# General Information

This is an INDIVIDUAL piece of work contributing towards the module assessment. Deliverables should include access guide to model deployment and a report document, which includes (critical discussion and analysis, research, and snapshot evidences of tasks carried out and justification of techniques used). Submission will be in the form of a MS word report (4000 words).

**PART A: FIFA World Cup 2022**

(The original case study and dataset can be found from Kaggle). The FIFA World Cup is the most prestigious football tournament in the world. The championship has been awarded every four years since the start of the tournament in 1930. In the tournament, 32 teams, including the host nation, compete for the title at different stadiums in the host country. This year, Qatar will host the 2022 tournament, for which the first match will be played in November 2022. Before the start of the tournament, to predict the winner of the tournament is an exciting discussion always and if this is done through machine learning, it is most likely that it is going to be true. To accomplish this task, the dataset provided includes a complete overview of all international soccer matches played since 90s. On top of that, the strength of each team is provided by incorporating actual FIFA rankings including several other attributes. The original data can be downloaded from the following URL (Please signin/signup using person or organizational email to download).

The original dataset contains several extra attributes and missing values. As ML developer, this is your task to analyze and preprocess the data using appropriate machine learning techniques. Later, to propose and develop a machine learning model (an application deployed on cloud) **to predict the winner of the tournament. You prediction may be a**

**probability of winning for the five teams.** In order to achieve this, you are required to accomplish following task on the AWS.

A typical workflow of machine learning on the cloud involves, data preprocessing, selection and training of model, model evaluation and deployment for prediction. During coursework, instructor will guide you how to perform different machine learning tasks on the AWS.

**Note:** This case study is based upon a challenging dataset provided on *Kaggle.*

# Assessment Criteria/Marking Scheme:

The work will be marked out of 100 in line with the University’s marking grades and according to the following assessment criteria:

|  |  |
| --- | --- |
| **Description** | **Marks** |
| **Task 1: Cloud feasibility study – 10% (suggested word limit for this section is 500 words)** | |
| In your report, explore and analyse different cloud platform for commercial machine learning solutions deployment in terms of feasibility to the given scenario. The discussion should be based on state-of-the-art literature.  aws collab | 10 |
| **Task 2: Data analysis and opportunity identification – 10% (suggested word limit for this section is 500 words)** | |
| In your report, explore and present facts about given data and a comprehensive discussion on similar problems and techniques used to solve those.  Graphs analysing | 10 |
| **Task 3: Data pre-processing – 20% (suggested word limit for this section is 1000 words)** | |
| Generate data for the model training after executing appropriate pre-processing steps like wrangling, ingestion and transformation etc. using AWS Data Wrangler UI or Python. Only perform whatever steps are essential and provide reasoning for that  in the report.  Answer according to collab not aws  Whatever you have done in the pre processing | 20 |

|  |  |
| --- | --- |
| **Task 4: Model selection and training - 30% (suggested word limit for this section is 1000 words)** | |
| Select an appropriate model and train it, Using the SageMaker Python SDK or Use Amazon SageMaker Built-in Algorithms. The SDK includes lot of different machine learning models, choose whatever is more appropriate to the problem and outcome of analysis performed in Task 1. Discuss the whole process in  detail. | 30 |
| **Task 5: Model evaluation and visualization - 15% (suggested word limit for this section is 500 words)** | |
| Evaluate the model, once you are done with training, evaluate it for the overfitting  etc. You can see what measure or visualization would be appropriate to showcase the performance of chosen model using following link. Discuss results.  Rc auc confusion matrix | 15 |
| **Task 6: Model deployment - 15% (suggested word limit for this section is 500 words)** | |
| Deploy your model using SageMaker Hosting Services (desired) or discuss step by step procedure for deployment without using SageMaker. Discuss in detail what best professional practices can be employed to comply with ethical and  privacy concerns.  Go to sage maker to get the discussion on the procedure - deployment procedure | 15 |
| **Total** | **100** |

**PART B: Big Data Analytics (Python)**

# General Information

This is an INDIVIDUAL piece of work contributing towards the module assessment. Deliverables should include access guide to model deployment and a report document, which includes (critical discussion and analysis, research, and snapshot evidences of tasks carried out and justification of techniques used). Submission will be in the form of a MS word report (4000 words).

**Case study:**

**Section 1:**

You are required to analyze the dataset “***Billionaires.csv***” using Python Tools for Big Data Analysis (The dataset can be found in the Assessment folder). Perform data analysis according to the regions and cities. As an outcome of your analysis, you have to answer the following research questions:

What are the top 10 countries with the highest number of billionaires?

What industries/sectors are most successful?

What are the main industries with the highest number of women billionaires?

What age range represents the highest and lowest number of billionaires?

Describe and analyse the additional factors that might influence wealth, e.g. inheritance.

Individual tasks which are required to be discussed in detail are presented below.

**Task 1: Problem Domain, Data Description, and Research Question (15 Marks).**

* Research state-of-the-art literature and describe the application domain to which the dataset belongs and provide a detailed description of the dataset.
* Formulate an appropriate hypothesis (NULL and alternative) based on correlation or other metrics, which can be tested using different statistical methods.

**Task 2: Solution Exploration (15 Marks)**

* Evaluate a range of approaches and technologies for developing big data applications.
* Describe the solutions and techniques applied to a similar problem.
* State your chosen methodological approach with justification.
* Support your research with relevant references.

**Task 3: Solution Development (35 Marks)**

* Perform data pre-processing if it is required.
* Provide a general analysis of your dataset (descriptive statistics) and answer the research questions. Include graphs and other types of visual representation that give essential information about the data components.
* Describe in detail the steps you have taken to reach your solution to the selected problem. Support your analysis with data visualization. Provide screenshots of the codding parts where appropriate.
* Present the results clearly and explicitly.
* Perform statistical significance testing on your data to accept or reject the NULL hypothesis.

**Task 4: Evaluation and Future Development (15 Marks)**

* Describe any potential impact of your results and how they can be used within the application domain.
* Include considerations for any limitations of your work.
* Detail a future work with an explanation of how the solution could be enhanced and developed further in terms of the latest technologies.

**Section 2 – Business Intelligence (Tableau)**

As a Data Analyst of the R&D of Amazon, you were asked to analyse the dataset obtained from Walmart, one of the leading supermarkets in the USA. The Dataset named “***Electronic Sales***” contains the product name, price, and order details (The dataset can be found in the Assessment folder).

You must use Tableau for the dataset analysis and for the completion of the following tasks.

**Task 1:** Using the parameter function, display a chart showing ten top-selling and ten low-selling products. **[3 marks]**

**Task 2:** Using an appropriate chart, display the Total/Average sales made in each city. The chart should allow the user to switch between the Sum and Average Sales. Place the result in descending order. **[3 marks]**

**Task 3:** You are asked to provide the stores’ performance on a weekly basis. These charts should demonstrate the average and total sales. **[4 marks]**

**Task 4:** The ordered items come with 6 months warranty from the date of purchase. Use the Date function to calculate the Warranty End date. **[4 marks]**

**Task 5:** Create an interactive dashboard with at least four sheets. The interactivity means that if you click on any sheet of the dashboard, the other sheets should simultaneously display the related information. **[6 marks]**

**Please note that each task should be supported by screenshots that demonstrate the stages of your process.**