**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | MSc. In Data Analytics |
| **Assessment Title:** | * Programming for DA * Statistics for Data Analytics * Machine Learning for Data Analysis * Data Preparation & Visualisation |
| **Lecturer Name:** | * Sam Weiss * Bharathi Chakravarthi * Muhammad Iqbal * David McQuaid |
| **Student Full Name:** | Maria Anany Casas Manosalva |
| **Student Number:** | 2023326 |
| **Assessment Due Date:** | 10-11-2023 |
| **Date of Submission:** | 10-11-2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

|  |
| --- |
| The present information of this project has been collected by The Central Statistics Office in Ireland, from the page <https://data.cso.ie> following the route: “Browse”, “Environment”, “Forestry”, “Afforestation Area”.  It is important to mention that the main goal of this report is to predict the area that will be afforested in the next years (2023-2030).  Explain the targets proposed by the European Union (8 000 ha per year) |

|  |
| --- |
| Programming for DA  First of all, it is necessary to do Exploratory Data Analysis (EDA) which is a methodology used in data science and statistical analysis to explore, analyse and visualise datasets to understand their structure, patterns and main characteristics. To do so, we will follow next steps: loading the data, exploring the information, debugging the data, final data processing, analysing the data.   1. Loading the data   In [1]: Import Pandas library which is used for data manipulation and analysis. It works with DataFrames which allow you to efficiently load, clean, transform, and analyse data.  In [2]: Read the CSV file named 'Afforestation\_area.csv' and store the data in a DataFrame named 'afforestation\_df' (convert the CSV file in a DataFrame).   1. Exploring the information:   In [3]: Display the object type of the 'afforestation\_df' which should be a DataFrame.  In [4]: The shape method shows the number of rows and columns of the DataFrame (5184 rows, 7 columns).  In [5]: The info method describes the data type of each column and the number of non-null values of each one. We can see that there are 5184 entries and 3687 non-null in the ‘VALUE’ column, so we have to decide what we will do with this missing values.  In [6]: the head method shows, in this case, the first 10 rows of the DataFrame.   1. Debugging the data:   In [7]: Delete the rows that contain 'Total Afforestation' since I want to analyse the plantations carried out for each specie and forest owner instead of analysing them in their entirety, this is to avoid generating redundancy or noise.  Filter the DataFrame to remove rows where the columns 'Species' or 'Forest Owner' contain 'Total Afforestation' and update the DataFrame with the requested conditions.  In [8]: Remove columns 'Statistic Label' and 'UNIT' due to they both have the same value: 'Afforestation Area' and 'Hectares' respectively.  The drop method deletes rows or columns in a DataFrame, axis=1 (remove columns instead of rows), inplace=True (modify the DataFrame in place).  In [9]: Rename the columns 'VALUE' and 'Forest Owner' creating a dictionary for then use the method rename to rebrand columns named 'VALUE' to 'Value\_ha' and 'Forest Owner' to 'Forest\_Owner'.   1. Final data processing   It can be shown that there are many NaN or null values in the 'Values\_ha' column (in the csv file they appear as empty data) and most of this information comes from 'Non-Farmer' and 'Public Sector'. Initially, I decided to replace these values to 0 because the afforestation mainly comes from 'Farmer' who receive loans for this work, but when I did the normalization, I got a high frequency of 0 values (about 1000 values) so it is better to remove the rows of null values to avoid this interference.  In [10]: The dropna method drops rows or columns with NaN values in a DataFrame, axis=0 (remove rows instead of columns).  In [11]: Create a dictionary to change some names shown in the 'County' and 'Forest\_Owner' columns.  In [12]: The replace method substitutes some values in the ‘County' and ‘Forest\_Owner' columns according to those found in the dictionary.  In [13]: Call the info method again to see the new number of rows obtained and if there are still null values, observing that the DataFrame contains 1594 entries or rows but the index values start from 5 to 5182. It means that the indexes need to be reset since rows have been deleted.  It can also be seen that the data type of each column is in accordance with their characteristics and we don't need to perform a conversion.  In [14]: The method reset reboots indexes to start from 0, drop=True (clear all indexes).   1. Analysing the data   In [15]: Import Numpy library which is used for used for mathematical calculations, statistics and lineal algebra operations. It works with n-dimensional arrays that are suitable for numerical operations.  Import Matplotlib library that is a powerfull tool used for creating 2D visualizations, plots and for representing data visually.  Import Seaborn that is a data visualization library built on top of Matplotlib and it simplifies the creation of informative and visually appealing statistical graphics.  In [16]: Create a bar chart to visualize the “Total Afforestation per County”, except for Ireland, by following these steps:   * Make a graph and set its size to 12 (wide) x 10 (high) * ‘Filtered’ variable is a boolean string that contains True if afforestation\_df['County'] != 'Ireland' * Filter the data in the 'County' column except 'Ireland' and calculate the sum of the values in the 'Value\_ha' column for each 'County' group * Generate a bar plot with the data contained in the 'afforestation\_county' variable, kind='bar' (create a bar chart), figsize=(12, 6) (Create bar chart dimensions: 12 (wide) x 6 (high)) * Set the graph title as "Total Afforestation by County from 2007 to 2022" with a font size of 20 * Set the x-axis label to "County" with a font size of 15 * Set the y-axis label to "Afforestation (ha)" with a font size of 15 * Rotate the x-axes labels by 90 degrees to improve readability * Prop = {'size': 10} (set the font size to 10) * Display the created bar plot   In [17]: Create a bar chart to visualize the “Total Afforestation by Counties per Year”, except for Ireland, by following the same structure as the previous chart with the difference that the data is grouped by ‘Year’ instead of ‘County’.  In [18]: Generate a bar plot to show the “Total Afforestation in Ireland per Year” to see if there is a significant difference with the “Total Afforestation by Counties per Year”. |

|  |
| --- |
| Machine Learning for Data Analysis  Machine Learning (ML) is a branch of artificial intelligence that allows the progressive improvement of tasks through machine learning with computers, which through algorithms and models can learn patterns and make decisions without being explicitly programmed. ML can be classified into 3 categories:   1. Supervised learning: the algorithms work with labeled data trying to find a function that, given the input variables, assigns them the appropriate output label. 2. Unsupervised learning: occurs when labeled data is not available for training. We only know the input data, but there are no output data that correspond to a certain input. 3. Reinforcement learning: This type of learning is based on improving the response of the model using a feedback process that it obtains from the outside world in response to its actions.   For my project, I want to predict the total afforestation in Ireland by county and per year (Value\_ha will be the variable to predict).  The type of Machine Learning that I will use for this assignment is Supervised since I have past information where I can train the machine so that it can predict future values. I will also use the Regression Model since the results I want to obtain are numerical (hectares to aforest = Value\_ha). |