# Analysis of chocolate bar ratings using Machine Learning

Wen Chaun Chang

Bhavesh Shinde

Pengda Ru

The George Washington University

Abstract

This paper is an analysis of a chocolate bar ratings datasheet using machine learning algorithms. The algorithm that has been discussed in this paper is the Decision Tree approach using gini index and entropy. The analysis is done based on 4 main parameters present in the datasheet. The datasheet had 50% of missing values in one of the parameter (Bean Type) which was eventually determined by using SVM. After analyzing, cocoa percent in the chocolate bar is the most important factor which decides the taste of the chocolate bar. The type of the bean used in the making of the chocolate bar does not have a significant role to play to decide its taste. The results also show that cocoa percent of more than 75% in a chocolate bar may ruin its taste. Company location and Bean origin also play a significant role in deciding the rating of the chocolate bar.

*Keywords*: Chocolate bar ratings, Decision tree, gini index, entropy, SVM.

Analysis of chocolate bar ratings using Machine Learning

# Introduction

In this paper, an analysis of a datasheet consisting of ratings of chocolate bars is carried out using machine learning techniques. The machine learning algorithm used in this paper is decision tree approach calculating the gini index and entropy separately and analyzing the results in both the cases. The datasheet consists of about 1800 different samples of chocolate bars with their company location, type of bean used, origin of the bean, percent of cocoa used in the bar and its rating based on a scale of 0 to 5. The datasheet did not show the type of bean used for about half the number of samples, so these values were missing in the datasheet used for the analysis. An approach to fill up the missing data is done using the Support Vector Machine (SVM) model. The available data for the type of bean used for the samples is considered as the training set data to build the SVM model. Based on this trained model, the type of bean used for the samples for which data was missing in the original datasheet was predicted with a fair enough accuracy rate for this SVM model. Thereafter filling the missing values in the datasheet, a processed datasheet was generated and used to build a decision tree model for the analysis of the chocolate bar ratings. In this paper, a decision tree algorithm is used to build a classifier model to analyze the ratings of chocolate bar based on different parameters such as company location, bean type, cocoa percent, bean origin etc. The decision tree approach can be carried out by two ways viz. calculating entropy and calculating gini index to determine the feature importance of various parameters. This paper calculates the feature importance of different parameters by both metrics of entropy and gini index. The results obtained by both the types of

calculations are almost similar with slight differences. The varied result is discussed further in the paper. The related decision tree is shown and discussed in the last part of this paper.

This paper talks about the variations to be made in different parameters related to the chocolate bar samples to classify their class and based on their class it can be determined as to which samples are better than the others.

# Method

### Data Collection and Data processing:

The datasheet used in this project is collected from Kaggle website from the following link (https://www.kaggle.com/rtatman/chocolate-bar-ratings). This datasheet consists of 1795 rows and 9 columns of data. The column data represents different attributes viz. company name, company location, specific bar origin, review date, cocoa percent, bean type, bean origin and rating. The row data represents 1795 samples of chocolate bars with different inputs for the above-mentioned attributes. This data consists of 887 missing values of the bean type column. These missing values were retrieved by applying SVM algorithm for the datasheet. The known values of the bean type column in the datasheet are trained for the SVM model. Based on this trained data, the missing data in the same column is predicted. The code for this SVM algorithm is shown in Appendix A. The data obtained upon applying the SVM module is inserted in the missing rows and we get a new datasheet eliminating the missing values and replacing them with the retrieved predicted values. This new datasheet is divided into training set and test set to train the decision tree algorithm to solve the main problem of the paper. The test size used is 0.01.

**SVM Methodology:**

SVM is introduced in this project to solve the problem of mass missing data. In the datasheet used for this project, there are as many as 887 missing values in the bean type column which counts to almost 50% of the data missing from that column. Therefore, Support Vector Machine is implemented to predict the missing values of the data. The data which is available in this column is used as the training set for the SVM algorithm and based on this trained model, the values for the missing columns are predicted. Support Vector Machine adopts Lagrange multiplier to simplify the calculation. Hence, the input data should be an indicator matrix. The framework used to solve this problem is scikit learn. The python code for this SVM algorithm is given on our Github. After running this code for the raw datasheet of chocolate bars ratings, a clean dataset with all the values filled up for the bean type column is obtained. The accuracy calculated for this SVM code comes out to be 67.4%. This newly obtained dataset is used to train and test the decision tree algorithm to solve our primary problem of the paper.

**Decision Tree Methodology:**

On running SVM for the missing values of the dataset, a new dataset with predicted bean type values for the missing data is generated. This dataset is split into two parts viz. training set and test set. The test size used is 0.01. The classification model used to solve the chocolate bar ratings classification is the Decision Tree model. To build the model four main parameters are considered viz. Cocoa percent, Company location, Bean origin and Bean type are used as variables and the target is the chocolate bars rating. Except for the target and the cocoa percent, all the other variables are considered as categorical variables. Therefore, the data processing involves another step which is transforming type into numerical array type. The decision tree can be solved using two metrics viz. entropy and gini index. Calculation of either entropy or gini

index gives the information regarding the most important parameter to be considered for building the decision tree. The framework used to build the decision tree is scikit learn and programming language used to code the model is python.

The algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely pure, the entropy is zero and if the sample is an equally divided, it has entropy of one. The following equation is about entropy calculation. According to the entropy evaluation, the information gain is based on the decrease in entropy after a dataset is split on an attribute. Therefore, constructing a decision tree is all about finding attribute that returns the highest information gain. The equation for entropy is given below:

$$\mathrm{E(s)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} -\mathrm{p_i} \log_2 \mathrm{p_i}$$

The Gini impurity, shown in the equation below, can be computed by summing the probability of an item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

$$\mathrm{G(s)} = 1 - \sum_{\mathrm{i=0}}^{\mathrm{n}} p_i^2$$

# Result:

The decision tree model built for the chocolate bars ratings is run for both the metrics i.e. entropy and gini index. The most popular node impurity measures used for the decision tree modeling are entropy and gini index. Two different python codes for entropy and gini index models are being run and two different results being obtained. The python code for this model is given on our Github. Both the results are discussed separately below.

**Entropy:**

The decision tree generated on running the code for entropy metric is shown as figure 1. The output of the decision tree shows that cocoa percent parameter has the maximum feature importance as compared to the other parameters. It shows that cocoa percentage of less than or equal to 90.5% is ideal for the chocolate bar to have the higher rating for the given sample. If the cocoa percentage exceeds the mentioned figure the rating of the chocolate bar keeps decreasing. Next to cocoa percentage, company location plays a key factor in determining the rating of the chocolate bar. The companies whose location is less than or equal to 32.5 (the set of such companies is mentioned in appendix D) can still manage to produce desirable chocolate bars considering that the cocoa percentage limit of 90.5% is exceeded. The parameter bean type does not contribute to any change in the results as the feature importance calculated for bean type parameter turned out to be 0.

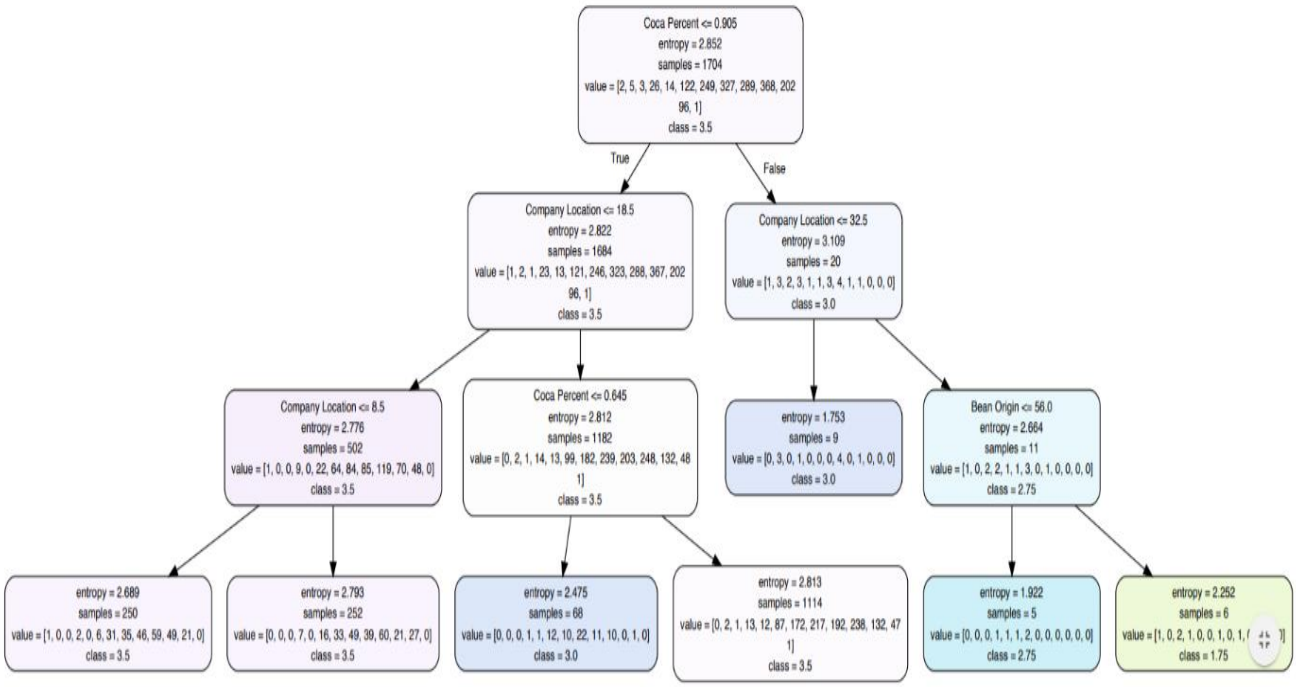


Figure 1: Decision Tree for Entropy

**Gini Index:**

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.

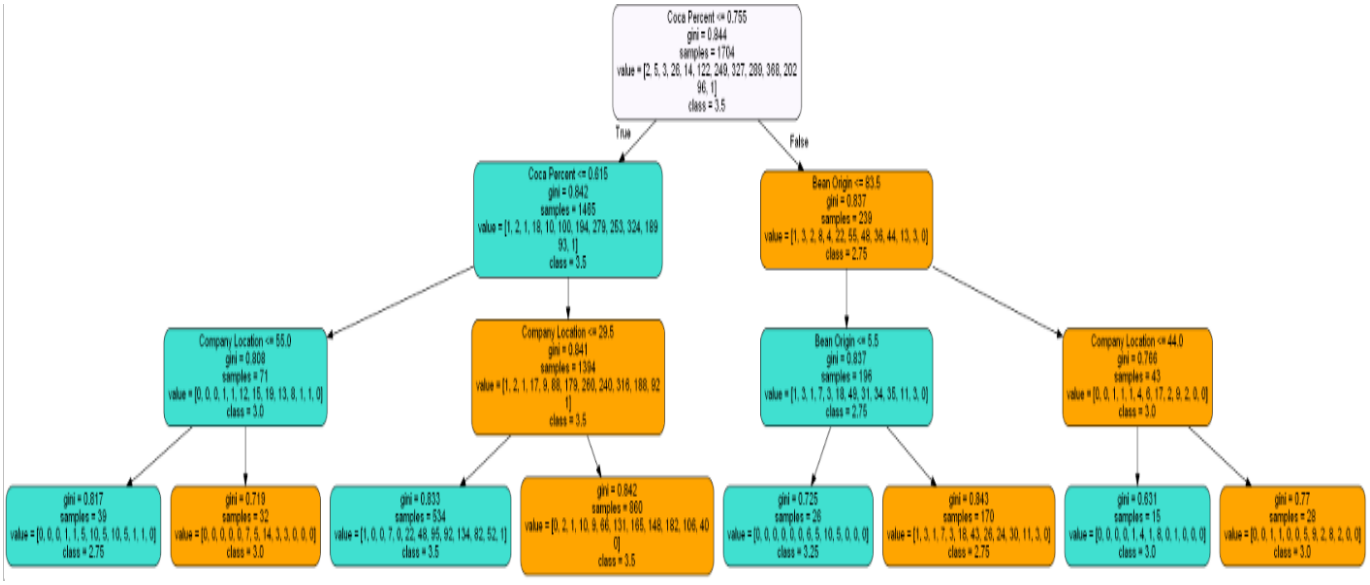


Figure 2: Decision Tree for Gini index

# Discussion:

A comparison between the two-metrics entropy and gini index is seen both showing almost similar results with some exceptions. The feature importance is calculated in both the metrics and it is seen the cocoa percentage holds the highest feature importance. The results for entropy model shows wider range of results for the class ratings of the chocolate bar whereas in gini index model the results do not vary much. But in both the models it is common that the class

rating of chocolate bar decreases on exceeding the higher limit of cocoa percentage. From this result, some information for successful making of dark chocolate can be extracted. For dark chocolate to maintain its taste and the rating of the bar both the models show the different paths for its success. Based on entropy, to make a dark chocolate with higher class ratings, the other factor considered important is company location within the range less than or equal to 32.5 (Refer Appendix A). 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 (Refer Appendix A). 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%.

Appendix A:

(1) Bean Origin <=5.5:

['Africa, Carribean, C. Am.' 'Australia' 'Belize' 'Bolivia' 'Brazil'

'Burma']

(2) Company location <=32.5

['Amsterdam' 'Argentina' 'Australia' 'Austria' 'Belgium' 'Bolivia'

'Brazil' 'Canada' 'Chile' 'Colombia' 'Costa Rica' 'Czech Republic'
'Denmark' 'Domincan Republic' 'Ecuador' 'Eucador' 'Fiji' 'Finland'
'France' 'Germany' 'Ghana' 'Grenada' 'Guatemala' 'Honduras' 'Hungary'
'Iceland' 'India' 'Ireland' 'Israel' 'Italy' 'Japan' 'Lithuania'
'Madagascar']

**References:**

(1) For python code:

https://github.com/amir-jafari/Machine-Learning/tree/master/Python-Algorithms/Classification/1-Decision_Tree

http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

(2) For data collection:

https://www.kaggle.com/rtatman/chocolate-bar-ratings