Assignment 3 (25 marks)
Programming for Data Science Module

You are required to perform a multi-class classification task on a rock dataset. Please download the data file in csv through the link in the assignment description. Read the data description and learning task below carefully. Then, complete 3 questions with Python programs. For each question, write python programs to solve, and include explanation on your results and findings as indicated in the question. Scikit-learn, Scipy and Pandas are likely to be used for implementation.

About the data set:
This is a real-world rock dataset that compiles more than 1700 magmatic rocks in southern Peru and northern Chile. These data samples document compositional variations of magmas since Jurassic time with a focus on the Neogene period, when major crustal thickening developed and its influence on magma composition was most pronounced. The dataset has 1788 rock samples (i.e. rows) and 101 descriptive features (i.e. columns) in total, and many have missing values.

About the classification task:
The task is to use machine learning algorithms to automatically categorize samples into several types -- ash, lava, ignimbrite, etc. In the data file, it corresponds to the "Sample_type" column.

About submission:
- Submit one Jupyter notebook file (.ipynb) only, including both codes and explanations required in each question (i.e. put codes in "code" cells, and descriptions in "markdown" cells).
- The deadline is **2nd of December (Thursday) 9am.**

Question 1. Read data into your python program. Perform data cleaning and feature engineering techniques to:
1) make sure there are no samples having missing values. (5 marks)
2) transform all the non-numeric features into numbers. (5 marks)

If you delete certain rows or columns, clarify what rows/columns are selected for deletion and why. For any feature transformation, indicate what transformation technique is used on which columns.

Question 2. Choose TWO algorithms in Scikit-learn for the classification task described above. Discuss which model performs better by:
1) Build one model from each algorithm using its default hyper-parameter setting based on 10-fold cross validation. (5 marks)
2) Compare the two models by discussing TWO performance metrics. The comparison should be based on an appropriate statistical test. Explain why you choose this test, and whether and why one model is significantly better than the other. (5 marks)

Question 3. Propose and implement a strategy to improve ONE performance metric of ONE of your models from Question 2. State briefly what strategy you use, why you believe the

strategy may work, and whether your strategy is effective based on the statistical test. (5 marks)

Please note that you will NOT get penalty, if the final statistical test result shows your strategy does not significantly improve performance.