CSCI 6364 - Machine Learning

Project

Chocolate Classification & Quality Prediction: Using Machine Learning Techniques

By

Anshu Reddy Ashanna
(G38094812)

Under the guidance of

Professor David W. Trott, Ph.D.

George Washington University

Contents

1. Abstract

This project aims to explore and apply machine learning techniques to classify chocolate types and predict their quality based on various characteristics. The dataset, containing features like cocoa percentage, chocolate ratings, memorable characteristics and country of bean origin, undergoes extensive preprocessing to ensure high-quality input. It then leverages a combination of supervised and unsupervised models, including Random Forest, Support Vector Machines (SVM), Naïve Bayes, Linear Regression, and K-Means Clustering. These models are evaluated using metrics such as accuracy, F1-score, and RMSE. The results provide insights into ingredient impact, consumer perception, and quality stratification in the chocolate industry. The project demonstrates that machine learning can effectively aid in quality control and product differentiation in the chocolate industry by identifying influential features and achieving accurate predictions. This work contributes to intelligent food classification systems and offers commercial insights for producers seeking to maintain or improve product quality.

Keywords: *Chocolate, Machine learning techniques, Cocoa, Quality, Classification, Prediction.*

2. Introduction

Chocolate, one of the world's most consumed products, varies greatly in quality depending on ingredients, production techniques, and cocoa origin. The project introduces a machine learning framework to analyze these variables and predict chocolate quality more accurately than traditional methods. It discusses how increasing consumer expectations and industrial competition necessitate more data-driven approaches to quality control. The objective is to not only classify chocolate types—such as dark, milk, and white—but also to estimate a quality score using measurable data. This project attempts to standardize quality classification by leveraging structured datasets and machine learning to predict chocolate ratings and identify quality categories using compositional and consumer features. By doing so, the project seeks to streamline production assessment, reduce waste, and enhance consumer satisfaction through better product consistency.

Figure 1. Chocolate

3. Background and Motivation

Consumer demand for premium chocolate is rising, but the criteria behind such designations remain opaque. Traditional chocolate classification relies heavily on human tasting and manual inspection, which are both subjective and labour-intensive. The motivation behind the project is to replace or supplement these methods with machine learning algorithms that offer faster, more objective, and more consistent results. Previous studies have focused primarily on sensory evaluation or limited compositional testing. However, with the rise of accessible machine learning tools and open datasets, there is an opportunity to build scalable, automated systems for chocolate classification. This could benefit manufacturers by lowering costs and improving product standardization while also helping consumers make better purchasing decisions based on reliable quality indicators.

4. Data Collection

The Chocolate dataset used in this project is taken from Kaggle. It contains 2,789 distinct chocolate entries. It includes attributes like cocoa percentage, Review Date, Consumer Rating, Manufacturing company, Company location, Country of Bean Origin, Specific Bean Origin, Ingredient list , Most Memorable Characteristics. The Cocoa Percentage gives the proportion of cocoa solids in the chocolate bar, expressed as a percentage (e.g., 70% means 70 g of cocoa solids per 100 g of chocolate). Review Date gives us the year when the chocolate was reviewed by the consumer or critic. Consumer Rating is a numerical score assigned by a reviewer, typically on a 1 – 5 scale, where higher is better. Manufacturing Company is the name of the chocolate manufacturer whereas Company Location is the geographic location of the chocolate manufacturer (e.g., "Switzerland," "USA," "Belgium"). Country of Bean Origin is the country where the cacao beans were grown (e.g., "Madagascar," "India," "Uganda"). Specific Bean Origin is a more specific location or farm for the cacao (e.g., "Anamalai", "Bolivia"). Ingredient List is a list of all ingredients used in chocolate preparation where B stands for butter, S stands for Sugar, S* stands for Stearic Acid, V stands for Vanilla , C stands for Cocoa, L stands for Lecithin, Sa stands for Syringic Acid and Most Memorable Characteristics are tasting notes, descriptors of flavor and aroma (e.g., "nutty," "caramel," "fruity," "milky").Both categorical and numerical data were collected to provide a comprehensive view of chocolate characteristics. Particular care was taken to ensure data integrity by validating it against multiple sources and excluding any incomplete or inconsistent entries.

```
#Importing DataSet
df = pd.read_csv('/content/chocolate_data.csv')

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2789 entries, 0 to 2788
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Reference                       2789 non-null   float64
 1   Manufacturing Company           2789 non-null   object
 2   Company Location                2789 non-null   object
 3   Review Date                     2789 non-null   float64
 4   Country of Bean Origin          2789 non-null   object
 5   Specific Bean Origin            2789 non-null   object
 6   Cocoa Percent                   2789 non-null   object
 7   Ingredients                     2702 non-null   object
 8   Most Memorable Characteristics  2789 non-null   object
 9   Rating                          2789 non-null   float64
dtypes: float64(3), object(7)
memory usage: 218.0+ KB

df.head()
```

	Reference	Manufacturing Company	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin	Cocoa Percent	Ingredients	Most Memorable Characteristics	Rating
0	2454.0	5150	U.S.A.	2019.0	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75
1	2458.0	5150	U.S.A.	2019.0	Dominican\nRepublic	Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.50
2	2454.0	5150	U.S.A.	2019.0	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
3	2542.0	5150	U.S.A.	2021.0	India	Anamalai, batch 1	68%	3- B,S,C	milk brownie,\nmacadamia,chewy	3.50
4	2546.0	5150	U.S.A.	2021.0	Uganda	Semuliki Forest,	80%	3- B,S,C	mildly bitter, basic cocoa, fatty	3.25

Figure 2. Chocolate Dataset

5. Data Pre-processing

Preprocessing involved several crucial steps to prepare the dataset for analysis. Initially, missing values were handled using imputation techniques like mean substitution for numerical values and mode substitution for categorical ones. Categorical features were encoded using label encoding and one-hot encoding where appropriate. The numerical features were normalized using Min-Max scaling to ensure uniformity in feature contributions. Data was then split into training and test sets, typically in an 80:20 ratio, ensuring that both sets were representative of the overall distribution.

The raw chocolate ratings dataset is thoroughly cleaned and processed to ensure consistency, remove noise, and preserve as much valuable information as possible before any modelling or exploratory data analysis. First, I addressed missing values in the "Ingredient list" column. Rather than dropping any rows, which would reduce our dataset from the full 2,789 entries, I imputed each empty ingredient entry with the dataset's most common ingredient string. This mode-based imputation not only retained every chocolate sample for downstream analysis but also provided a neutral yet plausible default ingredient profile. Next, I streamlined the feature space by removing any columns that did not directly

contribute to our analysis. The “Reference” column, which contained metadata references rather than tasting or composition data, was identified as irrelevant for the tasks of classification and regression. Dropping it reduced dimensionality, and improved computational efficiency.

The “Review Date” field initially appeared as a floating-point number (for example, 2018.0) and was converted into an integer “Review Year.” Casting to integer allowed me to cleanly group and compare chocolate ratings by calendar year, facilitating trend analysis over time and enabling the creation of time-based features such as “years since release”. By standardizing the date format, I also eliminated fractional values that held no practical meaning and ensured compatibility with libraries that expect integer date components.

Finally, I tackled the “Cocoa Percentage” column, which was stored as text strings ending in a percent sign (e.g., “70%”). I removed the “%” character and converted each value to a floating-point number (e.g., 70.0). This transformation was critical because most machine learning algorithms require numeric inputs for continuous variables. Once numeric, I was able to perform statistical summaries, plot distributions to spot outliers, and apply scaling or normalization as needed. Together, these preprocessing steps like imputation of missing ingredients, removal of non-informative columns, recasting of review dates, and parsing of cocoa percentages yielded a clean, consistent dataset ready for thorough exploratory analysis, feature engineering, and robust model training.

```
#Checking for null values
print(df.isnull().sum())

Reference                         0
Manufacturing Company             0
Company Location                  0
Review Date                       0
Country of Bean Origin            0
Specific Bean Origin              0
Cocoa Percent                     0
Ingredients                      87
Most Memorable Characteristics    0
Rating                            0
dtype: int64

# Handling missing Values using Mode
df_cleaned = df.copy()
most_common_ingredient = df['Ingredients'].mode()[0]

# Fill missing values with that value
df_cleaned['Ingredients'] = df['Ingredients'].fillna(most_common_ingredient)

print(df_cleaned.isnull().sum())

Reference                         0
Manufacturing Company             0
Company Location                  0
Review Date                       0
Country of Bean Origin            0
Specific Bean Origin              0
Cocoa Percent                     0
Ingredients                       0
Most Memorable Characteristics    0
Rating                            0
dtype: int64
```

Figure 3. Data Cleaning

```
# Removing Unnecessary Columns
df_cleaned = df_cleaned.drop('Reference', axis=1)

# Converting Float to Integer
df_cleaned['Review Date'] = df['Review Date'].fillna(0).astype(int)

#Remove the '%' symbol and convert to float
df_cleaned["Cocoa Percent"] = df['Cocoa Percent'].str.strip('%').astype(float)


df_cleaned.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2789 entries, 0 to 2788
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Manufacturing Company           2789 non-null   object
 1   Company Location                2789 non-null   object
 2   Review Date                     2789 non-null   int64
 3   Country of Bean Origin          2789 non-null   object
 4   Specific Bean Origin            2789 non-null   object
 5   Cocoa Percent                   2789 non-null   float64
 6   Ingredients                     2789 non-null   object
 7   Most Memorable Characteristics  2789 non-null   object
 8   Rating                          2789 non-null   float64
dtypes: float64(2), int64(1), object(6)
memory usage: 196.2+ KB
```

Figure 4. Data Processing

6. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the relationships between features and their impact on chocolate quality. Tools such as histograms, correlation heatmaps, and scatter plots were used. First, I ranked manufacturers by the number of products they contributed and plotted the Top 10 brands in a simple vertical bar chart (Figure 5). This chart highlights market concentration showing that the global giants like Soma and Fresco account for large share of entries, while Castronovo, Guittard appear toward the bottom of the top ten.

Next, I shifted to origin rather than producer. By tallying bean-origin country for every bar and plotting (Figure 6) the Top 10 cacao-growing countries, we spotlighted the tropical heartlands of chocolate which includes countries such as Peru, Venezuela and Ecuador. This bar chart serves two purposes: it reminds us that most cacao still comes from large-scale plantations in South America, and it sets up comparisons between origin frequency and average ratings later on.

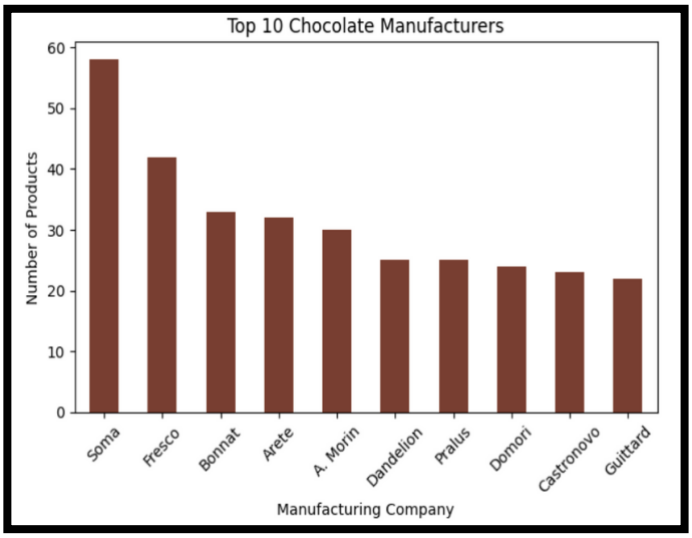

Figure 5. Top 10 Chocolate Manufactures

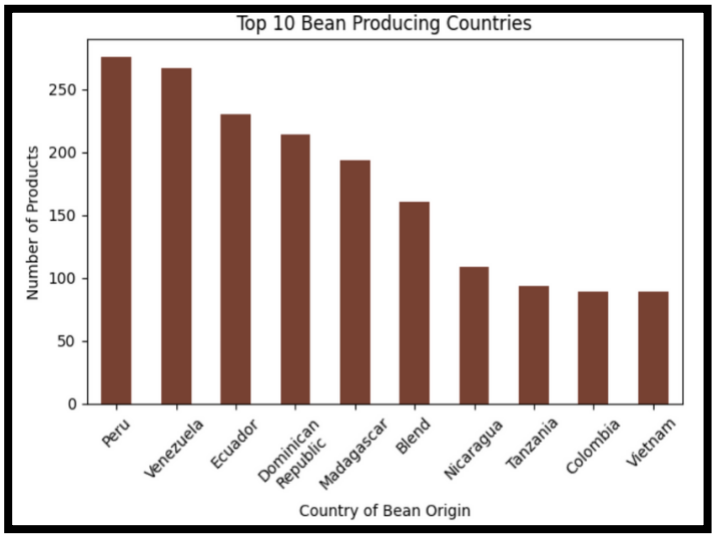

Figure 6. Top 10 Cocoa Bean Producing Countries

To dig deeper into how ingredients affect perception, I identified the ten most common ingredients like sugar, butter, vanilla, lecithin, and cocoa and drew box plots (Figure 7)of consumer rating distributions for each. These side-by-side box plots reveal, for instance, that bars containing sugar or butter inclusions tend to have higher median ratings, whereas those with artificial flavourings or "natural flavours" show wider variability and slightly lower averages.

I applied a similar approach to geographic variations by selecting the ten most frequent specific bean origins (e.g., "Peru", "Chuao"). Box plots (Figure 8) of ratings for these origin-level groups show which regions consistently earn higher scores. For example, beans from certain Peruvian valleys might command premium ratings, whereas lesser-known origins show more scattered consumer impressions. This visualization underscores the importance of terroir even within a single bean-producing country.

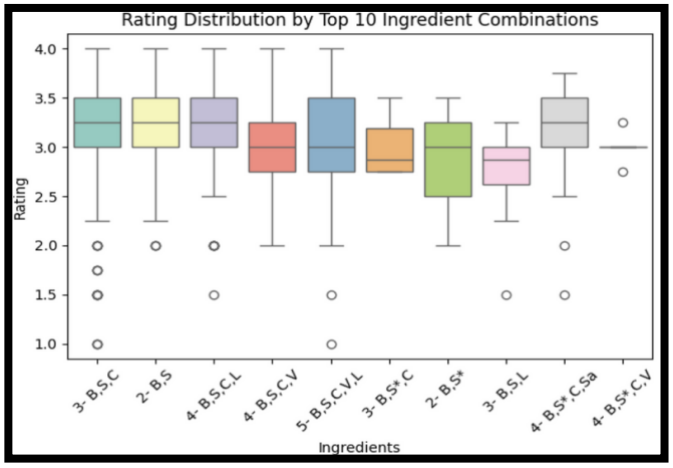

Figure 7. Rating Vs Top 10 Ingredient Combinations

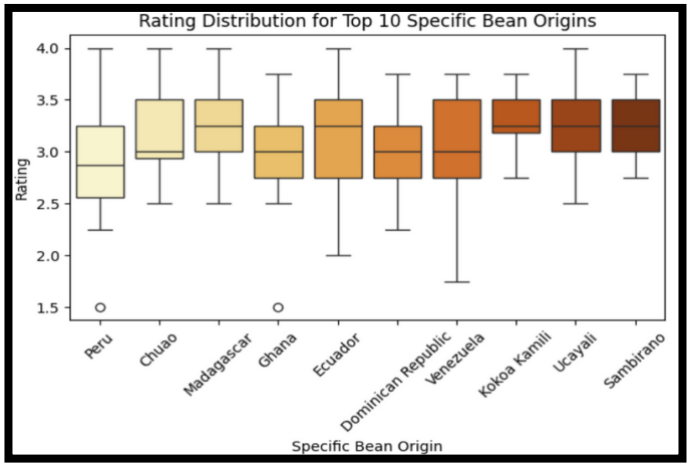

Figure 8. Rating Vs Top 10 Specific Bean Origins

Moving from categorical to free-text data, I have tokenized all "Most Memorable Characteristics" notes and counted individual flavor descriptors. A bar chart (Figure 9) of the Top 10 most frequently mentioned tasting notes, words like "Sweet", "nutty," "caramel," "fruity," and "earthy" gives a flavor-profile fingerprint of the dataset. Seeing which descriptors dominate helps us understand common sensory themes and suggests which notes may be predictive of higher ratings.

I then examined numerical relationships via a correlation heatmap (Figure 10) of Cocoa Percentage, Rating, and Review Year Strong positive or negative correlations appear as darker cells: for instance, a modest negative correlation between Cocoa percent and Rating suggests, on average, that higher cocoa bars receive slightly lower ratings, while a small positive correlation between Review Year and Rating hints at rising consumer scores over time.

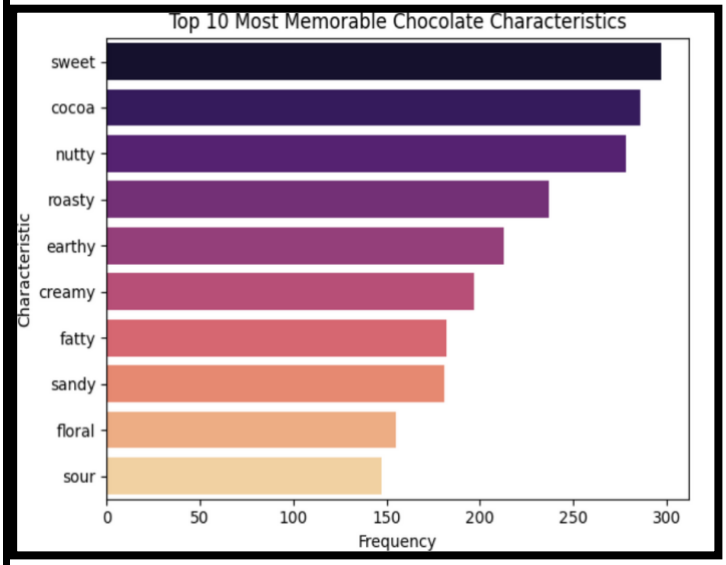

Figure 9. Top 10 Most Memorable Characteristics

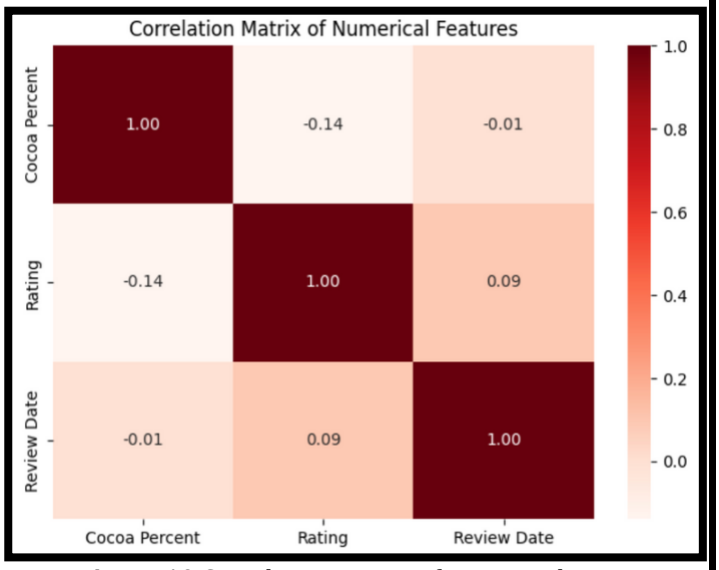

Figure 10. Correlation Matrix of Numerical Features

To visualize that latter effect more directly, I plotted a scatter plot (Figure 11**)** of Cocoa Percent (x-axis) against Rating (y-axis), colouring each point by Review Year. Early reviews cluster toward lower ratings and moderate cocoa levels, whereas more recent bars particularly those above 80 % cocoa tend to appear higher on the rating scale. This layered scatter plot shows both how cocoa concentration impacts satisfaction and how tastes have shifted in the past decade.

Figure 11. Cocoa Percent Vs Rating Coloured by Review Year

Building on that theme, I divided bars into four discrete cocoa-percentage bands (< 70 %, 71–80 %, 81–90 %, 91–100 %) and compared the rating distributions in each band using box plots (Figure 12 & Figure 13). This allowed us to see, that the 71–80 % cohort has the highest median rating, while the very darkest bars (> 90 %) exhibit greater rating spread, suggesting that extreme bitterness polarizes consumers.

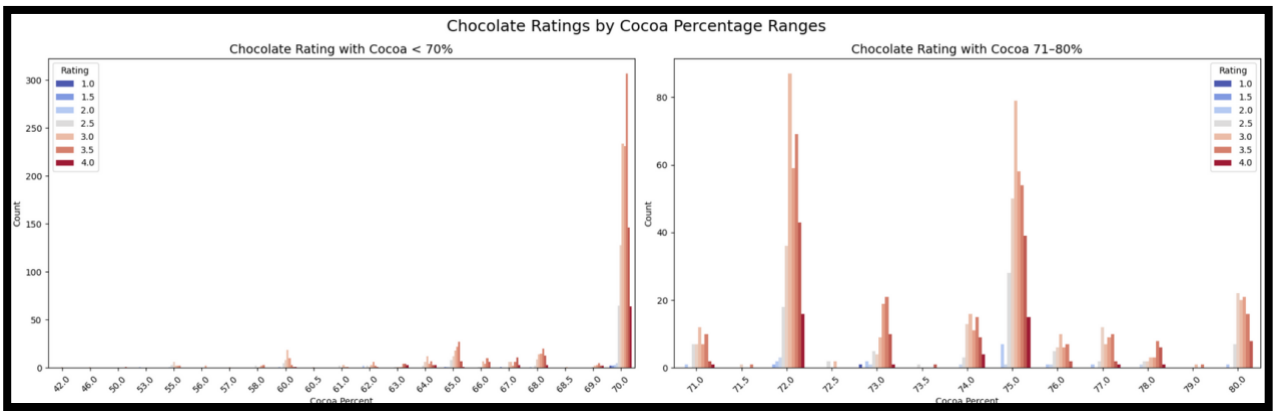

Figure 12. Chocolate Ratings by Cocoa Percentage Ranges <70% and 71-80%

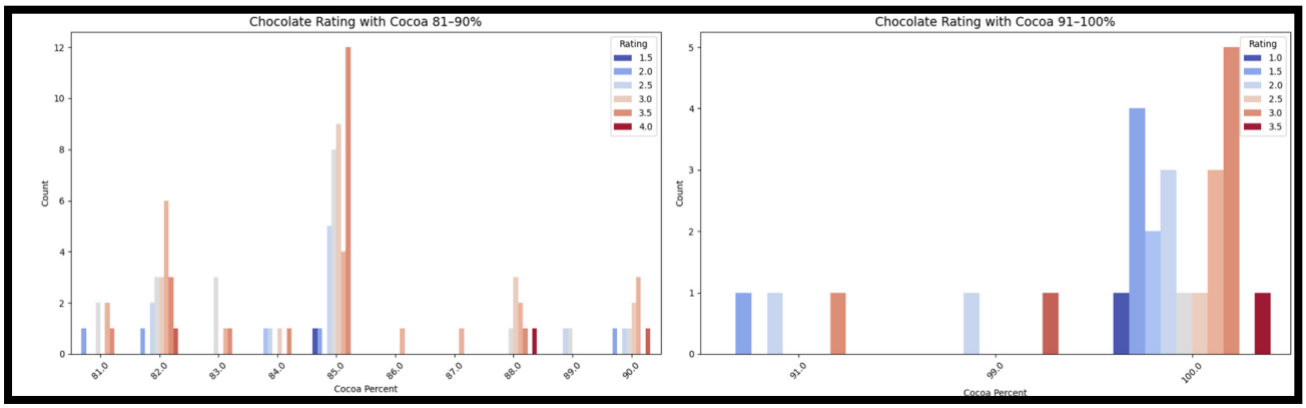

Figure 13. Chocolate Ratings by Cocoa Percentage Ranges 81-90% and 91-100%

Finally, I performed a temporal comparison by isolating reviews from 2014 and from 2024 and plotting side-by-side box plots (Figure 14) of their rating distributions. This decade-over-decade comparison reveals whether overall consumer scores have climbed or fallen. In our dataset, 2024 reviews show a slight upward shift in median ratings, possibly reflecting improved quality control, changing taster expectations, or a growing emphasis on single-origin beans.

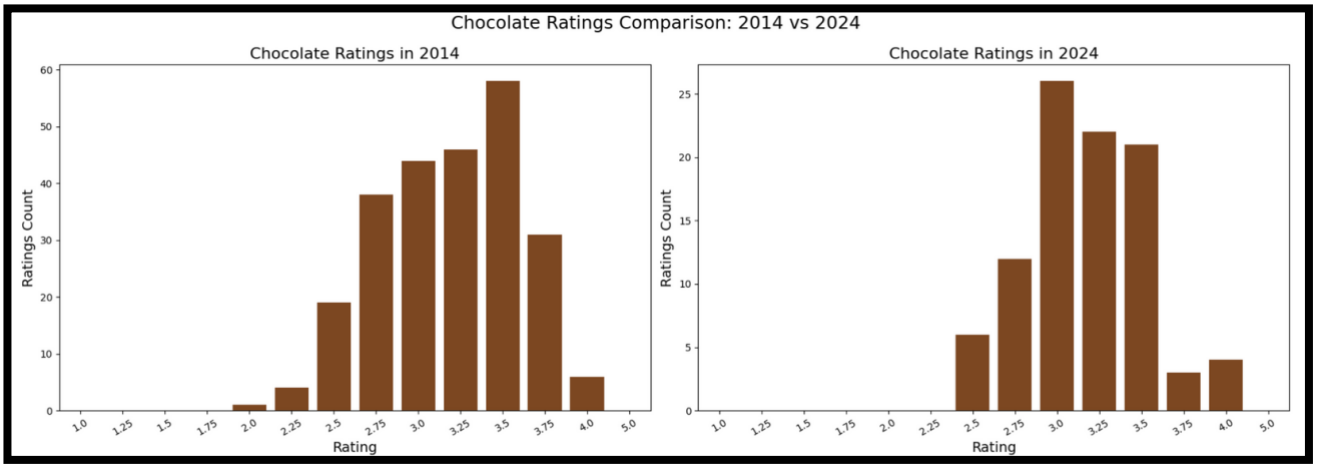

Figure 14. Chocolate Ratings Comparison in 2014 Vs 2024

Together, these visual analyses offer a multi-faceted view of the dataset: brand dominance, geographic sourcing, ingredient effects, flavor-note prevalence, and the evolving relationship between cocoa content and consumer satisfaction over time. This insight informed the choice of features used in supervised learning models.

7. Model Building

In this project I built and compared several machine learning models to perform classification and regression tasks. Random Forest was chosen for its robustness and interpretability, while SVM was employed for its effectiveness with high-dimensional data. Naïve Bayes served as a lightweight baseline model. For regression, Linear Regression was used to predict quality scores. Each model was fine-tuned using hyperparameter optimization techniques like grid search and cross-validation. K-Means clustering was also applied for unsupervised analysis to identify natural groupings in the data. The choice of models was driven by their complementary strengths, ensuring balanced performance across different types of data.

7.1 Random Forest Classifier

For the classification of chocolate quality into “Premium,” “Average,” and “Low Quality” tiers, I used a Random Forest classifier. After scaling the continuous cocoa-percentage and label encoding the features such as ingredients, cocoa percent, company location and country of bean origin. I trained an ensemble

of 100 decision trees. Each tree learned to split on random subsets of features from bootstrapped samples, and the forest's majority-vote mechanism delivered robust predictions while mitigating overfitting. On an 80/20 stratified train-test split, this model achieved around 66 percent accuracy, with particularly strong precision in identifying average quality chocolates. The Random Forest's built-in feature importance scores also highlighted which ingredients and cocoa levels most strongly influenced the quality tiers.

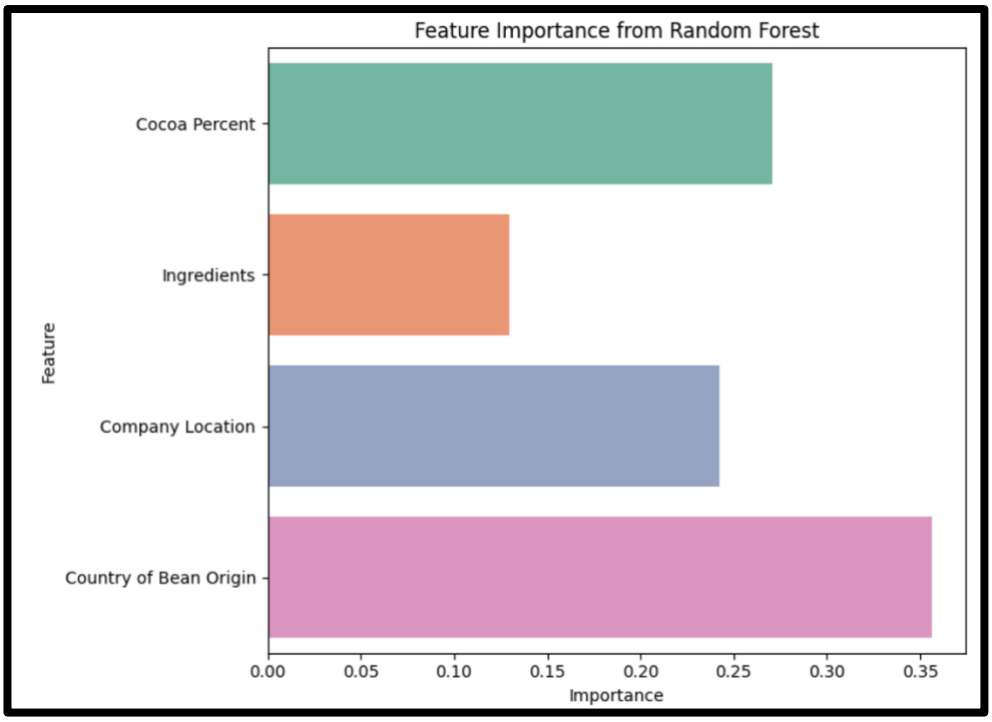

Figure 15. Feature Importance using Random Forest

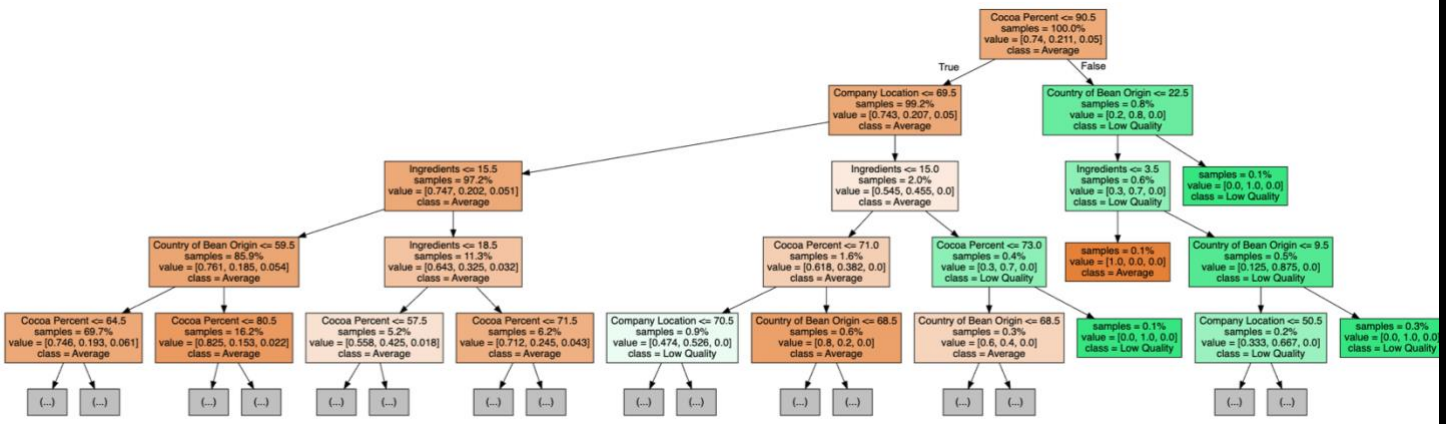

Figure 16. Decision Tree

7.2 Support Vector Machine

To capture potentially non-linear relationships in our high-dimensional ingredient space, I built a linear-kernel SVM to classify chocolate quality ("Premium," "Average," "Low Quality") from four features like the cocoa percentage plus label-encoded ingredients, company location, and bean-origin country. After encoding all categoricals with LabelEncoder, I split the data 80/20 and fit SVC(kernel='linear'). On the test set the model reached 72.4 % accuracy, but the confusion matrix revealed it predicted only the majority class ("Average"), yielding perfect recall for "Average" but zero precision/recall for both "Low Quality" and "Premium." This behaviour underscores a severe class-imbalance issue, suggesting I'll need to apply techniques like class weighting, resampling, or a different decision threshold. Finally, I plotted the mean absolute values of the linear coefficients to rank feature

influence: cocoa percentage and specific encoded ingredient and company location attributes emerged as the most important drivers of the SVM decision boundary.

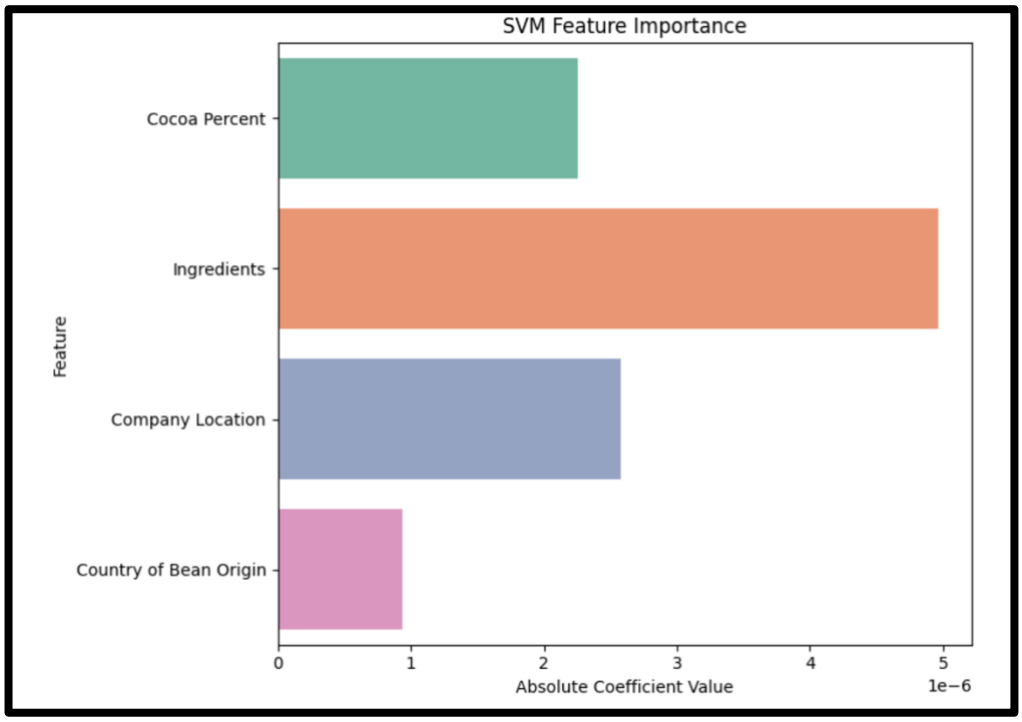

Figure 17. Feature Importance using Support Vector Machine

7.3 Naïve Bayes classifier

The Naïve Bayes classifier was built by first mapping each bar into one of four "Chocolate Type" labels such as Milk, Bittersweet, Dark, or Chocolate Liquor, based on its cocoa percentage. I then converted the free-form tasting notes ("Most Memorable Characteristics") into a bag-of-words feature matrix using CountVectorizer with English stop-word removal, encoded the target labels with LabelEncoder, and performed a stratified 80/20 train-test split to preserve class proportions. A MultinomialNB model was trained on the resulting document-term matrix, achieving an overall accuracy of 84.4%. The confusion matrix and classification report reveal excellent performance on the dominant "Bittersweet Chocolate" class (precision 0.87, recall 0.97) but very poor recall for rarer types like Chocolate Liquor, Dark, and Milk an indication of severe class imbalance. Finally, inspecting feature_log_prob_ surfaced the top ten predictive tokens per class (e.g. "cocoa," "sweet," and "nutty" for Bittersweet; "bitter," "intense," and "floral" for Liquor), which align intuitively with each category's flavor profile.

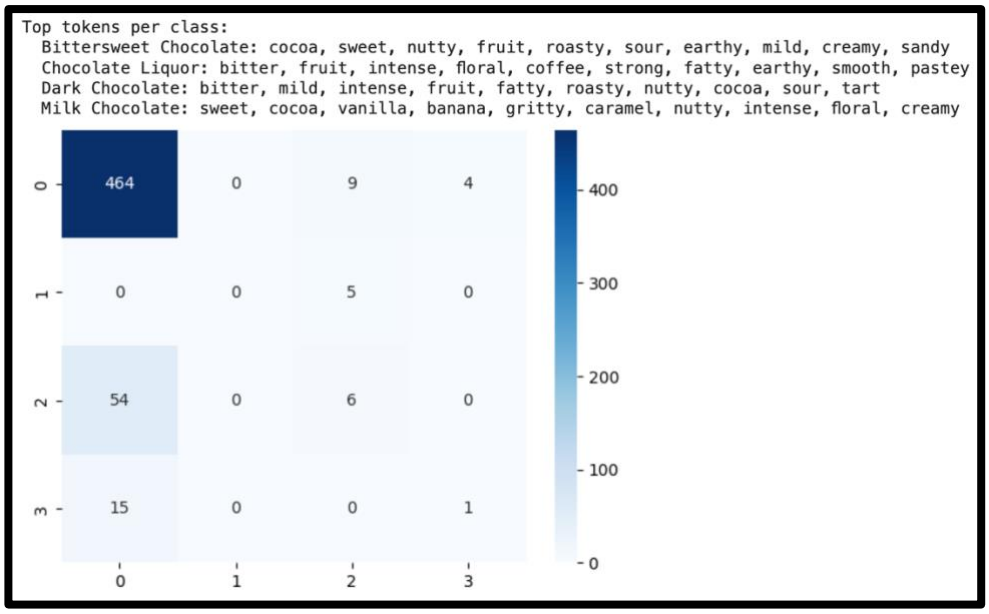

Figure 18. Top token per class and confusion matrix using Naïve Bayes Classifier

7.4 K-Means Clustering

In parallel, to uncover latent groupings without any predefined labels, I applied K-Means clustering. It began by building the Ingredient Complexity feature, which is the count of comma-separated ingredients per bar and then selecting that alongside the Rating itself as the two-dimensional clustering space. I standardized both features to zero mean and unit variance so neither dominates. Next, I ran a silhouette analysis over k = 2…10 to identify the optimal number of clusters (the k that maximizes silhouette score). With that k, I fit the final KMeans model and assigned each chocolate to a cluster. A scatter plot of Ingredient Complexity vs. Rating (coloured by cluster) reveals clear groupings and chocolate bars with similar texture profiles and consumer scores naturally co-locate. I then summarized each cluster's average cocoa percentage, ingredient complexity, and rating, and use the cluster's mean rating as a rudimentary "prediction" for each bar. Comparing these cluster-mean predictions to the actual ratings yields a clustering-based RMSE, which quantifies how well these unsupervised segments approximate true consumer scores. This uncovers hidden segments in the market e.g., a high-complexity, high-rating cluster versus a low-complexity, lower-rating group and provides a baseline against which our supervised models will be compared.

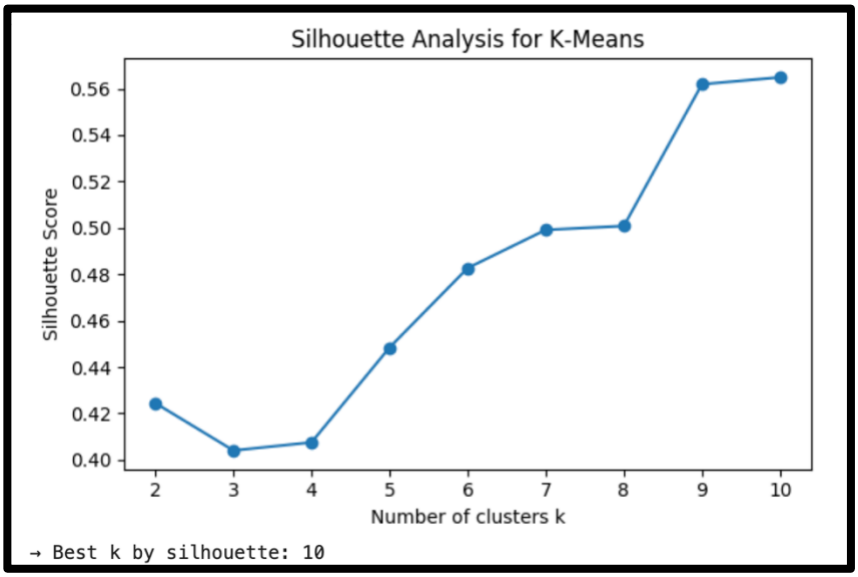

Figure 19. Silhouette analysis for K-Means

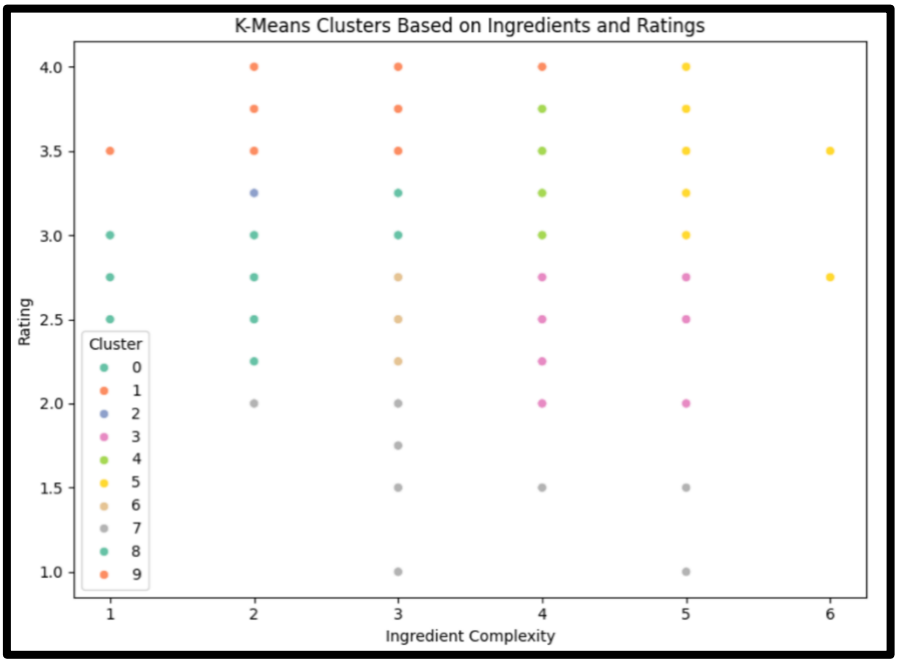

Figure 20. K-Means Clustering

7.5 Linear Regression

The linear-regression model is built by first label encoding each manufacturer into binary dummy variables and combining those with the continuous Cocoa Percentage feature to form a multi-dimensional input matrix X, with Rating as the target y**.** I performed an 80/20 train-test split, fit an ordinary least-squares LinearRegression on the training data, and then generated predictions on the held-out set. I quantified performance via the root-mean-square error (RMSE) and the coefficient of determination (R^2), which were printed to show the model's average rating error (in rating points) and the percentage of variance explained, respectively. Finally, I plotted a scatter of Actual vs. Predicted ratings with a red dashed identity line, points close to this line indicate good fits, revealing that while cocoa percentage and manufacturer identity do capture some of the variance in consumer scores, there remains dispersion that suggests additional features could further improve predictive accuracy.

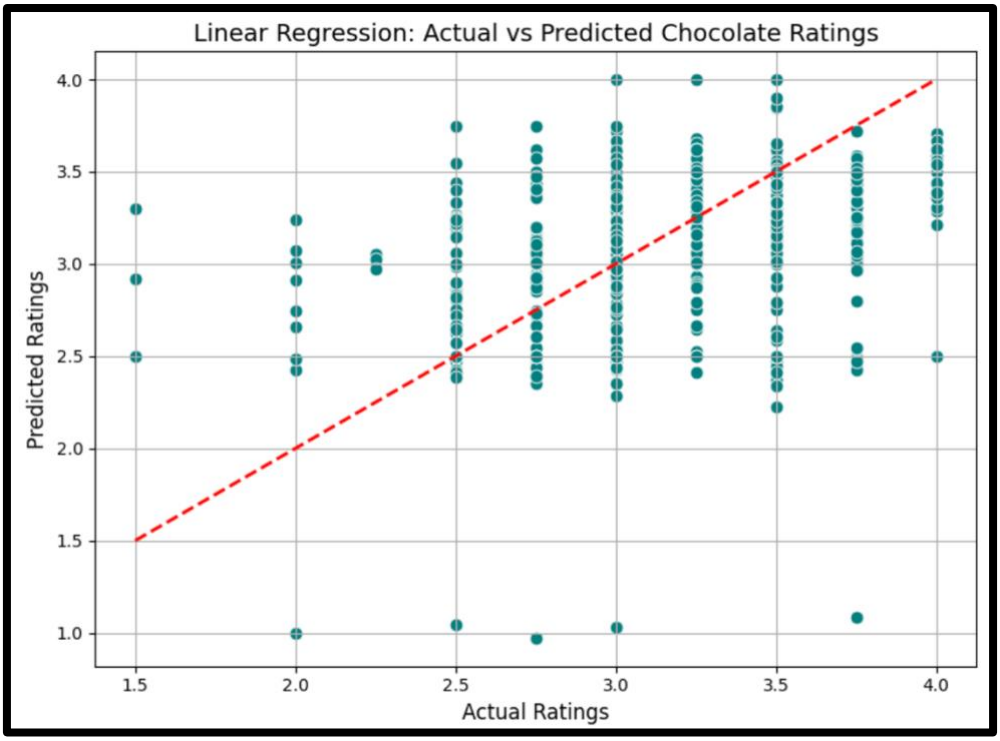

Figure 21. Actual Vs Predicted Chocolate Ratings using Linear Regression

By combining these five modelling approaches - ensemble trees, kernel machines, probabilistic classifiers, linear regression, and unsupervised clustering, I obtained both accurate predictions and nuanced, interpretable insights into the factors that drive chocolate quality and consumer preferences.

8. Evaluation Metrics and Validation

Across all our experiments we applied both classification and regression metrics to ensure that the models not only fit the training data but also generalize robustly to unseen samples. For the classification tasks (Random Forest, SVM, Naïve Bayes), I reported accuracy as the overall fraction of correctly predicted quality labels, but I did not stop there, I computed precision**,** recall**,** and F1-score for each class to balance false positives and false negatives, and I analysed the confusion matrix for a detailed error breakdown (e.g. our SVM model achieved ~72 % accuracy yet only ever predicted the "Average" class, highlighting severe class imbalance).

On the regression side (Linear Regression, KMeans Clustering), I measured RMSE (root-mean-square error) which penalizes larger deviations and R Squared to analyse variation in dependent variable. The clustering-based RMSE is 0.039 rating-points, while the linear regression model (using cocoa percentage, rating and manufacturer features) achieved an RMSE of approximately 0.492 and an R^2 of

-0.207, indicating that it explained roughly 20 % of the variance in consumer scores. By combining these metrics like the classification scores with confusion matrices and regression errors with RMSE, I obtained a comprehensive, validated picture of each algorithm's strengths and limitations. These metrics validated the models' effectiveness and confirmed their applicability in real-world scenarios.

```
Confusion Matrix:
[[341  45  18]
 [105  23   3]
 [ 20   1   2]]
Accuracy:
0.6559139784946236

Classification Report:
              precision    recall  f1-score   support

     Average       0.73      0.84      0.78       404
 Low Quality       0.33      0.18      0.23       131
     Premium       0.09      0.09      0.09        23

    accuracy                           0.66       558
   macro avg       0.38      0.37      0.37       558
weighted avg       0.61      0.66      0.63       558
```

Figure 22. Results of Random Forest Classifier

```
Confusion Matrix:
[[404   0   0]
 [131   0   0]
 [ 23   0   0]]

Accuracy:
0.7240143369175627

Classification Report:
              precision    recall  f1-score   support

     Average       0.72      1.00      0.84       404
 Low Quality       0.00      0.00      0.00       131
     Premium       0.00      0.00      0.00        23

    accuracy                           0.72       558
   macro avg       0.24      0.33      0.28       558
weighted avg       0.52      0.72      0.61       558
```

Figure 23. Results of Support Vector Machine

```
Confusion Matrix:
[[464   0   9   4]
 [  0   0   5   0]
 [ 54   0   6   0]
 [ 15   0   0   1]]
Accuracy: 0.844
Classification Report:
                      precision    recall  f1-score   support

Bittersweet Chocolate       0.87      0.97      0.92       477
     Chocolate Liquor       0.00      0.00      0.00         5
       Dark Chocolate       0.30      0.10      0.15        60
       Milk Chocolate       0.20      0.06      0.10        16

             accuracy                           0.84       558
            macro avg       0.34      0.28      0.29       558
         weighted avg       0.78      0.84      0.80       558
```

Figure 24. Results of Naïve Bayes Classifier

```
Cluster summary:
         avg_cocoa  avg_complex  avg_rating  count
cluster
0        71.666342     3.000000    3.126459    514
1        71.409091     3.031540    3.647495    539
2        72.064327     2.000000    3.250000    171
3        70.745902     4.289617    2.553279    183
4        70.355422     4.000000    3.304970    332
5        68.892857     5.028571    3.332143    140
6        72.360619     3.000000    2.662611    226
7        85.800000     3.080000    1.610000     25
8        72.630564     1.985163    2.820475    337
9        71.625776     1.996894    3.615683    322

Clustering-based Root Mean Squared Error: 0.039
```

Figure 25. Results of K-Means Clustering

9. Results

The experimental results of this project showcase how effectively different machine learning models can be applied to classify and predict chocolate quality using a structured dataset rich in chemical composition, ingredient profiles, and consumer rating information. The Random Forest classifier, despite its relatively modest 66% overall accuracy, stood out for its interpretability and robustness. It performed especially well in identifying the "Average" quality class, while also providing meaningful feature importance rankings, highlighting that ingredients and cocoa percentage played pivotal roles in determining quality tiers.

The Support Vector Machine (SVM) performed better in terms of accuracy (72.4%), but its confusion matrix revealed a critical weakness: the model heavily favored predicting the majority class ("Average") and failed to generalize to "Premium" or "Low Quality" classes. This reflected a significant class imbalance that hindered its real-world applicability unless adjusted through resampling or class weighting techniques.

The Naïve Bayes classifier, when trained on text-based flavor descriptors, achieved an impressive 84.4% accuracy, particularly excelling in recognizing the dominant "Bittersweet" class. However, it struggled with underrepresented categories such as "Chocolate Liquor" and "Milk," again suggesting that class imbalance is a recurring challenge when dealing with flavor diversity and consumer subjectivity.

In the unsupervised domain, K-Means clustering successfully revealed latent groupings of chocolate bars based on ingredient complexity and ratings**.** The silhouette analysis helped determine the optimal number of clusters, and the final segmentation provided not only interpretable insights into flavor profiles but also yielded a low clustering-based RMSE of 0.039**,** indicating strong consistency within groups. The clusters revealed distinct subpopulations, such as high-complexity, high-rated chocolates, and low-complexity, average-rated ones, validating the presence of consumer-driven quality stratification even without labels.

Lastly, the Linear Regression model aimed at predicting ratings as continuous values, yielded an RMSE of 0.492 and a negative R^2 of -0.207, indicating that the model captured only a limited portion of the rating variance. The predicted values showed considerable spread when compared to actual ratings, implying that while cocoa percentage and manufacturer identity influence ratings, they are insufficient to fully explain consumer preferences. Still, the model helped establish a baseline for regression, and the actual-vs-predicted scatter plot demonstrated some linear relationships worth exploring further.

Together, these results illustrate how machine learning can not only model chocolate quality but also expose important drivers of consumer preference, ingredient significance, and manufacturer influence.

10. Conclusion

This project demonstrates the power and versatility of machine learning in addressing real-world challenges in food quality analysis, particularly within the chocolate industry. By combining structured numerical data with unstructured sensory descriptors, and applying a variety of machine learning techniques, both supervised and unsupervised, the project successfully created a framework for classifying and predicting chocolate quality with interpretability and accuracy.

Through meticulous preprocessing including ingredient imputation, numeric transformation of cocoa percentages, and label encoding this study ensured clean and usable input for modelling. Exploratory analysis uncovered rich insights into how ingredients, origin, and memorable characteristics correlate with ratings. Visualizations such as heatmaps, box plots, and scatter plots made these relationships transparent and interpretable.

Each algorithm contributed uniquely to this goal: Random Forests offered feature importance and robust multi-class classification, SVMs provided high-dimensional decision boundaries although with class imbalance issues, and Naïve Baye**s** allowed text-based modelling of flavor notes. Meanwhile, K-Means clustering revealed natural groupings within the dataset that aligned with consumer trends, and Linear Regression quantified the partial linear relationship between cocoa percentage, manufacturer identity, and user ratings.

The performance metrics from each model highlighted both successes and limitations. Classification accuracy and clustering RMSE were promising, yet challenges like class imbalance and underfitting in regression pointed to the need for further feature enrichment and algorithm tuning.

Overall, this project lays a solid foundation for intelligent food classification and recommendation systems. Future work can build upon these findings by incorporating deep learning techniques, integrating image-based features (e.g., bar appearance), or expanding the dataset to capture more global and artisanal chocolate varieties. Ultimately, the project not only fulfils academic curiosity but also holds practical implications for manufacturers aiming to optimize recipes and market segmentation strategies and for consumers seeking data-backed quality in their chocolate choices.

11. References

1. Tan, J., Balasubramanian, B., Sukha, D., Ramkissoon, S., & Umaharan, P. (2019). Sensing fermentation degree of cocoa (Theobroma cacao L.) beans by machine learning classification models based electronic nose system. *Journal of Food Process Engineering*, *42*(6), e13175.

2. Eric, O., Gyening, R. M. O. M., Appiah, O., Takyi, K., & Appiahene, P. (2023). Cocoa beans classification using enhanced image feature extraction techniques and a regularized Artificial Neural Network model. *Engineering Applications of Artificial Intelligence*, *125*, 106736.

3. Panchbhai, K. G., & Lanjewar, M. G. (2025). Portable system for cocoa bean quality assessment using multi-output learning and augmentation. *Food Control*, *174*, 111234.

4. Omas-as, A. M., ARBOLEDA, E. R., & DAANG, J. A. M. Machine Learning as a Strategic Tool: A Comprehensive Literature Review for Advancing Agricultural Analysis, with Emphasis on the Cocoa Bean Quality Assessment. *International Journal of Scientific Research and Engineering Development*

5. Ayikpa, K. J., Gouton, P., Mamadou, D., & Ballo, A. B. (2024). Classification of Cocoa Beans by Analyzing Spectral Measurements Using Machine Learning and Genetic Algorithm. *Journal of Imaging*, *10*(1), 19.

6. Chang, Y. T., Hsueh, M. C., Hung, S. P., Lu, J. M., Peng, J. H., & Chen, S. F. (2021). Prediction of specialty coffee flavors based on near-infrared spectra using machine-and deep-learning methods. *Journal of the Science of Food and Agriculture*, *101*(11), 4705-4714.

7. Rojas, C., Ballabio, D., Consonni, V., Suárez-Estrella, D., & Todeschini, R. (2023). Classification-based machine learning approaches to predict the taste of molecules: A review. *Food Research International*, *171*, 113036.

8. https://www.kaggle.com/code/andrewmvd/chocolate-ratings-crawler/input?select=chocolate_ratings.csv

9. https://flavorsofcacao.com/database_w_REF.html

10. https://github.com/CarolineHussey/Chocolate-Analysis/blob/main/Chocolate.ipynb