MCAST

ASSESSMENT AND INTERNAL VERIFICATION FRONT SHEET (Individual Criteria)

(Note: This version is to be used for an assignment brief issued to students via Classter)

Course Title	IT4-A5-21-Advanced Diploma in IT (iGaming)		**Lecturer Name & Surname**	**Frankie Inguanez**	
Unit Number & Title	**ITIGM-406-2107 Advanced Data Analytics and Development**				
Assignment Number, Title / Type	**2, Data Analysis Case Study**				
Date Set	08/01/2024	**Deadline Date**	**22/01/2024**		
Student Name		**ID Number**		**Class / Group**	**IGA-4.2A**

Assessment Criteria	Maximum Mark
AA1.4 Calculate various types of data analysis.	7
AA2.2 Solve various types of data analytical procedures.	7
AA3.2 Organise data using a database management system and a programming language.	7
AA3.3 Produce data visualisations from one or more data sets within the iGaming sector.	7
SE3.4 Compile a report based on specific KPI's.	10
KU4.1 Identify the steps required in the data analytics methodology and related workflow.	5
AA4.3 Develop a statistical analysis report through the use of a programming language and an interactive computing platform.	7
SE4.4 Construct a workflow and a schedule plan for identifying and extracting specific information based on the given requirements.	10
Total Mark	60

Notes to Students:

- This assignment brief has been approved and released by the Internal Verifier through Classter.
- Assessment marks and feedback by the lecturer will be available online via Classter (Http://mcast.classter.com) following release by the Internal Verifier
- Students submitting their assignment on Moodle/Turnitin will be requested to confirm online the following statements:

Student's declaration prior to handing-in of assignment

- ❖ I certify that the work submitted for this assignment is my own and that I have read and understood the respective Plagiarism Policy

Student's declaration on assessment special arrangements

- ❖ I certify that adequate support was given to me during the assignment through the Institute and/or the Inclusive Education Unit.
- ❖ I declare that I refused the special support offered by the Institute.

Scenario

You have been engaged to create a predictive model that is able to determine the overall score of football players in the FIFA game. The dataset is being provided to you on a GitHub repository. You are free to use any assistive tool of your choice, even Generative Artificial Intelligence (GenAI) tools such as OpenAI ChatGPT or Google Bard. If using GenAI tools, it is important to keep 1 thread for your entire assignment and you are to provide a copy or share a link of the entire thread with your submission.

Tasks

1. Create a **GitHub** account if you do not already have one. Create a private Python repository for this assignment and name it: **com.mcast.adad2023.a02.<surname>_<name>** such that Joe Borg would name it: **com.mcast.adad2023.a02.borg_joe**. Make sure you set the visibility to private. Share with your lecturer with username **FrankieInguanez**. You are to commit your work regularly.
2. Create a Jupyter notebook and prepare several markdown cells for every stage/step in the machine learning predictive **data analytics methodology** (Data Acquisition, Data Cleaning, Data Exploration, etc.). The next steps will request that you add code cells for each step.
3. In the **data acquisition** step:
    a. Research how you can download a file on a git repository via Python. Adapt the code so that you download the following file: https://raw.githubusercontent.com/frankieinguanez/com.mcast.adad2023/main/football_striker_score.csv .
    b. In a markdown cell document what you needed to do in order to download the file, including installation of any specific packages for your Python environment.
    c. In a code cell write code to display the first 5 rows of the dataset.
    d. In a code cell write code to display some general information such as total rows, total columns, column names, non-null count and data type per column name.
4. In the **data cleaning** step:
    a. In a code cell write code to display any issues with the data (presence of null values).
    b. In a separate code cell remove any columns that have no values. Display information about the data frame to confirm.
    c. In a separate code cell remove columns that are irrelevant (the identifier column). Display information about the data frame to confirm.
    d. In a separate code cell display rows that have null values. Remove the rows with null values. Display information about the data frame to confirm.
5. In the **data exploration** step:
    a. In a code cell write code to extract statistical information about each variable.
    b. In a markdown cell write the minimum, maximum and average age.
    c. In a code cell write code to display the statistics of the top 10 overall ranking players?
6. In the **data visualisation:**
    a. In a code cell display histograms and scatter plots for each variable.
    b. In a separate code cell display boxplots for each variable.
    c. In a markdown cell identify which variables have outliers and which do not.
    d. In a separate code cell display the correlation between each variable.

e. In a markdown cell provide an interpretation of the correlation between each predictor with the target variable (overall). Remember that a variable can have no, weak, moderate or strong correlation with another based on a range of values. Also, the correlation can be positive or negative.

7. In the **data modelling** step:
    a. In a code cell write code to create a multiple linear regression model using all predictors, where you predict the overall score of each player. Following the creation of the model write code to display statistical information about each model.
    b. Write a markdown cell to determine with justification if the model is accepted, what percentage of the variation in the target is explained and which predictor to drop for the next model with justification.
    c. Write another code cell to create your second multiple linear regression model with the reduced predictors. Display information about the model.
    d. Write a markdown cell to determine with justification if the model is accepted, what percentage of the variation in the target is explained.
    e. In another markdown cell choose the best model with justification.
8. In the **prediction step:**
    a. Write a code cell to create a function that will be able to predict the overall score of a player.
    b. Write another code cell that will print the overall score of a player using the previous function and the following statistics (if your ideal model does not use any one of the provided variables then you can omit it):

Age	Acceleration	Aggression	Agility	Balance	BallControl	Composure	Crossing	Dribbling	Finishing
32	90	78	86	87	88	86	77	82	91

Grading Criteria

Criteria	Task	Marks
SE4.4	Create a GitHub repository for this assignment. Share with your lecturer	_ / 4
	Clone the repository on your system and commit your work daily. Include a meaningful comment with each commit.	_ / 6
KU4.1	Create a Jupyter notebook with markdown cells representing the steps of the data analytics methodology.	_ / 5
AA3.2	Research how to download a CSV file from an URL and adapt for your project.	_ / 2
	Display information about the dataset.	_ / 1
	Display the first 5 rows of the dataset.	_ / 1
	Add all researched content in repository or a link to the GenAI thread.	_ / 3
AA2.2	Identify the presence of null values.	_ / 1
	Remove any columns that are null.	_ / 1
	Remove any rows that have null values.	_ / 1
	Remove any redundant columns.	_ / 2
	Add all researched content in repository or a link to the GenAI thread.	_ / 2
AA1.4	Write code to generate statistical information to answer question 5a.	_ / 2
	Write a markdown cell to answer question 5b.	_ / 2
	Write a code cell to answer question 5c.	_ / 1
	Add all researched content in repository or a link to the GenAI thread.	_ / 2
AA3.3	Display histograms and scatter plots for each variable.	_ / 1
	Display boxplots for each variable.	_ / 1
	Display the correlation across each variable.	_ / 2
	Add all researched content in repository or a link to the GenAI thread.	_ / 3
AA4.3	Interpret the correlation of each predictor against the target.	_ / 2
	Create 2 multiple linear regression models.	_ / 2
	Display descriptive information for each model.	_ / 1
	Recommend the ideal model with justification	_ / 2
SE3.4	Create a function that used the parameters of the best model to predict.	_ / 5
	Predict the overall score for the provided statistics and display output.	_ / 2
	Add all researched content in repository or a link to the GenAI thread.	_ / 3