Assessment Brief

Module Title:	Tools for Data Analytics
Module Code:	B8IT106
Assessment Title:	Continuous Assessment – Python
Assessment Type:	Python Practical Assignment
Individual/Group:	Group (2 or 3 members)
Assessment Weighting:	20%
Issue Date:	29/09/2019
Due Date/Time:	**Deadline – 11:55 pm on Sunday 20/10/2019**
Mode of Submission:	MOODLE

Learning Outcomes to be assessed

1. Ability to analyze datasets using Python

Assessment Overview

This assignment is divided into two parts. Each part requires you to build a machine learning model using the given dataset.

Part 1: Random Forest **(60 Marks)**

Dataset

The dataset ***Spruce.csv*** contains cartographic data for observations made over different 30m × 30m patches in the forests of Alberta, Canada. This dataset has 15,120 observations, with 44 input variables (cartographic variables) and 1 target variable (Tree_Type). Description of these variables is as follows:

Elevation - Elevation in meters
Slope - Slope in degrees
Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway
Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points
Soil_Type (38 binary columns, 0 = absence or 1 = presence) - Soil Type designation
Tree_Type - Tree Type designation

The target label Tree_Type can be either Spruce (meaning Spruce tree was found predominant in the observed patch) or Other (meaning trees other than Spruce were found predominant).

Task

Canada's forest department wants to build a classification model to predict locations of Spruce trees. Spruce trees are of particular interest because of their several uses in medicine, timber and paper industries. Using the given dataset, construct a classification model using random forest in Python.

Perform all necessary data preparation steps that you may deem fit and choose an appropriate train-test split. Perform necessary hyperparameter tuning to construct an optimal model. Ultimately, arrive at your best random forest model that incorporates a subset of significant features.

In addition to providing the python code file, you are required to provide analysis of your approach and results in a pdf report.

Your code and analysis should cover the following points:

1. Data Preparation. [10]

2. Model Evaluation Strategy – Are you focusing on classification accuracy or false positives or false negatives for model evaluation? You should provide logical justification behind your approach. [15]

3. Model Building and Testing – tuning, performance testing for model built using all features and for model built using a subset of significant features. [10]

5. Identifying Best Model – What is your final (best) classification model? What are its features and parameters? How does it perform on test set? [15]

6. Generating Guidelines. Based on information obtained about predictive power of various features used in your final model (using feature_importances), what guidelines can you prepare for Alberta's forest department for quick identification of possible locations with Spruce trees? [10]

Part 2: PCA & K-Means Clustering **(40 Marks)**

Task

Implement PCA and K-Means clustering on the given dataset.

In addition to providing the python code, you are required to provide analysis of your approach and results in a pdf report.

Your code and analysis should cover the following points:

1. Data Preparation. [5]

2. PCA Implementation to visualize dataset. How much variance is explained by the first two principal components? Is the resultant plot a good representation of data? [10]

3. Elbow Plot Creation. What does the elbow plot tell you about the number of clusters (K)? [5]

4. K-Means Implementation. What do you reckon are the correct number of clusters in the dataset? What are those clusters in your opinion? Did elbow plot help you in any way? Did PCA visualization help you in any way?

[20]

Students must submit: -

1. A single Python Code File (.py) covering Part 1 and Part 2 of the assignment. It should be named as –

Surname1_Surname2_Surname3.py

2. Data Analysis Report (.pdf) containing description of approach, results and conclusions for Part 1 and Part 2 of the assignment. It should be named as –

Surname1_Surname2_Surname3.pdf

Both files should be put into a zipped folder, which should be named as –

Surname1_Surname2_Surname3.zip

Submission Guidelines

1. There is no prescribed word limit for Data Analysis Report. Use your own judgement, and cover all the above-mentioned points in a clear and concise fashion**.**

2. Include full names and student numbers of all group members in both files. In the code file, you should mention these as comments. In the report, you should mention these on the cover page.

3. Only one submission per group is required. Any one group member can submit the assignment.

4. It is the responsibility of students to form groups of their choice. Each group can have 2 or 3 members. No group can contain more than 3 members. Any submission with more than 3 names will be liable to a 25% grade penalty per additional member. Note that all group members will receive same marks and same penalties (if any).

Assessment Criteria

Each part will be graded according to the following criteria:

1. Quality of code (correctness and completeness) [Weightage – 40%]

2. Quality of analysis in report (description of approach, presentation and interpretation of results, conclusion) [Weightage – 60%]

General Assessment Submission Requirements for Students:

1. Online assignments must be submitted no later than the stated deadline.

2. All relevant provisions of the Assessment Regulations must be complied with. Students are required to refer to the assessment regulations in their Student Guides and on the DBS Quality Assurance Handbook Guide.

3. Extensions to assignment submission deadlines will be not be granted, other than in exceptional circumstances. To apply for an extension please contact the course administrator.

4. Students are required to retain a copy of each assignment submitted.

5. Assignments that exceed the word count will be penalised.

6. Dublin Business School penalises students who engage in academic impropriety (i.e. plagiarism, collusion and/or copying). Please refer to the referencing guidelines on Moodle for information on correct referencing.

Late Submission

- Assignments submitted after the deadline published in the assessment specification, including any extension, are deemed to be 'late' and are penalized as follows:
- Where an assignment is submitted between 1 and 14 days late 2 a day marks are deducted.
- Where an assignment is more than 14 days late it is annotated at the discretion of the lecturer, but no marks can be awarded.
- Where the assessment is undertaken in a group, the piece of work should be submitted in its complete entirety, and any penalty for late submission incurred applies to all group members.