|  |  |
| --- | --- |
| Computing Project Specification Form | **glyndwrLogo** |

**Student Surname: Collins email address: s18001869@mail.glyndwr.ac.uk**

**Student Forenames: Daniel James**

**Course: BSc Computer Science**

**Supervisor: John Worden, Bindu Jose**

|  |
| --- |
| **Project Title:** Domain Driven Data Mining: Predictive Analysis Techniques |

|  |
| --- |
| **Project Aims and Objectives:**  Develop a data pipeline consisting of the following steps:  - Takes information from various online sources and stores it in a SQLite3 database  - Clean this data, removing outliers and imputing missing data  - Generate features that have reasonable predictive power (features that tell us something about what we hope to predict)  - Use the features in a predictive model to predict the outcome of sporting events.  - Evaluate the predictions of this model |

|  |
| --- |
| **Research Areas of Study:**  predictive analysis, data science, programming, SQL |

|  |
| --- |
| **Methodology:**  I am going to be following a waterfall methodology. I have chosen this methodology as I believe my project has clear and fixed requirements; there is a clear divide between the stages of the project which will help me to organise work.  I will also be implementing a SCRUM methodology. I will conduct 30 day sprints; each sprint will be discussed with my project supervisor and recorded in a project supervision log book.  Project Outline:  - Undertake a literature review of available methodologies and tools used for predictive analysis techniques and models. Notably, the following topics will be explored:  - Data Storage  - SQLite3 database design  - Data Cleaning  - Handling missing data  - Outlier removal  - Feature Engineering  - Explore makes a good feature, and research state of the art feature engineering techniques  - Model building and evaluation  - Explore the best performing models in the industry, explore their strengths and weaknesses in relation to this task  - Explore methods of evaluating models, including developing an effective baseline to compare models against.  - Evaluate the data sources available online, that could be used in the project. Design a database structure that can support this data in the optimal way.  - Design scripts to extract data from online sources and store it in the SQLite3 database.  - Develop scripts for cleaning the data and generating features  - Pick a list of candidate models, and test each model at predicting the outcome of sporting events. Evaluate these models against a baseline and pick the best performing one.  - Provide a set of documentation and a user-guide for the model  - Consider how the model could be improved further in the future |

|  |
| --- |
| **Project Outcomes and Deliverables:**  A report detailing the stages of my project  A poster demonstrating my project and progress  A github repository consisting of python scripts and an sqlite3 database |

|  |
| --- |
| **Project Timetable:**  The final report submission date is Friday 7th May and the project demonstration is on Thursday 13th May. I will be liaising with my project supervisor on a weekly basis, where I will ask any questions as well as providing updates on progress. Information from meetings will be recorded in the project log book.  Prepare poster for poster party: 8 days  Write literature review: 20 days  Analyse and compare tools: 7 days  Design and create SQLite3 database: 15 days  Design and develop Python model: 25 days  Implement, evaluate and test model and provide recommendations: 15 days  Document model and create user-guide: 8 days  Consider future adjustments and uses: 7 days  Prepare for demonstration of artefact: 10 days  There is also a MS Project file detailing the project schedule. |

**Student Signature: D.Collins Date: 04/12/2020**

**Agreed by Supervisor**