**Introduction:**

The project is an analysis of the dataset of chocolate bar ratings using supervised machine learning techniques. The algorithm used to analyze this problem is the decision tree approach. Firstly, we have processed the dataset as it had many missing values. For that we used SVM algorithm to predict the missing values based on the known values which we used to train the SVM model. On completion of processed dataset, we built a decision tree based on two of the most popular node impurity measures in decision tree viz. entropy and gini index model. In this project the work load was shared between 3 project partners namely Bhavesh Shinde, Wen Chaun Chang and Pengda Ru. Each of us were responsible to carry out different problems in the project. Firstly, to extract the data we needed to apply SVM algorithm which was done by Wen Chaun Chang. She looked after everything related to building the SVM model and getting the processed required data. The decision tree algorithm was applied for the processed data by means of entropy and gini index. Pengda Ru was responsible to build the model for entropy and get the desired decision tree based on his model. I was responsible to build the gini index model and obtain a desired decision tree and results for the same. Finally, the project report writing was divided into 3 parts as: Bhavesh Shinde (Abstract, Introduction, Data Analysis, Discussion), Wen Chaun Chang (Methodology) and Pengda Ru (Results). The power point presentation was made by Wen Chaun Chang and Pengda Ru.

I was looking after the gini index model, so I will describe my work in brief.

**Methodology:**

The datasheet generated after SVM was a processed dataset which could be used to build the decision tree model directly. This dataset consisted of 1795 rows and 9 columns. The 9 columns indicate various parameters which determine the ratings of the chocolate bar. To build the gini index model I have considered 4 main parameters viz. Bean Type, Bean Origin, Company Location and Cocoa percentage. Except for the target and the cocoa percent, all the other variables are considered as categorical variables. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The Gini impurity can be computed by summing the probability Pi{\displaystyle p\_{i}} of an item with label i {\displaystyle i} being chosen times the probability {\displaystyle \sum \_{k\neq i}p\_{k}=1-p\_{i}} Pk = ∑ 1-Pi of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category. The equation to calculate gini index is given below:

I have split the dataset into training data and test data. The test size is kept as 0.01. The gini index model is built using python code. I have drafted a python code which will be found in my Github. A command to calculate the feature importance is added to the code. Also, the code to find the accuracy of the result can be seen. The maximum depth of the decision tree is chosen as 3 to avoid more complications in the results.

**Results:**

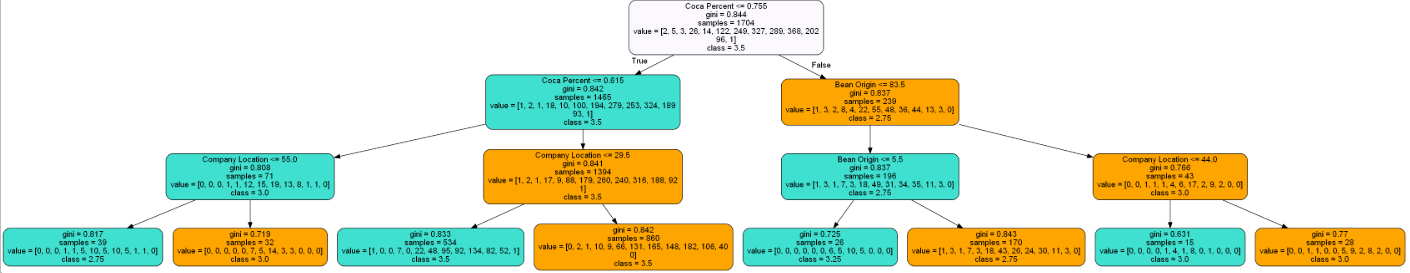
The decision tree generated on running the python code for gini index model is shown in figure 2. A similarity in the result of both metrics is seen as the most important parameter according to this metric also is cocoa percentage. But there are few dissimilarities in the results as well. The ideal cocoa percentage for the chocolate bar to have higher ratings is shown to be about 75%. If the cocoa percentage exceeds the ideal limit of 75% then bean origin in certain range can help to maintain the desired level of chocolate bar rating. The next main feature next to cocoa percentage is company location which is the same as entropy result. In the decision tree of this model, it is seen that when the branches of the decision tree are increased, bean type plays a small part in the outcome of the results unlike the entropy model where it did not appear in the decision tree at all. In this model, the accuracy rate for the entire output obtained is higher than that of entropy.

Figure 2: Decision Tree for Gini index

**Discussion:**

From the results obtained for the gini index model it can be seen that the important feature taken into consideration is the cocoa percentage. The desired percentage of cocoa in the chocolate bar is about 75%. Based on gini index, for a successful dark chocolate, factor which is considered important is the bean origin of less than or equal to 5.5. The accuracy of the results obtained from both the models varies a bit. Entropy giving out the accuracy to be 44.4% whereas gini index sows accuracy rate as 55.5%. The accuracy for the decision tree with depth=3 was found to be 55.55%. It was seen that if we increase the depth of the decision tree the accuracy kept on reducing. The lower the test size, the higher would be the accuracy. Until the depth of 3 layers, Bean type was not seen as a decisive parameter in the decision tree. But as we increase the depth of the decision tree above 3 we can find traces of consideration of bean type in the decision tree.

**Conclusion:**

The model for the chocolate bar ratings was successfully built using the gini index metric and a decision tree was created to show the results based on the gini model. Although the achieved accuracy could have been higher but the obtained accuracy rate of 55.55% is fair enough. It was found that gini model worked better than the entropy model.