**Research Problem:**

The focus was to predict the rating of the chocolate based on the features such as characteristics and ingredients of the chocolate by using various regression models and selecting the best model’s prediction. Just to note this project was done in Jupyter Notebooks using Python.

**Exploratory Data Analysis:**

As we already have the graphs in the notebook just to elaborate, the first chart was the correlation heat map followed by the bar charts. Now coming to the bar charts, the initial chart is about what ingredients of the seven are used in the majority, looking at which it was clear that Sweetener, Sugar, Beans as well as Cocoa Butter made the chocolate likable. After that is the chart that tells us what characteristics made the chocolate likable or unlikable. Then there is the comparison between the countries as well as the country of origin, the first graph depicts that most of the chocolate companies are in the USA than the other. The second graph tells us the country of origin of the beans. The next graphs are the comparison between company location vs rating and number of ingredients vs rating.

**Data Tidying and Methods used:**

The data was loaded as a CSV after downloading it from Google, then checked the shape of the dataset which had 2130 rows and 33 columns,

The next step was to check if we had any NULL values or not, after the observations it was clear that there were 70 Null’s in the column named “number\_of\_ingredients”. Then all the NULL values were replaced with ‘0’ using the fillna() function, we also found that the number of ingredients was not accurate because there is one ingredient which is “Sweetener” was a part of all the chocolates that the dataset had. So instead of saying that chocolate has ‘0’ ingredients, we added ‘1’ to the column which now indicates that every chocolate now has one ingredient which is the sweetener. After this operation on the column if there was 0 in the column that means now it’s 1, if there was 1 now means the updated value is 2, and so on. One issue that we faced while performing this operation was, we were not able to add 1 to the column because it was a float data type initially, to overcome this we then changed the datatype of the column to integer (int).

Two other analytical or transformation techniques we used for this task were Chi-Square and Label Encoder. A little elaboration on them is as follows,

**Chi-Square Test:**

A statistical formula for comparing two or more statistical variables is the chi-square formula. It is used for data that include variables spread over several categories and is indicated by the symbol "2."

**Formula:**

χ2 = ∑(Oi – Ei)2/Ei,

χ2 = chi-squared

{O}\_i = observed value(actual value)

E\_{i} = expected value

This analysis was used to extract what features are dependent and independent, then we considered only the dependent variables and considered them as our initial training set.

**Label Encoder:**

Label encoding is the process of transforming categorical variables into a binary form so that it is easier for models to understand the data, especially for regression. The operation of those labels can then be better determined by machine learning techniques. It is a significant supervised learning data transformation step for the structured dataset.

**Regression Models:**

Different models used for the prediction in our analysis were Linear Regression, SVM, KNN Regressor, Decision Tree Regressor, and the Gradient Boosting Regressor. Root Means Squared Error as well as Mean Squared Error was used as the metrics for checking the accuracy of the models. As per my understanding and knowledge, it is said that if the achieved value is in the range of 0.2 to 0.5 then the model is performing well, so as per this statement all the model’s accuracy was in the range. In comparison, among them, Gradient Boosting Regressor performed better than the others with an RMSE of ‘0.39’.

**Results:**

Linear Regression: RMSE – 0.39, MSE – 0.15

SVM Regressor: RMSE – 0.42, MSE – 0.18

KNN Regressor: RMSE – 0.44, MSE – 0.19

Decision Tree Regressor: RMSE – 0.46, MSE – 0.21

Gradient Boosting Regressor: RMSE – 0.39, MSE – 0.21

Submission CSV:

As requested, this CSV file has around 401 rows including the column names

**Future Work:**

Due to the time constraint, the analysis was at the base level, in future if given a chance, including the hyperparameter tuning as well as choosing what are the best parameters using Grid Search CV (which takes time) are the concepts that can be considered to enhance the performance of the models. One more point to add normalization is also one of the techniques which can be included.

GitHub Link to the project is given below, It will be really appreciated if someone from the reviewing or the judging team can take a look at the work done by our team. Thanks.

[DataIku Hackathon Project](https://github.com/pranav3993/DataIku-Hackathon-Pace-2022/blob/main/COCO.ipynb)