Project Overview

You are provided with a dataset chocolate.csv on chocolate bars. Your goal is to develop a machine learning model which takes the properties of a specific chocolate bar (e.g. the percentage of cocoa, the origin of beans), and output the rating. The dataset contains the relevant information of a number of chocolate bars, along with expert ratings as the ground truth.

Data source:
The dataset is from Brady Brelinski, Founding Member of the Manhattan Chocolate Society. The data is also used in a Kaggle competition.

Column Description:

#	Column Header	Description of the data
1	Company (Maker-if known)	name of the company (string)
2	Specific Bean Origin	the geographical origin for the chocolate bar (string)
3	REF	a value indicating when the review was entered in the database. A higher value indicates more recently entered (integer)
4	Review Year	the year of the review published (integer)
5	Cocoa Percentage	cocoa percentage of the chocolate bar (string)
6	Company Location	the country of the manufacturer (string)
7	Rating	expert rating for the chocolate bar (float). This is the label to be predicted by the model. It is a number from 1 (lowest quality) to 5 (highest quality)
8	Bean Type	the type of cocoa bean used (string)
9	Broad Bean Origin	the broader geographical origin of the cocoa bean (string)

Dataset Dimension:

- Samples (rows): 1500
- Attributes (columns): 9 (including the target: rating)

Tasks

Your team will need to accomplish the following tasks. You should apply the suitable techniques covered in the lectures and tutorials.

1. Perform **data pre-processing**. This includes but is not limited to checking typos, dealing with missing values and creating dummy variables.
2. Conduct other analysis to explore the data. For example:
   - Identify the most predictive attributes.
   - Map out the chocolate rating geographically on a map.
3. Formulate the problem as a machine learning task.
4. Select **TWO** learning algorithms based on the previous task.
5. Perform data partitioning. This will split the data into the training data and the test data. The training data will be used for **model development**, with the test data for **performance evaluation**.
6. Perform **model development.**
   - List all your learning algorithms.
   - Assess each learning algorithm on the training data.
7. Perform **performance assessment**
   - Apply model M on the test data to get the prediction.
   - Calculate the accuracy and the confusion matrix.