# Premier League predictive learning algorithm (PLePA)

Contents

[Premier League predictive learning algorithm (PLePA) 1](#_Toc37488294)

[Glossary 1](#_Toc37488295)

[Section 1 Preparation and planning 2](#_Toc37488296)

[1.1 Project Description 2](#_Toc37488297)

[1.2 Revisions made and reasons why 2](#_Toc37488298)

[1.3 Tasks, subtasks, evaluation criteria 3](#_Toc37488299)

[1.4 Updated resource list 5](#_Toc37488300)

[1.5 Updated skills list 5](#_Toc37488301)

[1.8 LEPSI review 7](#_Toc37488302)

[References 8](#_Toc37488303)

[Appendices 8](#_Toc37488304)

[Appendix 1 – Resource list 8](#_Toc37488305)

[Appendix 2 – Skill list 8](#_Toc37488306)

|  |  |  |  |
| --- | --- | --- | --- |
| Appendix | Page | Name | Description |
| 1 |  | Resource list | A table to show the initial resources required. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Figure | Page | Name | Description |
| 1 |  | Structured-case life cycle | An image created to show the structured-case life cycle which will be used by PLePA. |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table | Page | Name | Description |
| 1 |  | Activity evaluation | The evaluation criteria for PLePA to know if it has been successful. |
| 2 |  | Updated resource list | A list of resources required which were not stated in Appendix 1. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Glossary

PLePA – Premier League predictive learning algorithm, the name of the project.  
Premier League – The highest league in English football.  
Structured-case lifecycle – The lifecycle chosen for PLePA, it consists of multiple Conceptual Frameworks, described below.  
Conceptual Framework, also known as CF – A conceptual framework is a chunk of the lifecycle, it has a plan, process, analysis and reflection piece.

# Section 1 Preparation and planning

## 1.1 Project Description

The project is to create a predictive algorithm for Premier League results. It will use historic data; it will use data since the 2014-15 season. It is now a trivial task to add more data if it is deemed needed because the work has been done on creating the Python scripts to insert the data and notes have been made on cleansing the data, which is a couple of simple steps.

The result is aimed at Premier League fans and people interested in the prediction of football results. It’s also something I have an interest in solving, I do a bit of sports betting so this could prove useful for this.

The aim is to produce four separate algorithms and determine which has the most accurate results. The four algorithms I will use are Support Vector Machines (SVM), Random Forest, Naïve Bayes and KNN. The goal is to have around a 75-80% pass rate, this will be a benefit to all football fans, fans who place bets and even possibly teams, that may be a bit far though.

If the project is unable to give good results after many iterations for each algorithm and tweaking the numbers, then it will serve as research for people who take on a similar project. There is no real issue if it not developed because it is more of a personal project and I am the only stakeholder.

There are different ways to achieve the results, the aim is to develop four separate predictive algorithms, this will give the opportunity to move on if one is not going well.

The stats data will be stored in a MySQL database and all the algorithm code will be written in Python. All code and documents will be stored in a GIT repository. A stretch goal would be to have an interface for user interaction or to pull the latest set of fixtures from a football website.

## 1.2 Revisions made and reasons why

The main scope has not been updated, PLePA is still planning to consist of the four algorithms and there are still stretch goals. It is too early to tell if the stretch goals can be reached, this will become apparent in the next couple of months because the work on the algorithms will start. There is now a potential change that may need to be made but it is too early to tell, each algorithm in PLePA was going to be used to predict the next weeks’ worth of fixtures, however, given the current state of the world with COVID-19, there is no certainty whether there will be any fixtures being played upon completion. Other than that, there has been no uncertainty or changes to the scope so far, everything has gone to schedule.

The solution, without stretch goals being achieved, will produce graphs and stats to show how accurate each algorithm is. Each algorithm will use the data insert into the database to help it learn, then the algorithm will predict results, this may have to be past results due to unsure whether there will be upcoming fixtures imminently.

A report will be delivered at the end of the development of PLePA, it will give a comparison between the four algorithms to determine the most favourable. If stretch goals are achieved, there will be an interface to choose teams and the program will provide the result with likelihood of it happening, alternatively, or additionally, PLePA will have a button click which pulls in the latest fixtures from a football website, the user can then select which of the four algorithms they want to use to predict those results.

## 1.3 Tasks, subtasks, evaluation criteria

There was a couple of updates to the task list, the completions of TMA’s were removed because, as informed by my tutor, these are milestones to reach but not part of the project. A reflection task was also removed because it was part of the course but not the project. Some tasks were added, one of the tasks added was to review legal, ethical, social and professional, although this is a course task, it is also an important task for the project. Creating a database reset script was also a task added, this was realised to be needed after inserting the same data multiple times due to small mistakes.

The tasks start to follow the lifecycle chosen, structured-case lifecycle, which consists of multiple conceptual frameworks (CF), at Conceptual Framework 1 – setting up the database and data. An image of the life cycle choice can be seen below in figure 1, this is just a snippet and there can be more CFs.

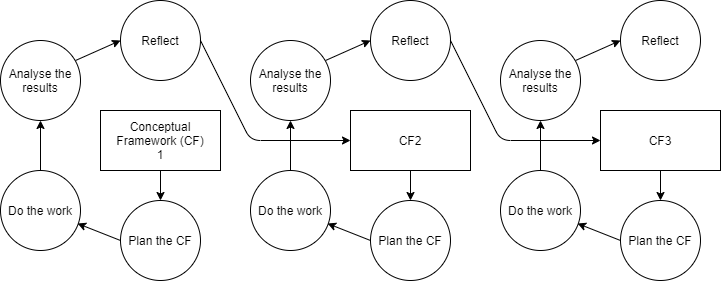


Figure 1 (Blagg, 2020) Structured-case lifecycle

* Define the goals and contents of my project.
* Research SDLC choices and decide on one for the project.
* Research difference between Oracle and MySQL and decide between the two.
* Think about how the database will be structured
* Look into Python modules which may be useful.
  + Distance calculation
* Set up base Python project with GIT version control
* Install database software
* Investigate similar studies for ideas. Document key findings.
* Find the best source for the Premier League statistics required.
* Investigate Machine Learning, ML, predictive algorithms and decide on 4 possibilities.
* Review legal, ethical, social and professional issues.
* Revaluate project after each TMA given feedback from tutor for TMA. Make sure project still makes sense.
* CF 1 – setting up the database and data
  + Gather data required.
  + Cleanse data.
  + Insert data into database.
  + Create a database reset script.
* CF2-5 – work on the four algorithms, each one is a separate CF.
  + Plan how algorithm will work.
  + Do some more research on top of what has already been done.
  + Code the algorithm
  + Test the findings
  + Evaluate
  + Produce graphs and report to show successfulness of the algorithms
* CF6 – develop a feed from a sports website to pull in the latest fixtures
  + This will feed the predictions for the upcoming fixtures and display on screen or email to a user.
* CF7 – develop a user interface
  + This will allow users to select two clubs and will display the predicted results
  + This could be developed further to allow the user to select which algorithm to predict with.

The progress to date has been updated for the evaluation criteria below in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Activity | Criteria | Evaluation | Progress to date |
| Premier league stats in the database. | Must have. | There will be a year’s worth of data in the database. | Five years’ worth of data has been inserted into the database, this activity has been fully completed. |
| Should have. | Three years. |
| Nice to have. | Five years. |
| A working predictive algorithm to predict the results. | Must have. | Two algorithms will have been developed and can be used. | The four algorithms have been chosen, SVM, Naïve Bayes, Random Forest and KNN. There has been work on researching the first algorithm to be developed, Random Forest. |
| Should have. | Three algorithms. |
| Nice to have. | Four algorithms. |
| An interface for user interaction. | Nice to have. | A user can load up an executable and select two clubs, they will then receive a prediction. | No progress. |
| Nice to have. | Further development could mean the user could select the algorithm to predict with and the results using the teams and the algorithm are displayed. | No progress. |
| A feed from a football website. | Nice to have. | The upcoming fixtures are pulled from a website and predicted using the chosen algorithm. | No progress. |

Table 1, Activity evaluation.

## 1.4 Updated resource list

The resource below, information from previous studies, has been added. This was overlooked as a resource, but it is very useful to see how other studies have investigated a similar problem. The initial resource list can be seen in Appendix 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resource | Why needed | When needed | Problems if not available | How to ensure availability |
| Information from previous studies | To understand what went well and not so well for similar projects. | Studied prior to any real task being completed, used throughout. | No knowledge from other teams experiences so PLePA may make similar mistakes as other studies which were not investigated. | Look on the internet and in the library early in the project. |

Table 2, Resource list

## 1.5 Updated skills list

There are no new skills identified for PLePA, the initial skill list can be seen in Appendix 2.

## 1.8 LEPSI review

|  |  |  |  |
| --- | --- | --- | --- |
| Act / Law / Guidance | Purpose | Relevance to my project or n/a and why not applicable | How my project will be affected |
| Data Protection Act 2018 (DPA) | Control how personal information is used by organisations, businesses or the government (UK Government, 2018) | I will not be using any personal information in my project, so this is not applicable for me. | My project will be unaffected. |
| Equality Act 2010 | Protection from discrimination in the workplace and society (UK Government, 2012) | I will not be looking at people in my project so this will not be applicable for me. | My project will be unaffected. |
| Computer Misuse Act 1990 | The act was brought into place to prevent people gaining unauthorised access to computer material, commonly known as hacking. (UK Government, 1990) | My project will likely not involve connection to the internet. It may do, if I reach my stretch goal to connect to a sports website for latest fixtures. | If I reach my stretch goal for connection to the sports website to get the latest fixtures, I will ensure the connection I have is secure and authorised. It likely will be authorised since the website will be publishing it. |
| Copyright | To protect your work by preventing it being copied, redistributed, adapted and put on the internet are some of the example (UK Government, n.d) | I will need to make sure the data that I use for past football results is not protected by copyright and can be used. | I will need to ensure I am able to use the data and not breaking copyright rules. |
| Freedom of Information Act | It gives the general public access to certain information on request from the public authorities. The public authorities are also obliged to publish certain information. (ICO, n.d) | This will not affect my project because I am not working for a public authority and I don’t require data from a public authority. | My project will be unaffected. |
| Protection from harm | Protect participants of studies from any harm, physical or psychological, particular care should be paid to children. (TM470 course team, 2012) | This will not affect my project because I’m not using participants. | My project will be unaffected. |
| Professional codes of practice and ethics | This allows participants to understand the purpose of the study, the researchers must state their intentions. (TM470 course team, 2012) | This will not affect my project because I’m not using participants. | My project will be unaffected. |
| BCS Code of Conduct | To set out standards across the board for all members of the BCS. An example is having respect for public health, privacy, security and wellbeing of others and the environment. (BCS, June 2019) | This will not affect me directly because I’m not a member of the BCS, however, I should try to comply with the standards set out by the BCS such as honesty with my skillset and acknowledgment to any borrowed source code used in my project. | I must ensure that borrowed code is acknowledged, and I am not overexaggerating my skillset. |

# References

UK Government (2018) *The Data Protection Act* [Online]. Available at <https://www.gov.uk/data-protection> (Accessed 1st April)

UK Government (2015) *Equality Act 2010: guidance* [Online]. Available at <https://www.gov.uk/guidance/equality-act-2010-guidance> (Accessed 1st April)

UK Government (1990) *Computer Misuse Act 1990* [Online]. Available at <http://www.legislation.gov.uk/ukpga/1990/18> (Accessed 1st April)

UK Government (n.d) *How copyright protects your works* [Online] Available at <https://www.gov.uk/copyright> (Accessed 1st April)

Information Commissioner’s Office (ICO) (n.d) *What is the Freedom of Information Act?* [Online]. Available at <https://ico.org.uk/for-organisations/guide-to-freedom-of-information/what-is-the-foi-act/> (Accessed 1st April)

TM470 course team (2012) *‘Legal, Social, Ethical and Professional issues’ in TM470 study material, The Open University, Milton Keynes* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1184585> (Accessed 1st April)

BCS, The Charted Institute for IT (June 2019) *Code of Conduct for BCS Members* [Online] Available at <https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf> (Accessed 1st April)

# Appendices

## Appendix 1 – Resource list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resource | Why needed | When needed | Problems if not available | How to ensure availability |
| Time | Complete any task for the project. | Throughout project. | The less time available, the fewer tasks can be completed. | Cannot ensure time but can try to get ahead on tasks in case of any complications. |
| PC | To complete all tasks, requires pc availability. | Throughout project. | Whilst no PC is available, most tasks will be unable to be achieved. | I have a desktop and laptop. As a last resort, I can also use my work laptop. |
| Python | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | It is installed on all PCs available. |
| MySQL | Store the data used for algorithms. | Set up in CF1 but will be required from CF1 – CF7. | The algorithms will have no database to read from. | It is installed on all PCs available. |
| Premier League data | The project is using data from the Premier League for the algorithms. | Pre CF1, it will then be stored in the MySQL database. | No data to be used for the algorithms. | Once gathered and stored in the database, store a copy of the database. |

## Appendix 2 – Skill list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Skill | Why needed | When needed | Problems if not available | How to ensure availability |
| Python coding | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | Research unused additional Python modules which are required. |
| SQL coding | Write the code for the database. | During CF1. | Unable to create the database structure and insert the data. | Used in everyday work, fluent in SQL. |
| Time management | Ensure milestones are met. | Throughout project. | Milestones may be missed; project will be incomplete. | Research good time management ideas and seek advice from people with good time management. |
| Report writing | To write TMA1-3 and EMA. | For all assignments but mainly the EMA. | The quality of writing for TMAs and EMAs will be lacking. | Research and check previous module’s advice on report writing and read course booklet on the subject |
| Researching | For several tasks to investigate best approach | Throughout project. | Project will suffer due to decisions being made without right information. | Look into ideas for researching, the best approaches to research and read course booklet on the subject |