BAWKOS PROJECT PLAN

Predicting Professional Football Match Outcomes with Artificial Intelligence

Abstract

'Football has many variables which contribute to the outcome of the result of a match. To identifying the likelihood of a team winning, losing, or drawing seems like it is luck if the correct outcome is selected. However, analysis of statistics shows that there is more to luck when predicting the outcome of a game – although it will always be impossible to have 100% certainty of the outcome.'

Team BAWKOS

University of Surrey

Jamie Wiggins, Joseph Kutler, Nkem Ogosi, Manraj Sidhu, Syed Ali Raza, Stefano Bussandri

Contents

1 Project Overview

1.1 Document Change Control

Current Version			
Version Number	Approval date as of	Printed on	Author
2.0	28/11/21	N/A	Team BAWKOS

Revision History			
Version Number	Date	Summary of Changes	Comments
0.1	29/10/21	Setup	Initial setup
0.2	02/11/21	Draft	First draft content added
0.3	04/11/21	Second Draft	Expanded upon first draft content after review
1.0	05/11/21	Final Draft	Finalised project plan following secondary review and formatted document
1.1	08/11/21	Update Project Vision/Mission, Objectives and Scope	No player data available in data sets. Project vision, scope, and mission altered for a better project outcome and investigation.
1.2	10/11/21	Added Functionality (SRS) section	Added in required functionality for the project.
1.3	10/11/21	Added Test Plan	Added in test plan.
1.4	20/11/21	Added new data set	Football-data dataset added to list. This is the chosen data set as seen in the report.
1.5	20/11/21	Removed background	Now found in report document in more detail.
2.0	28/11/21	Format and review document for final submission	N/A

1.2 Project Vision

To predict professional football match outcomes, win, loss, draw, as accurately as possible using artificial intelligence. The aim is to compare different methods of machine learning and compare their ability to be used to predict outcomes. At minimum the project aims to achieve a higher accuracy than the average human guess and strive towards matching/bettering industry standards for each model.

1.3 Project Goals and Objectives

No.	Goals	Objectives
1	Predict matches more reliably than people with good football knowledge.	To predict results of professional football matches.

2	Implement ideas and methods that help combat what would be considered an ‘upset’ or ‘shock’ result.	Have an accuracy of greater than 40%.
3	Compare different ML techniques to see which is best suited for predicting football match outcomes.	Rank different AI techniques for predicting football games.

1.4 Project Scope

The following table outlines the scope of the project. Out of scope requirements can be moved into scope should time allow and upon all team members agreement.

In Scope	Out of Scope
• Predict W/D/L (win/draw/loss) of professional football matches. • Create multiple models that account for team data, both historic and present.	• Predict individual statistics, such as goal scorers, players who make an assist and so on. • Predict scores of games with a high degree of accuracy - focus is on the outcome. • Identify potential upsets (Matches where the underdog wins) • Use a wide range of in-depth industry features including player statistics.

1.5 Major Milestones and Deliverables

Deliverable	Date
Project Plan	05/11/21
Version 1.0 Code	24/11/21
Report	24/11/21
Presentation	10/12/21

1.6 Project Cost and Source Funding

The project has no planned or required cost as it is student based and not needed. However, should the project become commercial, or the research become funded the following costs may need to be considered:

- Salaries of project team
- Software and/or software commercial licenses
- Marketing
- R&D (research and development)
- Trademarks, patent, copyright documentation
- Data sets
- Legal Licenses - dependent on the route of purpose of the project

The above could be funded by:

- An angel investor
- University (research based)
- Form a PLC and sell shares

- Form a Ltd and sell percentage of ownership to investors (shareholders)
- Crowd fund though organizations such as kick-starters

Note: Should the project get to a stage where costs are incurred, or funding is required this section would be expanded into further detail.

1.7 Dependencies

The table below outlines dependencies of the project, split into two sub sections, internal – within the project – and external – outside the project.

Internal	External
• Project members follow the agreed upon rules detailed in later sections. • Design, research, planning, and scheduling are in depth and have the required information for the development phase of the project.	• Data required is available and accessible. • Funding, if required, is obtainable. • Software is available for free and educational use.

1.8 Risk, Assumptions, & Constraints

1.8.1 Risks

The most significant risks to the project include the accuracy and availability of the data sets, inaccurate scheduling estimates, and the coronavirus. See section 20 for an in-depth breakdown of all notable risks to the project along with a contingency plan to minimise these risks.

1.8.2 Assumptions

The following is presumed to be true throughout the duration of the project:

- No funding or financial planning is required
- Required data will remain available and accessible
- All team members will follow the rules of engagement that have been set out
- There will be minimal disruption due to the events of the ongoing Covid-19 pandemic

1.8.3 Constraints

The following are restraint to the project:

- Financial - restricted to no budget due to being a student project.
- Timeline - small time window for project, in practice a longer period would be expected.
- Experience & expertise - final year university students, who have all worked a minimum of a year in the industry, but not specific to the PBA subject area, so these attributes are somewhat limited.

2 Team Charter

2.1 The Problem

Football has many variables which contribute to the outcome of the result of a match. To identifying the likelihood of a team winning, losing, or drawing seems like it is luck if the correct outcome is selected. However, analysis of statistics shows that there is more to luck when predicting the outcome of a game – although it will always be impossible to have 100% certainty of the outcome. Bookmakers consider a vast array of statistics and data, both quantitative and qualitative which are plugged into complex algorithms to determine the odds of matches. Raw data of which they may use, but not limited to:

- Expected goals by team
- Goals scored by the team
- Goals conceded by the team
- Expected goals by individual players
- Previous stats on goals scored against a given team
- Previous stats on goals conceded against a given team
- Formation and team strategies
- The effectiveness of team strategies against a given team
- Home record
- Away record
- Fundamentals - news around the team
- State of players' and coaches' personal lives

Football, like any sport is hard to predict as there is a huge uncertainty around a lot of the data. Even when stats alone suggest on paper one team should beat another this is still not guaranteed and building an algorithm that adjust for this is challenging. In later sections there is a short analysis of betting company's accuracy for the English Premier League Results so far which comes out at around ~40% accuracy. These companies have invested hugely into this area of the business and spent decades perfecting prediction algorithms, and still only produce an accuracy that is not massively better than a random guess.

Accounting for randomness is the biggest challenge to the project, factors include referees, venue, competition, as individuals there is access to news which the team will not be providing as an input. For humans we get a feeling, but what is this based on, and how can we put this into metrics. What do shocks all have in common? How do you predict Leicester winning the league in 2016, were there any indicators for this? All these questions are asked in order to determine a prediction and attempt to make an accurate prediction on a game.

2.2 Data Sets

Viable data sets for the project:

- Kaggle - https://www.kaggle.com/hugomathien/soccer
- Elana Sport - https://elenasport.io/doc/coverage
- Sport Data Api - https://sportdataapi.com/football-soccer-api/
- Serpapi - https://serpapi.com/sports-results
- Football-Data - https://football-data.co.uk/englandm.php

Data Set	Overview
Kaggle	The Kaggle dataset is a European football database consisting of player and team attributes which have been sourced by the football game FIFA. These attributes which have a cap up and until 100, despite these attributes are determined on prior years performances some statistics can be inflated based on a player performing well last year and so these attributes can be deemed misleading. This database also has a given teams' formation and their respective play style; this is useful as a teams' formation when compared to another team's formation can influence the way the game ends as some formations can counter other

	formations this also applies to a teams' play style. On top of this, the database holds match data and goals scored, conceded for a home/away team, fouls, cards, crosses, possession, shots-on and more.
Elana Sport	ElenaSport allows countries, league, stage, stand, line-up, team, player, player stats, fixtures all to be queried and returned by the API despite having all of this information we don't have enough information on individual player statistics or individuals which would be very helpful in working out expected goals and goals being scored by individuals, metrics like this which massively contribute to working out and being able to accurately predict the results on matches.
Sport Data API	Sport Data Api is very similar to the ElenaSport API with all the information that there is access to. The additional information accessible is bookMakers ID, Markets, Odds, Referees and Rounds. The odds, markets and bookMakers gives a clear indication of how bookmakers and what the market is suggesting for the odds of the game in question which provides a statistic which can be used when comparing the likelihood of a result when being compared to the techniques we decide to deploy. Rounds and Referees can play a big factor as there are some teams who suffer from anomalous results which are due to competitions.
Serpapi	Serpapi - Google Sports API provides the information for score lines, team, league position, date of kick off, competition, round in the competition. This is yet again similar to previously mentioned datasets however, the Google Sports API links the YouTube videos which are the highlights for a given football match, if the project was to extend the requirements beyond the initial scope, then an option could be to take video footage and analyse it to determine why and how a given team won a game and how that can transition and be applied to given upcoming results.
Football-data	Football-data provides an expected range of features for a publicly available data set. This includes; additional stats beyond the game scores, such as, cards, attendance, referee, offsides and so on. It also comes in csv format, one per season which makes the data easy to manage when using R.

2.3 Initial Research

2.3.1 Data Modelling Approaches

After conducting extensive research into previously existing solutions in the field of football statistical modeling, we have decided to use the following approaches:

- SVM - Supervised Learning
    - SVM's use labelled datasets to train algorithms to classify data or predict outcomes accurately. As input data is passed into the model, it makes adjustments to the weights attached to each item of data until the model has been fitted appropriately. These patterns are subsequently used to classify/predict unseen data. A visual example of how an SVM works can be seen below [5].

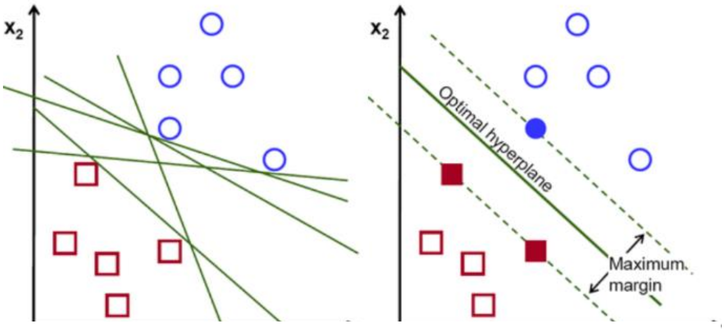


Figure 1: SVM Example [19]

    - SVM has been selected as one of our models as it is a strong, general use machine learning algorithm that can be applied to many different situations. SVMs are also very well supported in R, and so can provide ease of use for our project. This model will be combined with a K-fold Cross Validation evaluation method which will allow us to minimise the measure of error for the solution.

- Random Forest
  - Random Forest is a flexible, easy to use machine learning algorithm which is known for producing above average results. It works by building a network of decision trees that are trained in conjunction with one another to increase the overall result. A visual example of how a Random Forest works can be seen below [6].
  - The project will benefit from deploying a Random Forest modeling approach as this covers both classification and regression problems, both of which will be considered within our project.

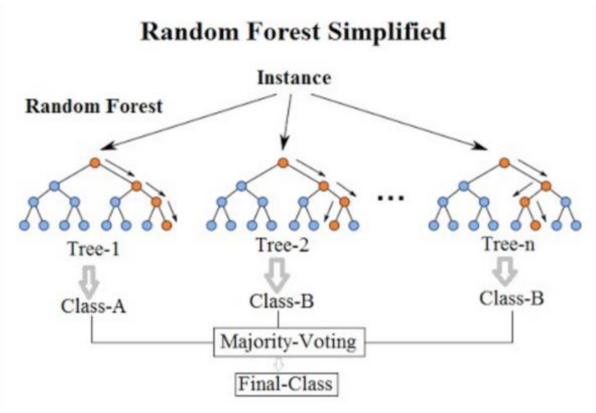


Figure 2: Random Forest Example [20]

- KNN
  - KNN is a simple model that performs classification based upon the class of its k-nearest points based on a distance metric. The output of a KNN is a class membership. An object is classified by the vote of its neighbors, with the object being assigned to the class most common with its k nearest neighbors. A visual example of how a KNN model works can be seen below [7].
  - Our project will utilise the KNN model as it will allow us to experiment with an alternative approach to supervised learning.

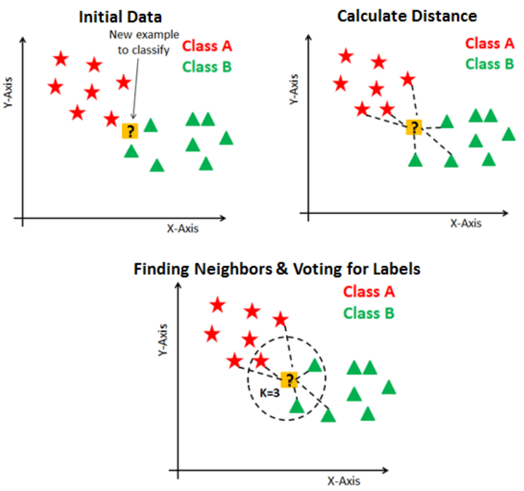


Figure 3: KNN Example [21]

- Multiple Logistic Regression (MLR)
  - While simple logistic regression refers to the regression application with one dichotomous outcome and one independent variable, multiple logistic regression gets applied when there is a single dichotomous outcome but more than one independent variable. These are coded in terms of values between 0 and 1, which results in a probability value being the final output.

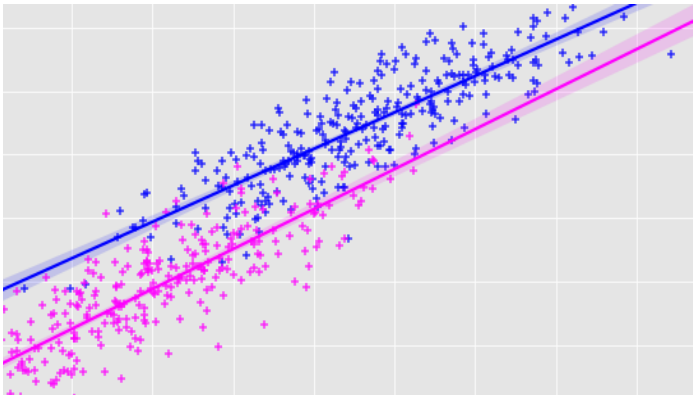


Figure 4: MLR Example [22]

- Using multiple logistic regression in the context of our project will be a valid model as the approach of MLR is well suited to the dataset that we will be working with.
- Having multiple values that can be coded to a percentage outcome and using each value to predict the final result could be an ideal model and certainly one that should and will be explored within this project.
- The team has decided to use a wide variety of approaches so that we can compare and contrast the degree of success of each model and settle upon an optimal project solution.

2.3.2 Expectations

The expectation of the final solution is that our project will be able to predict the outcome of a premier league game given the teams playing and game statistics. The expectation is that each model will achieve a higher percentage success rate than the benchmark of 33% (a random guess) and intend for the solution to be able to compete with the success rate of previous models using existing machine learning approaches.

2.3.3 Evaluation Criteria

Criteria	Success Criteria	Achieve by
Business Understanding	Every team member understands both primary and secondary objectives of the project.	26/11/21
Data Understanding	Every team member understands the data that will be collected, and why we are collecting it.	03/11/21
Data Preparation	Data is pre-processed into a format that is usable for R when needed. All erroneous values are removed from the dataset.	08/11/21
Modelling	A model can predict a match outcome successfully	14/11/21
Results	One or more of the models used is able to predict match outcomes at a rate higher than 33%.	24/11/21
Deployment	The final solution and report are submitted in a timely manner.	26/11/21

Individual model evaluation methods can be found in the report document.

2.4 Rules of engagement

2.4.1 Expected Behaviour

It's important that as a team we are able to meet consistently to get the maximum amount from the sessions since this will be something that would happen twice a week whereby all team members will be present. As a team to get the most out of these sessions we want to follow these guidelines to ensure our project is successful:

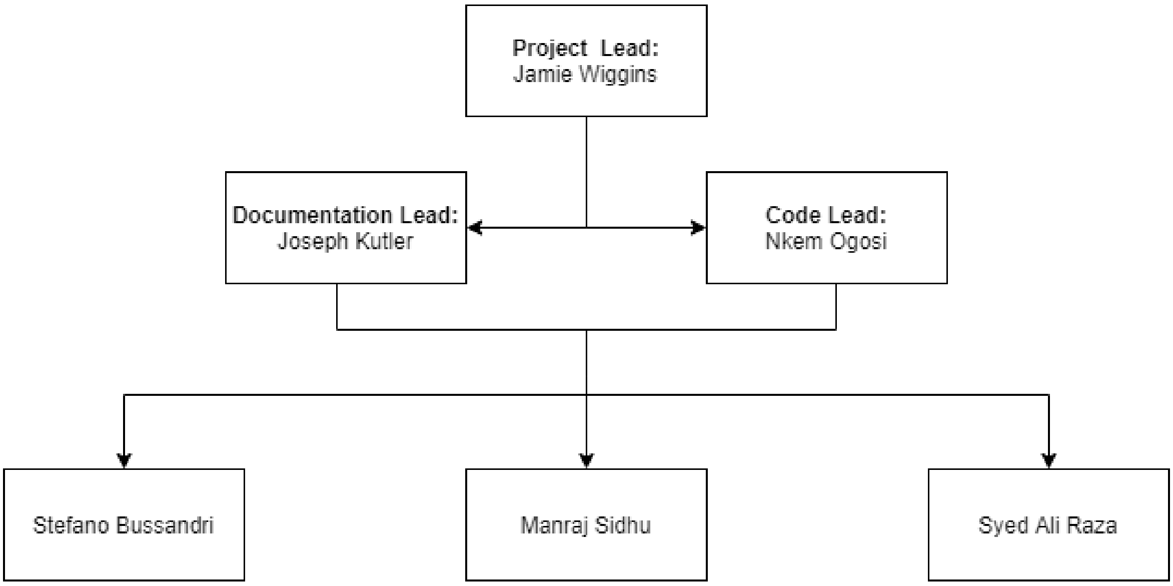
- Arrive on time
- Speak one at a time and don't interrupt someone who is talking
- Everyone has the opportunity to contribute in each session
- We arrive with a team agenda or plan every meeting
- Record attendance at meetings to ensure no one is felt excluded

2.4.1.1 Operating Protocol

- Update the JIRA board when you are doing a task, so the team is notified that you are performing the given task.
- Comment on the Google Drive on documents to provide feedback on what team members have done and questions you may have for colleagues.
- Communicate via Discord and WhatsApp group chat if there are any urgent questions with regards to the project.

2.4.2 Roles

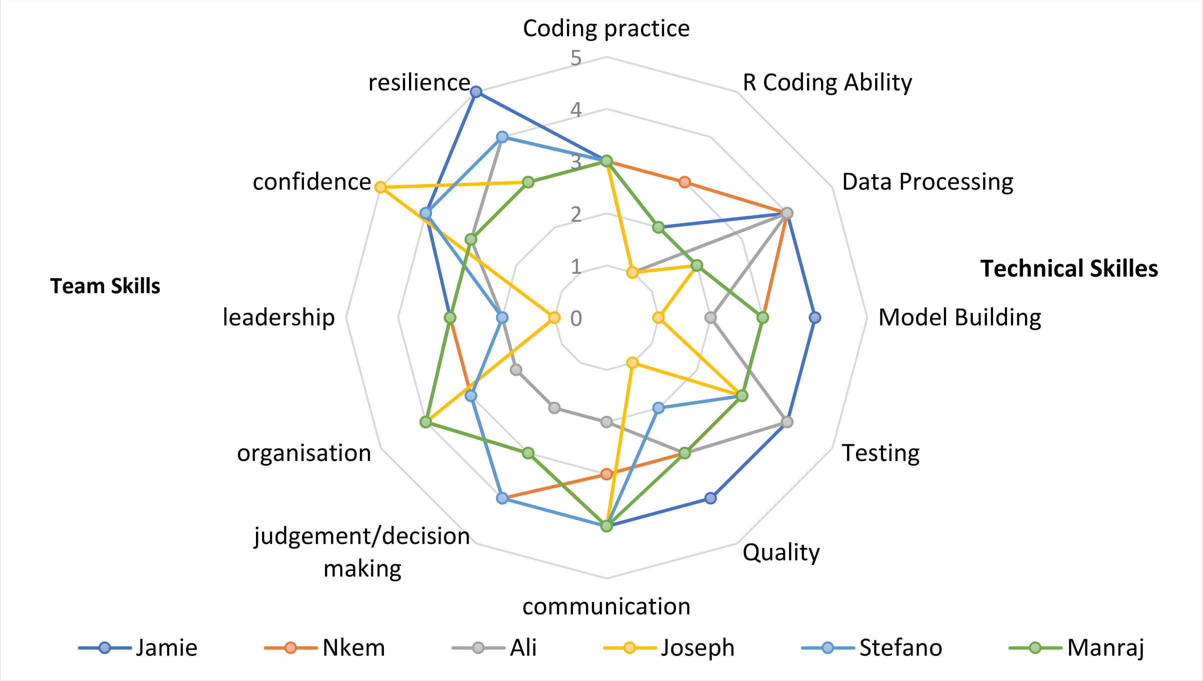
Below depicts the structure and members main roles within in the project.



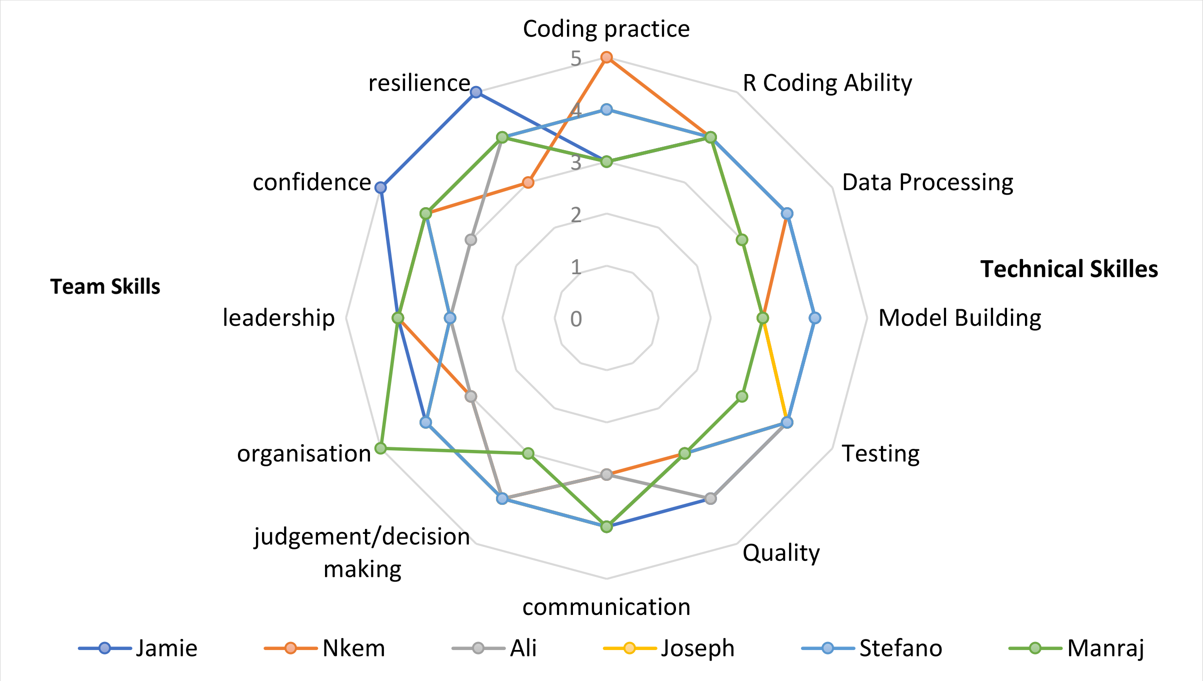
2.4.3 Self-Evaluation

Project members will evaluate themselves out of five on a variety of skills across the project. These will be split broadly between 'Team Skills' and 'Technical Skills' which we will have periodic reviews of throughout the duration of the project.

Start of the project:



End of project:



3 Model Requirements

3.1 Functional Requirements

The following section details the functional requirements of the project.

1. **Data Cleaning**
    a. The dataset shall be cleaned of all outliers

   b. The dataset shall have irrelevant fields removed
   c. Structural errors shall be fixed
   d. Missing data shall be handled in any of the following ways:
      i. Drop observations that have missing values
      ii. Input missing values based on other observations
      iii. Alter the way the data is used to effectively navigate null values

2. **Data Preparation**
   a. The dataset shall contain all required derived fields
   b. The dataset shall be made up of normalised data
   c. The dataset shall contain up-to-date values
   d. The dataset shall be future proof
   e. The dataset shall be provided in an accessible format
3. **KNN Model**
   a. The model shall be able to be trained upon a section that makes up 70% of the given dataset
   b. The model shall be able to be tested upon a section that makes up 30% of the given dataset
   c. The model shall be able to make a prediction on unseen data based upon the training dataset
4. **Random Forest Model**
   a. The model shall be able to be trained upon a section that makes up 70% of the given dataset
   b. The model shall be able to be tested upon a section that makes up 30% of the given dataset
   c. The model shall be able to make a prediction on unseen data based upon the training dataset
5. **Neural Network Model**
   a. The model shall be able to be trained upon a section that makes up 70% of the given dataset
   b. The model shall be able to be tested upon a section that makes up the remaining 30% of the given dataset
   c. The model shall be able to make a prediction on unseen data based upon the training dataset

3.2 Non-functional Requirements

This section covers any non-functional requirements of the website and can be subject to the opinion and judgement of the user/tester.

6. **System prediction accuracy**
   a. The system shall produce a prediction accuracy equal to or higher than our primary benchmark of 33.3%.
   b. The system shall produce a prediction accuracy equal to or higher than our secondary benchmark of 42.42%.
   c. The system shall produce a prediction accuracy equal to or higher than our stretch goal of 43%.
7. **Code efficiency**
   a. The system shall produce a prediction within 30 seconds of running

8. **Code effectiveness**
    a. The code will be well commented
    b. The code will contain meaningful variable names
    c. The code will be readable to a reasonable degree
9. **Scalability**
    a. The system will be made in such a way that it is scalable for future developments

4 Testing Plan

4.1 Overview of Testing

Testing for the project is conducted at the end of the creation of the algorithms and models. The testing is done so that the parameters for each respective implementation of KNN, Random Forest, NN to optimise the models.

4.2 Testing Goals, Objectives, and Outcomes

No.	Goals	Objectives
1	Implement robust machine learning models that are functional and give the desired outcome, regardless of accuracy.	95% pass rate on all tests run and no failed tests prevents providing a relevant outcome to preform analysis and evaluation.

4.3 Assumptions

- Matches should be in the same game week when rescheduled
- Dataset is accurate and match results are presented correctly
- User can run the R script
- Sufficient time to test
- Every test that passes implies that the work is working correctly within the appropriate environment (running r scripts within R studio)

4.4 Principles

- All tests are to be focused on the aforementioned project scope and outcomes
- Tests are to be designed to fulfil the acceptance testing and Blackbox testing requirements
- Tests which fail are to be resolved prior to the internal deadlines
- Testing to be repeated across sprints (Regressions)
- Each sprint will have clearly established testing metrics

4.5 Data Approach

4.5.1 Split Test and Train

Each model will first be trained on 80% of the dataset to fit the model and then tested on the remaining 20% to evaluate the model. Splitting the dataset to hold a portion solely for training and testing is used on machine learning algorithms to allow for predictions to be made on the data that has not be used to train the model. This is done to get the accuracy of the models on new, unseen data (data not provided within the training dataset) to best represent how the model would perform in practice.

4.5.2 Cross Validation Testing

The project will follow a K-Fold cross validation method which is an approach that allows for the entire dataset to be provided as both training and testing data to mitigate the model from being overfitted since we have a limited dataset.

The number of fixed folds is 5 and the project will be deploying a cross validation approach in most applicable cases. Shuffling over the training dataset based on the number k-folds there are, is done to reduce the amount of bias shown in the model. This occurs because the new training and testing set will be more representative of the data in its totality since it is iterated over based upon the number of folds that have been defined. However, this leads to the algorithms needing to be run k times.

The model be presented with unseen data and a prediction of a particular matchup resulting in a home win, away win or a draw will be made for an upcoming game week of matches.

4.6 Strategies

The following strategies are being used to test:

- Black box testing:
  - Testing the functionalities of the algorithms and focus solely on input and output of the models.
- Acceptance testing:
  - To ensure the algorithms and models run smoothly and based on the criteria predetermined once the complete of the development has finished. Acceptance testing will be used to validate everything that is done throughout the lifecycle of the project.

4.7 Test Criteria

Testing condition	Description	Additional Information
All tests run successfully	All tests are required to run regardless of result should cover test requirement and metric established that sprint	
Most tests run successfully	Many tests run but defects are holding back tests which are ought to be resolved in this sprint.	
Tests fails but no high priority tests fail	Tests have failed and those in question will be given a priority.	
All tests fail	Failed tests must be reported as bug/defect and be updated on the Jira board.	
Tests to be approved	All tests must be checked by another team member and approved on the JIRA board	

4.8 Validation & Defect Management

The validation and management of defects during this project will be matched against the test criteria. On Jira the priority tag will be used to represent the difficulty and the importance of the defect that needs to be managed. The priority will be established and differentiated in multiple ways from high to low.

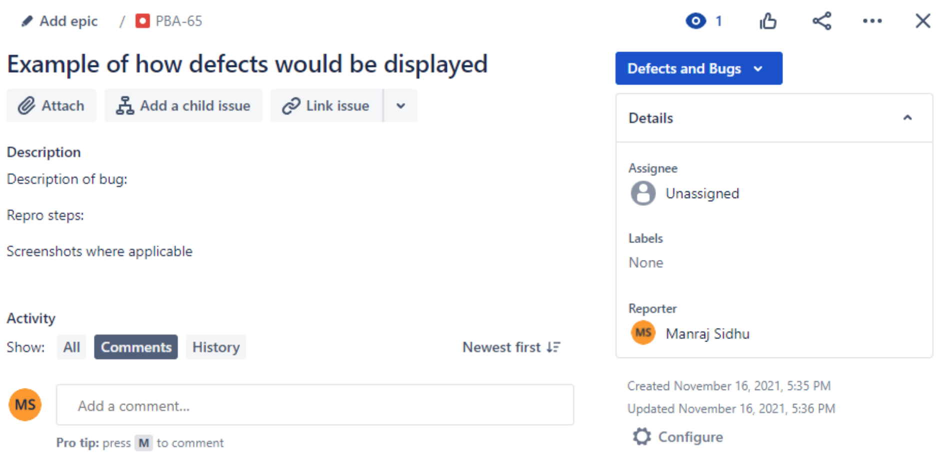
High - Tests of this priority are considered at the highest importance due to their impact on the project and the schedule. If the tests interfere with these outcomes or alters the pathing it is determined to be a high priority test.

Medium – Tests which impact some of the Blackbox tests but does not disrupt the testing of the other models to be tested in the sprint.

Low – Tests fail but do not have any impact any of the predetermined requirements.

4.9 Defect Tracking and Reporting

When tests fail or there are models which are not demonstrating the output as expected, the individual who has been working in that section of the code will be required to do the following on JIRA. In the Defects and Bug column on the board, team members would need to give a Title for the bug, a description of what the bug/defect is, reproduction steps if and where necessary in the code and screenshots of the code and where the error has occurred.



Example of how defects would be displayed

Description

Description of bug:

Repro steps:

Screenshots where applicable

Assignee: Unassigned

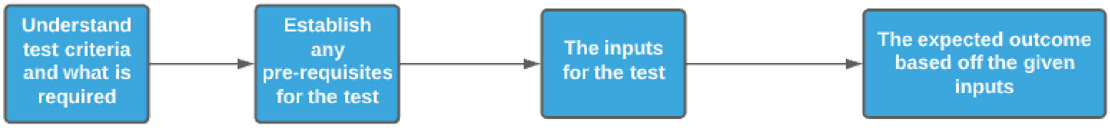
Labels: None

Reporter: Manraj Sidhu

Created November 16, 2021, 5:35 PM

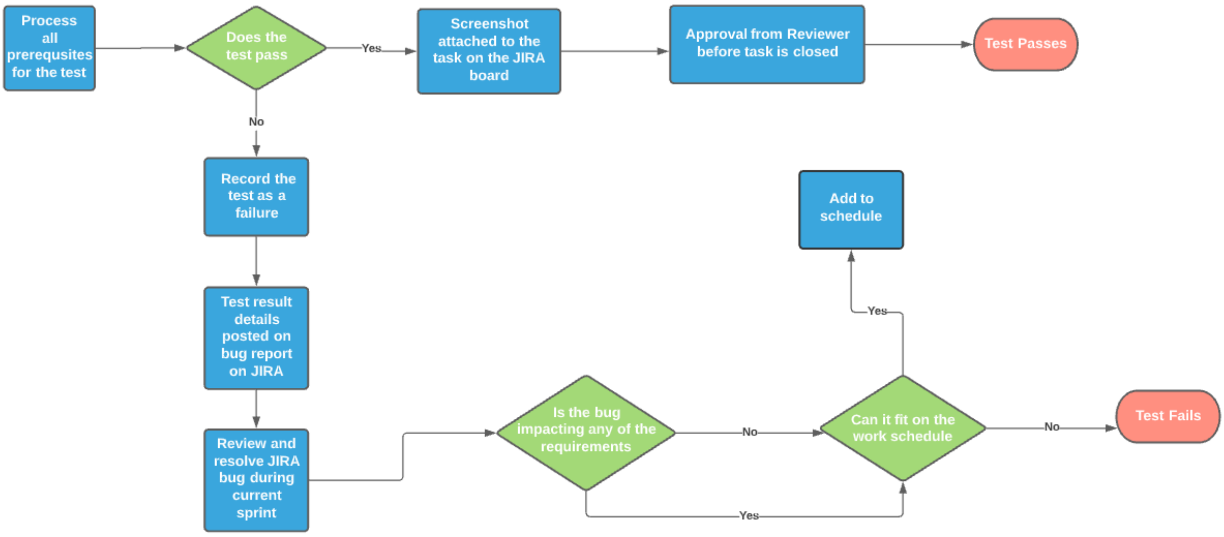
Updated November 16, 2021, 5:36 PM

4.10 Testing Design Process



Understand test criteria and what is required
Establish any pre-requisites for the test
The inputs for the test
The expected outcome based off the given inputs

4.11 Test Execution Process


Process all prerequsites for the test
Does the test pass
Yes
Screenshot attached to the task on the JIRA board
Approval from Reviewer before task is closed
Test Passes
No
Record the test as a failure
Test result details posted on bug report on JIRA
Review and resolve JIRA bug during current sprint
Is the bug impacting any of the requirements
No
Yes
Can it fit on the work schedule
Yes
Add to schedule
No
Test Fails

4.12 Testing Risks

Risk	Impact	Prob	How to mitigate
Tests delayed due to circumstances out of our control that results in implementation taking a significant time more than that scheduled	M	M	Deadline affected by the risk of circumstances out of our control will result in changes to the schedule taking this new event into account.
Model fails to compute and does not provide the accuracy measures required	H	M	Refer to resources found in the drive and links provided by team members. If a team member is still struggle, they can look to their colleagues for support where possible.
Tests outcomes and expectations are not clearly defined which results in errors occurring	L	H	The tests requirements are created and discussed with members hence by following these defined test designs, following this would mitigate that
Project goes beyond scope or changed in its scope – older tests no longer required	M	L	Changes to the scope would be addressed in the scheduling and implementation of these tests. These changes should be documented in each section.

4.13 Tests Against Scope

The following table outlines the scope of the project and what is to be achieved from the scope.

4.14 Tests Against Functional Requirements

The table below outlines the functional requirements for the project from section 3.

Feature Description	Requirement Numbers	Comment
1 – Data Cleaning	1a-1d	
2 – Data Preparation	2a-2e	

3 – KNN Model	3a-3c	
4 – Random Forest Model	4a-4c	
5 – Neural Network Model	5a-5c	

4.15 Tests Against Non-Functional Requirements

The table below outlines the non-functional requirements for the project from section 3.

Feature Description	Requirement Numbers	Comment
6 – System prediction accuracy	6a-6c	
7 – Code efficiency	7a	
8 – Code effectiveness	8a-8c	
9 - Scalability	9a	

5 Project Management

5.1 Methodology

5.1.1 CRISP-DM

5.1.1.1 Understanding

CRISP-DM is the CRoss-Industry Standard for Data Mining. It is a six-process model which describes the application of the data science life cycle. CRISP-DM allows you to plan and implement machine learning into a project. It does this using the below standards:

- Business understanding – Looks at the requirements for the project and any business specific needs.
- Data understanding – Details the data requirements for the task and how appropriate any existing data is.
- Data preparation – Discusses how data will be prepared for processing.
- Modelling – Details which modelling techniques will be used.
- Evaluation – Performs an analysis of the above modelling techniques and why they are the best techniques for the given task.
- Deployment – Looks at how accessible the completed data will be for necessary stakeholders.

5.1.1.2 Business Understanding

The team will spend time thoroughly understanding what it is that we truly want to accomplish from our solution so that we can avoid expending a great deal of effort producing the right answers to the wrong questions. This will be achieved by discussing and agreeing upon the primary objectives of the project, followed by the outlining of secondary objectives and feasible stretch goals. Doing so will allow us to uncover important factors that can influence the outcome of the project. Additionally, the team is going to outline and identify key risks that could pose a threat to the success of the project and implement effective contingency plans to counteract the degrading effects of negative unforeseen circumstances.

5.1.1.3 Data Understanding

Adding to the foundation of the Business Understanding phase, the team will continue to follow the CRISP-DM method as we enter the Data Understanding phase where we will identify, collect, and analyse the data sets that can help us to accomplish the project's goals. After we have collected, described and explored the data we will verify the data quality and, if the quality is blemished, we will document any of the issues that have been identified. We have chosen our primary source of data to be a dataset provided by Kaggle. Alternatively, if it becomes clear that this is not a viable dataset, we have identified that we could use any of the secondary datasets mentioned in section 16.

5.1.1.4 Data Preparation

Given that the team now understands the intricacies of the data that will be used, we will spend a considerable amount of time correcting, imputing or removing erroneous values. The team will also consider deriving new attributes from pre-existing data that can enhance our final analysis. This will be followed by the integration and formatting of data as necessary in preparation for the modelling phase.

5.1.1.5 Modelling

The team has decided to take an iterative approach to technical model building and assessment which will repeat until we strongly believe that we have found the best models. This will consist of the team being able to generate a test design, build and assess the model, and proceed through the CRISP-DM lifecycle whereupon we will further improve the model in future iterations.

5.1.1.6 Evaluation

In order to remain consistent with the CRISP-DM methodology, in this phase the team will look more broadly at which model best meets the business needs and what to do next. The first method of determining whether the models are appropriate is to assess whether the models meet the business success criteria whereupon we can figure out which model should be approved for the business. This will be followed by a rigorous review process where the team will consider what is next for the project, whether it be to proceed to deployment or iterate further.

5.1.1.7 Deployment

The final phase of the CRISP-DM lifecycle, deployment, is incredibly important and it is imperative that the team follows this method closely. This consists of outlining and defining plans to do with the deployment itself, monitoring and maintenance, and conducting a retrospective inquest into the success factors of the project.

5.2 Agile

5.2.1 Understanding

AGILE is a methodology which focuses on allowing for better adaptability to changing needs of the team over the course of a project. This is done through breaking down the project into smaller increments which are much easier to manage. Once the project is under way the team member will work through a continuous process of planning, executing and then evaluating the smaller increments.

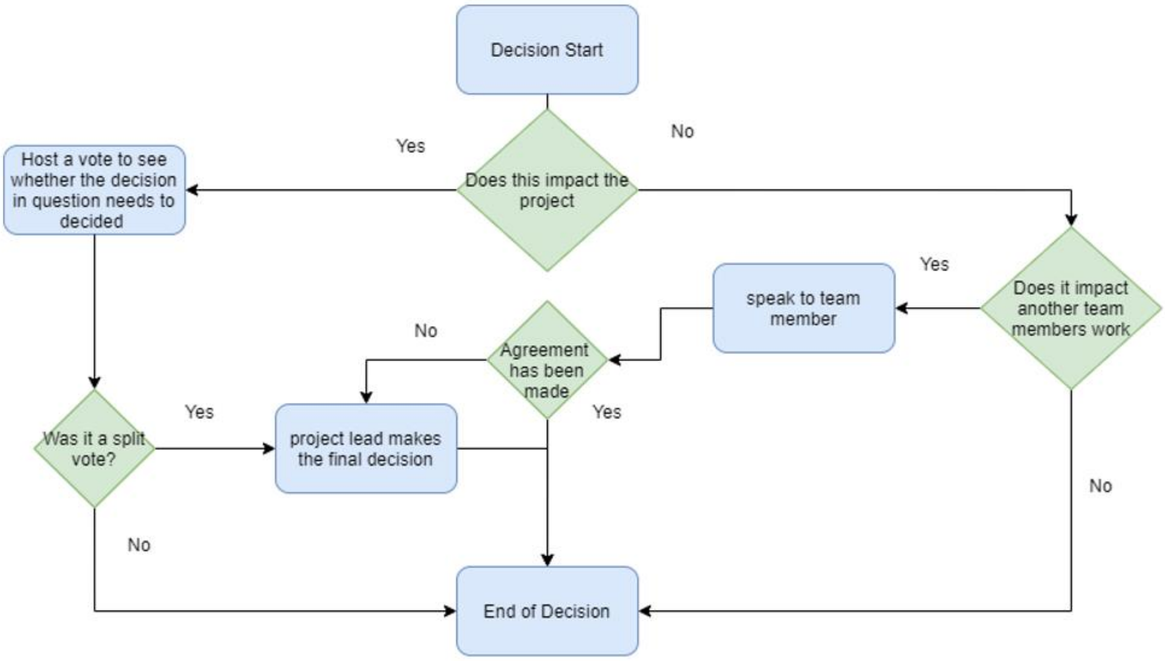
5.2.2 Application - Scrum ban

The team has opted to utilise a combination of the Scrum and Kanban methodologies so that the project can take advantage of having both a parallelised and serialised approach when it comes to implementing our solution. The Scrum methodology encourages teamwork and structure, with our group having a heavy emphasis on embracing the sides of a business analytics project outside of just the programming aspect. To complement this approach, we are using Kanban as this brings a focus on continuous delivery, improved productivity and enhanced flexibility which is hugely beneficial due to the often tightly packed schedule of a university student. This approach will be characterised by the varying length sprints that make up the Scrum aspect, combined with the use of a Kanban board to ensure that all team members are well informed of the tasks that they need to complete. The documented weekly implementation structure represents all the tasks from the Kanban board that were completed during each weekly scrum. We have taken this approach because it minimises the time spent needing to plan the next steps as there will be improved team cohesion.

By strictly following the methodologies we have decided upon regarding data handling and team management, we will create a fluid and efficient workflow that enables the project to achieve its goals.

5.3 Decision Making Process

The following flowchart depicts the decision-making process within the team and how split decisions will be handled.


Decision Start
Does this impact the project
Yes
No
Host a vote to see whether the decision in question needs to decided
Does it impact another team members work
Yes
speak to team member
Agreement has been made
No
Yes
Was it a split vote?
Yes
project lead makes the final decision
No
No
End of Decision

5.4 Resources

Primary Data set	Football-Data: https://football-data.co.uk/englandm.php
Secondary Data set	SerpAPI: https://serpapi.com/sports-results Elena Sport: https://elenasport.io/doc/coverage Sport Data API: https://sportdataapi.com/football-soccer-api/ Kaggle: https://www.kaggle.com/hugomathien/soccer
R Documentation	R Docs: https://www.rdocumentation.org/
External Documentation	Stack Overflow Python Docs Ladbrokes Weekly Pre-Match Predictions Blog Stack Exchange
Internal Documentation	Surrey Learn – Module content

5.5 Lines of Communication

The table below outlines the lines of communication and scheduled meetings and group sessions of which who is expected to attend.

Who	How	When	What	Why
All Team Members	Weekly meeting face-to-face	Wednesday 2pm - 5pm	In person group work. I.e. Documentation & coding.	Keep the team on track and to be able to discuss things about code
All Team Members	Weekly meeting online - Discord	Sunday 8pm - 8:30pm	Weekly retrospective	Plan for upcoming weekly work
Paired work	Ad hoc face-to-face or online	Whenever needed	Peer supported documentation, development or testing	Aids real time peer review and allows pairs to more easily communicate issues

5.6 Record Keeping

Doc Storage	Google Drive
Documentation	Google Docs, Adobe PDF
Code	GitLab
Tracking & planning	Jira

6 Work Plan

6.1 Project Schedule

6.1.1 Milestones and Deliverables

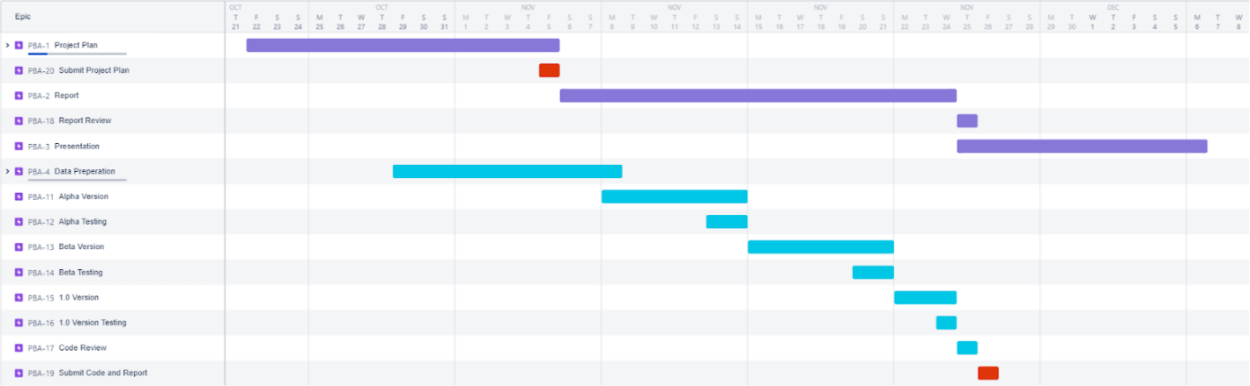
Deliverable	Due Date
Project Plan	05/11/21
Data Collection and preprocessing	08/11/21
Alpha Version	14/11/21
Beta Version	21/11/21
Version 1.0	24/11/21
Report	24/11/21
Presentation	06/12/21

Sprint	Milestone	Description	Due Date

Sprint 1	Project Proposal	Initial project description and outline	22/10/21
	Project Plan	Document containing project, team charter, and a work plan.	05/11/21
	Sprint 1 - Data collection and preprocessing	Collect data needed and process it into a usable form	08/11/21
Sprint 2	Sprint 2 - Alpha version	Build an alpha version of the project	14/11/21
	Alpha version testing	Test the alpha version of the project and the model	14/11/21
Sprint 3	Sprint 3 - Beta version	Implement changes to the alpha version based on the testing	21/11/21
	Beta version testing	Test the beta version of the project and the model	21/11/21
Sprint 4	Sprint 4 - 1.0 version	Implement changes to the beta version based on the testing	24/11/21
	1.0 version testing	Complete any final changes to the project and the final testing	24/11/21
	Report	Complete the report	24/11/21
	Code and report review	Review the final version of the project and the report	25/11/21
	Code and Report	Submit the final version of the code and the report	26/11/21
Sprint 5	Presentation	Presentation of the project	06/12/21

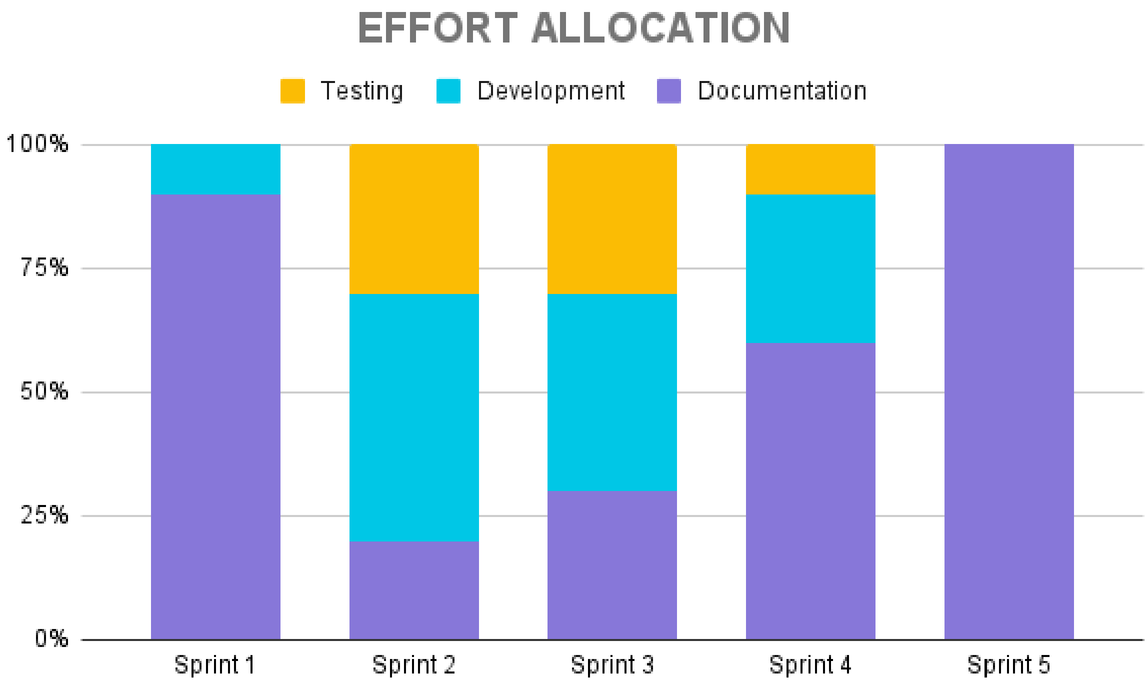
6.2 Initial Schedule

The figure below shows a Gantt chart breakdown of the project schedule. It is split into two different sections, documentation (purple) and development (blue). This is an initial schedule so may be altered throughout for improved efficiency or results upon review in meetings.



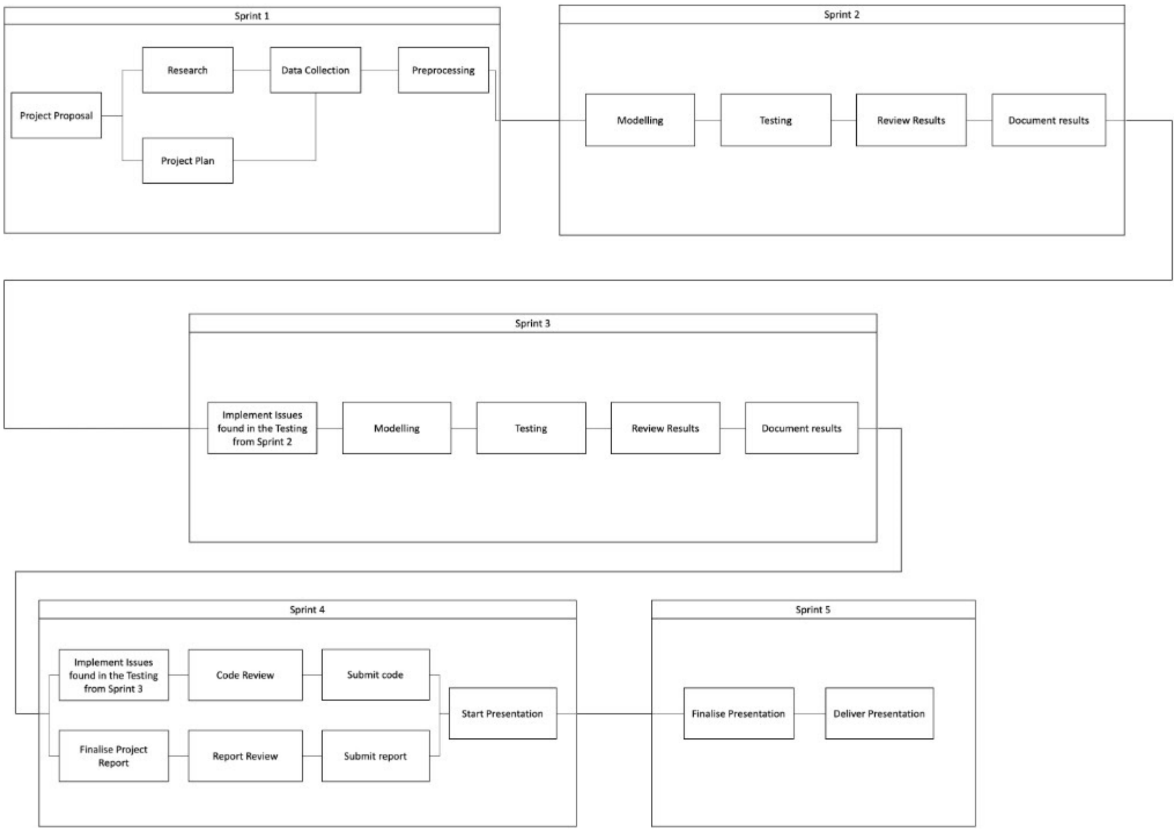
6.3 Effort Allocation

The figure below presents the effort allocation of the team that should be applied to each sprint. This may be subject to change dependent on team reviews but provides an outline of the expected workload split throughout the project's sprints.



6.4 WBS and Logic flow

The figure below outlines an example of the breakdown of each sprint. However, these tasks are not final as the agile approach will allow us to review will result in reviews at the end of each sprint and planning for the next sprint which may lead to changes in the WBS. These will be reflected in updates to the WBS throughout the project's duration.



6.5 Risk Overview

6.5.1 Risk

The team is aware of potential risks that can be deemed to be of high impact to the project and have outlined the necessary steps to ensure that these issues are minimised and otherwise contained.

1. Accuracy and availability of data - poor data or lack of will have a large negative impact on the success of the project. This is because the algorithm is dependent on good quality data to ensure it produces good quality results.
2. Coronavirus - the global pandemic has introduced many new challenges and changes to the way people work. There has been great emphasis on home working and online communication and even though the impact is not as present now as it was last year it still poses a major potential threat to the project and its deadlines due to many unknown quantities around it.
3. Inaccurate scheduling estimates - poses a risk to the development of the project, as could result in missing major milestones and deadlines, which could have a potential negative impact on the success of the outcome of the project.
4. Requirement's inflation - during the progression of the project more features may emerge that were not initially identified which may affect scheduling, resulting in missed deadlines or a poor-quality outcome.
5. Unavoidable risks - such as, changes in legal legislation in relation to the software, or data that is being used.
6. Poor Testing - failure to test the application thoroughly and fairly could result in poor feedback and misleading results and may lead to bugs being missed or a lower rate of success upon 1.0 release.

6.5.2 Risk Analysis

No.	Prob	Impact	Control Measure
1	L	H	Data will be well researched and investigated to ensure the highest quality data is used and available for the project.
2	M	H	The project team has both online and in person lines of communication in place. The team is set up in such a way that should a transition to pure home working be required with no in person contact it will be possible.
3	L	M	The team must remain flexible to adapt to possible changes and improvement of the product, and willing to put in more time to make them feasible or be willing to dispense of previous requirements in favour of better ones.
4	M	M/H	Produce realistic and fair estimates and adjust and react when required in a timely manner to reduce the effect of changing timelines.
5	L	H	Research and remain up to date on relevant regulations and be willing to adapt to unexpected changes.
6	L	H	Plan for extensive testing throughout development and ensure all work goes through quality assurance checks. Any bugs or issues will be logged and tracked to make sure they are dealt with.

key: H = high, M/H = medium/high, M = medium, and L = low

7 References

[1] "premier-league-explained," [Online]. Available: https://www.premierleague.com/premier-league-explained. [Accessed 16 November 2021].

[2] "machine-learning-algorithms-for-football-prediction," [Online]. Available: https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8. [Accessed 19 November 2021].

[3] "Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches," [Online]. Available: https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf. [Accessed 19 November 2021].

[4] "Premier League Fans Summary," [Online]. Available: https://insight.gwi.com/hs-fs/hub/304927/file-2593818997-. [Accessed 20 November 2021].

[5] “share-of-sports-betting-great-britain,” [Online]. Available: https://www.statista.com/statistics/917088/share-of-sports-betting-great-britain/. [Accessed 20 November 2021].

[6] “how-david-prutton-rates-pundit,” [Online]. Available: https://www.football.london/reading-fc/how-david-prutton-rates-pundit-15149366. [Accessed 21 November 2021].

[7] “World_Football_Elo_Ratings,” [Online]. Available: https://en.wikipedia.org/wiki/World_Football_Elo_Ratings#cite_note-Laesk_et._al.-1. [Accessed 22 November 2021].

[8] “football_rankings,” [Online]. Available: http://lasek.rexamine.com/football_rankings.pdf. [Accessed 22 November 2021].

[9] “epl-week-4,” [Online]. Available: https://www.ladbrokes.com.au/blog/soccer/epl-week-4/. [Accessed 14 November 2021].

[10] “premier-league-betting-tips-paddy-power-best-bets,” [Online]. Available: https://news.paddypower.com/football-tips/2021/11/19/premier-league-betting-tips-paddy-power-best-bets/. [Accessed 14 November 2021].

[11] “english-premier-league-betting-preview-liverpool-and-man-city-cap-an-intriguing-weekend-of-fixtures,” [Online]. Available: https://uk.sports.yahoo.com/news/english-premier-league-betting-preview-liverpool-and-man-city-cap-an-intriguing-weekend-of-fixtures-120404239.html?guce_referrer=ahr0chm6ly93d3cuz29vz2xllmnvbs8&guce_referrer_sig=aqaaaetzzh_icx8w_ewxbd19qsvdqvyxn0pcfip15x5. [Accessed 14 November 2021].

[12] “sports-results,” [Online]. Available: https://serpapi.com/sports-results. [Accessed 3 November 2021].

[13] “opta,” [Online]. Available: https://www.statsperform.com/opta/. [Accessed 10 November 2021].

[14] “englandm,” [Online]. Available: https://www.football-data.co.uk/englandm.php. [Accessed 10 November 2021].

[15] “international-football-results-from-1872-to-2017,” [Online]. Available: https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017. [Accessed 12 November 2021].

[16] “ratings,” [Online]. Available: https://football-data.co.uk/ratings.pdf. [Accessed 22 November 2021].

[17] “predicting-forecasting-football,” [Online]. Available: https://mercurius.io/en/learn/predicting-forecasting-football. [Accessed 19 November 2021].

[18] “opencv.org,” [Online]. Available: https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html. [Accessed 3 November 2021].

[19] "opencv.org," [Online]. Available: https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html. [Accessed 3 November 2021].

[20] "medium.com," [Online]. Available: https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d. [Accessed 2 November 2021].

[21] "datacamp.com," [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Accessed 2 November 2021].

[22] "towardsdatascience.com," [Online]. Available: https://towardsdatascience.com/simple-and-multiple-linear-regression-with-python-c9ab422ec29c. [Accessed 1 November 2021].