circuits.

1	season	round	circuit	driver	grid	constructor	won	weather
2	2023	1	Bahrain International Circuit	Verstappen	1	Red Bull	1	Sunny
3	2023	1	Bahrain International Circuit	PÃ©rez	2	Red Bull	0	Sunny
4	2023	1	Bahrain International Circuit	Alonso	5	Aston Martin	0	Sunny
5	2023	1	Bahrain International Circuit	Sainz	4	Ferrari	0	Sunny
6	2023	1	Bahrain International Circuit	Hamilton	7	Mercedes	0	Sunny
7	2023	1	Bahrain International Circuit	Stroll	8	Aston Martin	0	Sunny
8	2023	1	Bahrain International Circuit	Russell	6	Mercedes	0	Sunny
9	2023	1	Bahrain International Circuit	Bottas	12	Alfa Romeo	0	Sunny
10	2023	1	Bahrain International Circuit	Gasly	20	Alpine F1 Team	0	Sunny
11	2023	1	Bahrain International Circuit	Albon	15	Williams	0	Sunny
12	2023	1	Bahrain International Circuit	Tsunoda	14	AlphaTauri	0	Sunny
13	2023	1	Bahrain International Circuit	Sargeant	16	Williams	0	Sunny
14	2023	1	Bahrain International Circuit	Magnussen	17	Haas F1 Team	0	Sunny
15	2023	1	Bahrain International Circuit	de Vries	19	AlphaTauri	0	Sunny
16	2023	1	Bahrain International Circuit	HÃ¼lkenberg	10	Haas F1 Team	0	Sunny
17	2023	1	Bahrain International Circuit	Zhou	13	Alfa Romeo	0	Sunny
18	2023	1	Bahrain International Circuit	Norris	11	McLaren	0	Sunny
19	2023	1	Bahrain International Circuit	Ocon	9	Alpine F1 Team	0	Sunny
20	2023	1	Bahrain International Circuit	Leclerc	3	Ferrari	0	Sunny
21	2023	1	Bahrain International Circuit	Piastri	18	McLaren	0	Sunny
22	2023	2	Jeddah Corniche Circuit	PÃ©rez	1	Red Bull	1	Sunny

Successful Predictions made before the Canada and Spain race as can be seen below



Race Winner Predictor

Powered by data. Inspired by speed.

Deploy

Race Settings

Select Race Location:

Circuit Gilles Villeneuve

Select Race Year:

2025

☐ Include Qualifying Data

Selected Race: Circuit Gilles Villeneuve

Predict Winner

Prediction Panel

Predicted Winner: Russell (Win chance: 24.00%)

Historical Stats

- 2024: Verstappen
- 2023: Verstappen
- 2022: Verstappen

Weather Impact Insights:

Rainy races historically increase unpredictability by 32%.

Above is the prediction for the Canada race made on the 15/6/25

Powered by data. Inspired by speed. Deploy

Race Settings

Select Race Location:

Circuit de Barcelona-Catalunya

Select Race Year:

2025

☒ Include Qualifying Data

Selected Race: Circuit de Barcelona-Catalunya

Predict Winner

Prediction Panel

Predicted Winner: Piastri (Win chance: 24.00%)

Historical Stats

- 2024: Verstappen
- 2023: Verstappen
- 2022: Verstappen

Weather Impact Insights:

Rainy races historically increase unpredictability by 32%.

Above is the prediction for the Spain race made on the 1/6/25

5.3

- My solution effectively reaches the goals previously defined as it allows users to select the race circuit and season. It accurately predicts the winning driver for an upcoming race given that the qualifying data is provided and shows past winners at the same circuit accurately, hence the software solution addresses its goals

5.4

- If I had more time I would have taken more variables into consideration for the prediction model, such as length and number of pit stops during the races, tyre type (hard, soft etc) and driver rankings to produce a more accurate model which would have reduced the biasing of the model.

6. Feedback And Reflection

6.1

- "The ability to include qualifying data makes the predictions feel more tailored and realistic." - Aryan Joshi
- "The layout is intuitive and easy to navigate. I love how the predictions are clearly displayed!" - Akshat Khurana
- "Consider adding a feature for users to compare predictions across different tracks or seasons." - Atharva Malik

6.2

- Throughout the project I learned the following:
 - The agile process that was used in the development of this application
 - How to design and develop a user-friendly interface
 - How to extract and clean larger amounts of data
 - How machine learning works and how to use it
 - How biases work within the context of machine learning
 - How different testing methodologies work and are used (Unit, Regression, User)
 - The Importance of planning a project using the steps outlined in the SDLC
 - How to deploy an application to a public cloud as a finished product
- In this project I learned skills relating to the concept of machine learning and how the behaviour of machine learning is driven through the manipulation of data and how to manage my time when developing a larger scale project.