QUALITY ANALYSIS OF TEA WITH THE HELP OF MACHINE LEARNING ALGORITHMS

Anjali Mathur[1], G. Prem Rishi Kranth[2], Laharika Basava[3], Maneesh Alla[4]

[1,2,3,4] *Koneru Lakshmaiah Education Foundation, Guntur(A.P)-522502*

---***---

Abstract - *Tea is a beverage that is always in high demand. Taste of tea is one of a major reason behind that. Taste depends on multiple factors like, type of tea-leafs, ingredients in tea, temperature of boiling, etc. Immense flavor of tea can easily analyze by the color, aroma and texture of tea while serving. Basically this flavor originates from tea leafs. There are multiple features that affect the quality of tea leafs like, briskness, appearance, type, color, aroma, chemical components like Caffeine, Poly-phenolic compounds, etc. If we have the details of these features in advance, we can predict and analyze the quality of tea, prepared by those leafs. Machine learning is a sub branch of artificial intelligence that works on existing data and tries to predict the future data. In the proposed research work, quality analysis of tea is determined by the significant effect of ingredients on the taste of tea and check the probability of getting a quality of tea by using machine learning algorithms. Different grades and qualities exist for all plant products and characters such as color, aroma and texture are used as indicators of the product's quality. Involving all such features, we are using following machine learning algorithms to measure tea quality: Decision tree-to determine the root cause of tea quality, Naïve Bayes to determine probability of getting different tea quality, Random Forest- to produces several decision trees, Support Vector Machine (SVM) and K-means clustering.*

***Key Words*: Decision tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM), K-means clustering.**

1. INTRODUCTION

The goal of Machine Learning is to understand the structure of the data and fit that data into models that can be understood and utilized by the people. It is a field of study that gives computers the ability to learn without being explicitly programmed. Machine Learning is considered with computer programs that automatically improve performance through experience. Machine learning algorithms can be broadly classifies in three categories:-

- Supervised learning: These algorithms are trained by giving examples. The algorithm will receive input as different data sets from those data sets the machine will be trained. Major learning algorithms are Decision Tree and Naïve Bayes.
- Unsupervised learning: These algorithms will not have any historical data. Here the system will not be provided with the right answer. Major learning algorithms are Association Rules and K-means clustering.
- Reinforcement learning. This is mainly used in robotics and navigation where trial and error methods are used. Major learning algorithms are Utility learning and Q learning.

Tea is made [15] from the infusion of the leaves of Camellia sinensis, the tea plant. After tea leaves are harvested, they undergo different treatments to produce different kinds of tea. Black teas have been fully fermented creating their black appearance. Oolong teas are fermented only partially and are lighter in color. Green teas are not fermented at all and should have no dark color. Tea leaves are also graded for size, with the tree main grades of black tea being orange pekoe, pekoe and souchong (from largest to smallest). The best green teas will have dried buds of unopened leaves.

In the datasets, there are several attributes that can affect the taste or the quality of the tea. The list of attributes that affect the quality of tea are as follows:-

- Caffeine :- Caffeine is a naturally occurring stimulant that can help us stay alert and awake. Consuming caffeine may also have some health benefits, such as improved brain function.
- Polyphenolic :- These chemical compounds affect the flavor and mouth feel and are speculated to provide potential health benefits.
- Catechins :- It is one of the polyphenol and a type of disease-fighting flavonoid and antioxidant. It is used for stimulate the nervous system.
- Appearance :-there are several factors like the broken and unbroken leaf and shape and size also the color of leaf

- Aroma :-the aroma of the tea while dry and while wet. Green tea should have a light, fresh, soothing fragrance, from a light orchid to a chestnut smell. Black tea should have a sweet, floral fragrance, and the aroma should linger. The aroma of dry oolongs can range from peach to osmanthus flowers.
- Touch:- the tea leaves are smooth or coarse, whether or not it crumbles easily, and whether it is heavy or light.
- Type:- In tea we are having different types of tea like the green tea and black tea.
- Quality:- whether the tea is good, bad, or on the average.
- Briskness:- Sometimes it's mean the strength of a flavor, the kick in the face quality a tea.

2. Literature Survey

A research work [6] was carried out with a method of sensory evaluation using fuzzy logic. It was the analysis of tea liquor made out of dried CTC tea with the help of linguistic data (i.e., excellent, very good, good, satisfactory, fair, not satisfactory) for certain attributes or parameters like touch, mouth feel, type, smell. In this work authors used certain algorithms like fuzzy logic, sensory evaluation, triangular fuzzy number, extended product of fuzzy numbers with the tea quality parameters. Finally with some of the parameters and datasets the highest accuracy tea liquor is known as the best quality. The work obtained both good and bad quality of tea using fuzzy logic.

One research work [4] was carried out on fruit pulp to improve the quality. It was based on the preparation of jam with coconut where to improve the coconut economy with the combination of pine apple pulp and guava pulp. In that research work the physicochemical characteristics were used to find the process that is the fundamental in analyzing the characteristics of food during its processing. Sensory evaluation was used to obtain a complete analysis of various properties of food as perceived by human sense. From the data given that was the percentage of pine apple pulp and guava pulp the statistical analysis was determined. The result pertaining to physiochemical characteristics of coconut jam were evaluated once in 30 days. Hence, the jam prepared at optimum condition of tender coconut pulp, pine apple pulp, guava pulp which showed good sensory acceptability and it was stored in glass bottle and plastic containers at room and refrigerated conditions for 6months.Finally the pleasing taste of coconut based jam which posses highly nutritive value, safe and fit for consumption was ready.

According to a research work [7] there are dozens of different classifications of the quality of tea, most with their own modifiers and addenda. These are usually boasted on the packages of most gourmet tea companies. Tea grading is a confusing and commonly misunderstood subject. Most folks think that a high-grade tea will be superior to a medium grade. When grading most Orthodox teas, the starting point is Pekoe (P), or a relatively whole leaf tea. Chinese Tea Grading System Chinese teas are usually numbered, first being the highest grade and down from there. There's no set stopping point, but generally 7 or 9 is what most people deal with. Again, this is specific to the leaf style and shape and how perfectly that was executed in production. It says nothing to the quality of the flavor.

3. METHODOLOGY

1. Collect the data
2. Data preparation:
    a. Noise removal
    b. Remove outliers
    c. Remove invalid values
    d. Remove null values
3. Feature extraction: Remove redundant values and extract only the important features from the given data set.
4. Algorithm: Choose appropriate machine learning algorithm. Divide the dataset into training and testing.
5. Training: 70% of the dataset is given to training, the machine learning algorithm is trained based on the given examples.
6. Testing: 30% of the dataset is given to testing, if the accuracy is good for prediction.
7. Fine Turing: If the accuracy is not reed then adjust the weights and distances
8. Prediction: Predict the results if the results are correct stop the algorithm

Table 1: Dataset

Name	Description

Caffine	High=3
	Normal=2
	Low=1

Polyhenolic	High=3
	Normal=2
	Low=1

Catechins	High=3
	Normal=2
	Low=1

Appearance	Broken=2
	Unbroken=1

Touch	Smooth =1
	Coarse =2
	Smooth without crumbles =3
	Smooth with crumbles =4

Aroma	Sweet =1

4. Implementation of Algorithms

4.1 Decision Tree:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is a type of supervised learning algorithm that is used in classification problem. The algorithm contains Pre-defined target variables. It is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision. The decision trees takes input as object or situation described by set f properties & output as Yes / No. The learning trees are represented as decision trees and represented as set of if-then rules to improve human readability. Some advantages of using decision trees are easy to understand, useful in data exploration, less data cleaning required, it can handle both numerical and categorical variables. However over-fitting is one of major disadvantage.

Decision Tree Representation:

 Root node: Represents entire sample / population
 Splitting: Dividing a node into 2 (or) more sub nodes.
 Decision node: When sub node splits into further sub nodes.
 Leaf / Terminal node: Which cannot be split.
 Pruning: When we remove sub node of a decision node.
 Branch: Part of the decision tree

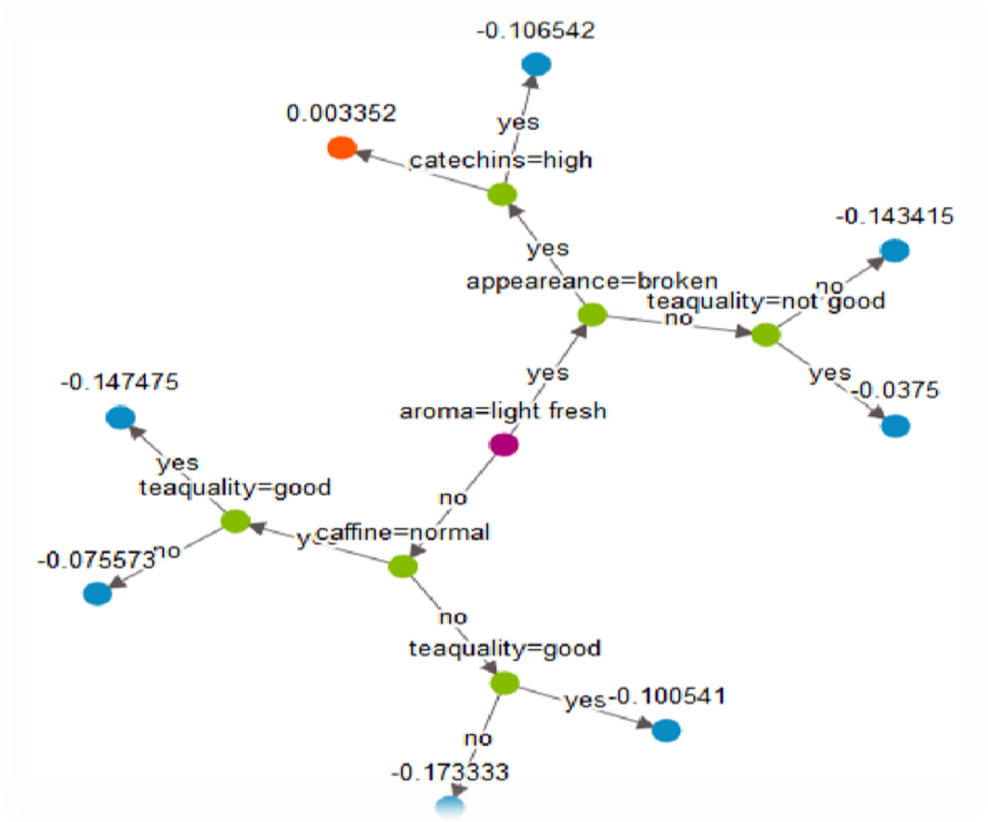


Fig 1 : Result of decision tree for green tea

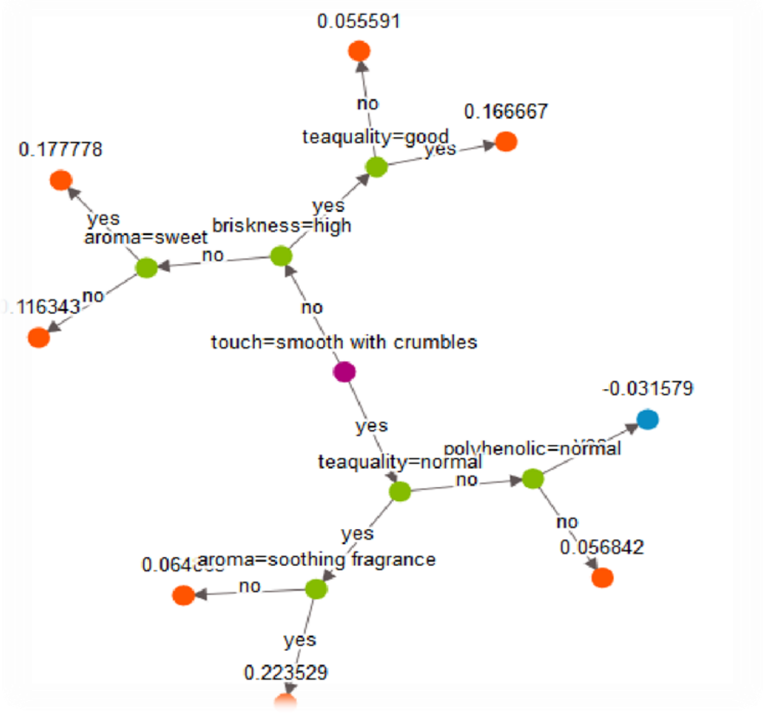


Fig 2 : Result of decision tree for black tea

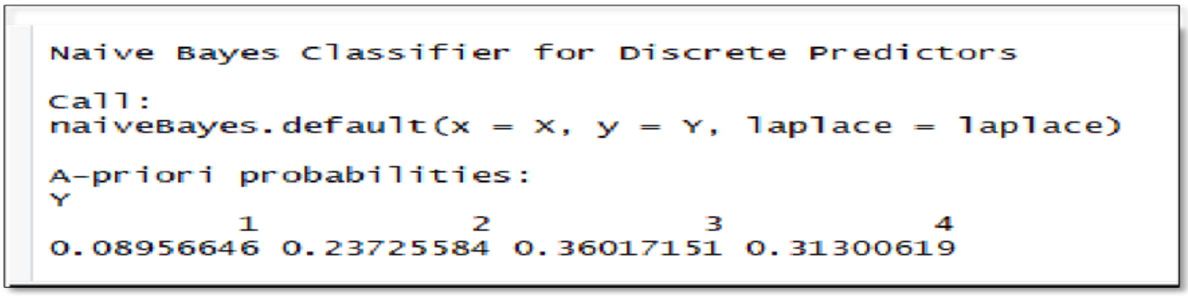
4.2 Naive Bayes:

Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability and the probabilities of observing various data.

$P(h/D)=(P(D/h)*P(h)$

A concept learning algorithm considers a finite hypothesis space H defined over an instance space X.□

Naïve bayes classifier is applied to learn tasks and each instance x is described by the conjunction of attribute values and target function f(x) of any value in some finite set V. It is based on the simplifying the assumption that all the attribute values are conditionally independent given the target value.



```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
         1          2          3          4 
0.08956646 0.23725584 0.36017151 0.31300619
```

Fig 3: Results of Naive Bayes

From the outputs generated by the Naïve Bayes it is crystal clear that that the tea quality is very good for almost 31% of the total dataset and good for 36% of the overall data. The tea quality is normal approximately 24% and not good with not even 1% of the overall data.

4.3 Random Forest:

Random forest algorithm is a **classification** algorithm. Ensemble classifier means a group of classifiers. Instead of using only one classifier to predict the target, In ensemble, we use multiple classifiers to predict the target. In case, of random forest, these ensemble classifiers are the randomly created decision trees. Each decision tree is a single classifier and the target prediction is based on the **majority voting** method. The majority voting concept is same as the political voting. Each person **votes** per one political party out all the political parties participating in elections. In the same way, **every classifier will votes** to one target class out of all the target classes.

Input: Data set and number of clusters.

Output: A set of decision trees.

Properties: Variable Importance, Relationship to nearest neighbors.

Random Forest pseudocode:

```
1.Randomly select “k” features from total “m” features. Where k << m
2.Among the “k” features, calculate the node “d” using the best split point.
3.Split the node into daughter nodes using the best split.
4.Repeat 1 to 3 steps until “l” number of nodes has been reached.
5.Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.
```

The beginning of random forest algorithm starts with randomly selecting **“k”** features out of total **“m”** features. In the next stage, we are using the randomly selected **“k”** features to find the root node by using the best split approach. We will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node. Finally, we repeat 1 to 4 stages randomly created trees forms the to create **“n”** randomly created trees. To perform prediction using the trained random forest algorithm uses the below pseudocode.

- Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
- Calculate the **votes** for each predicted target.
- Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

To perform the prediction using the trained random forest algorithm we need to pass the test features through the rules of each randomly created trees. Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Then the final random forest returns the x as the predicted target. This concept of voting is known as **majority voting**.

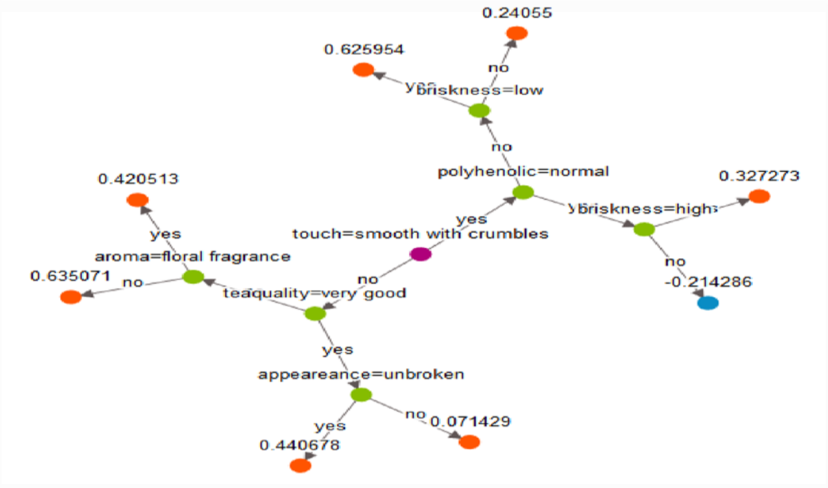


Fig 4: Results of Random Forest (1)

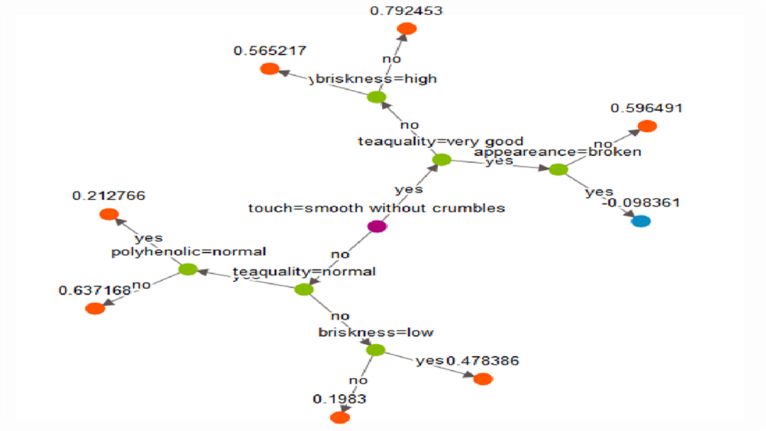


Fig 5: Results of Random Forest (2)

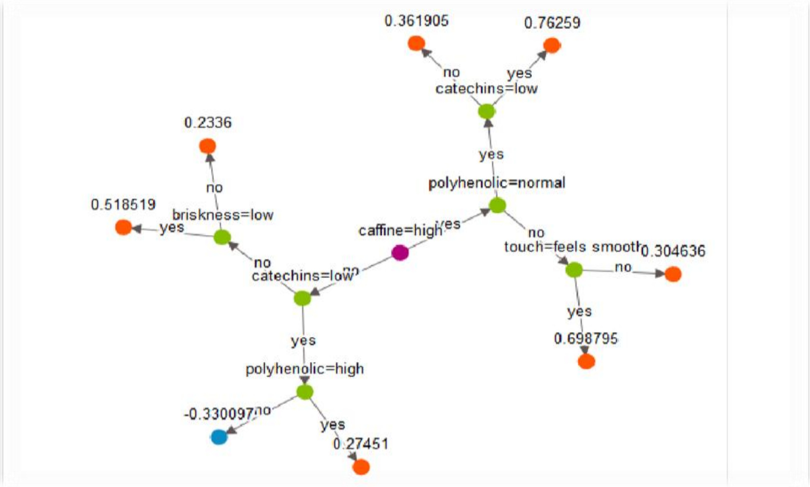


Fig 6: Results of Random Forest (3)

Random forest is an algorithm that produces several decision trees, here the random forest is performed with a condition of type is black tea. So, several trees have been generated accordingly, which means there might be several root nodes and so by using the majority voting we can set the root node more precisely. It is evident that the main root node is touch which consist the attribute smooth without crumbles or with crumbles.

- The factor that shows its major impact is touch in decision tree 1.
- The factor that states its self as a root node is caffeine in the decision tree 2.

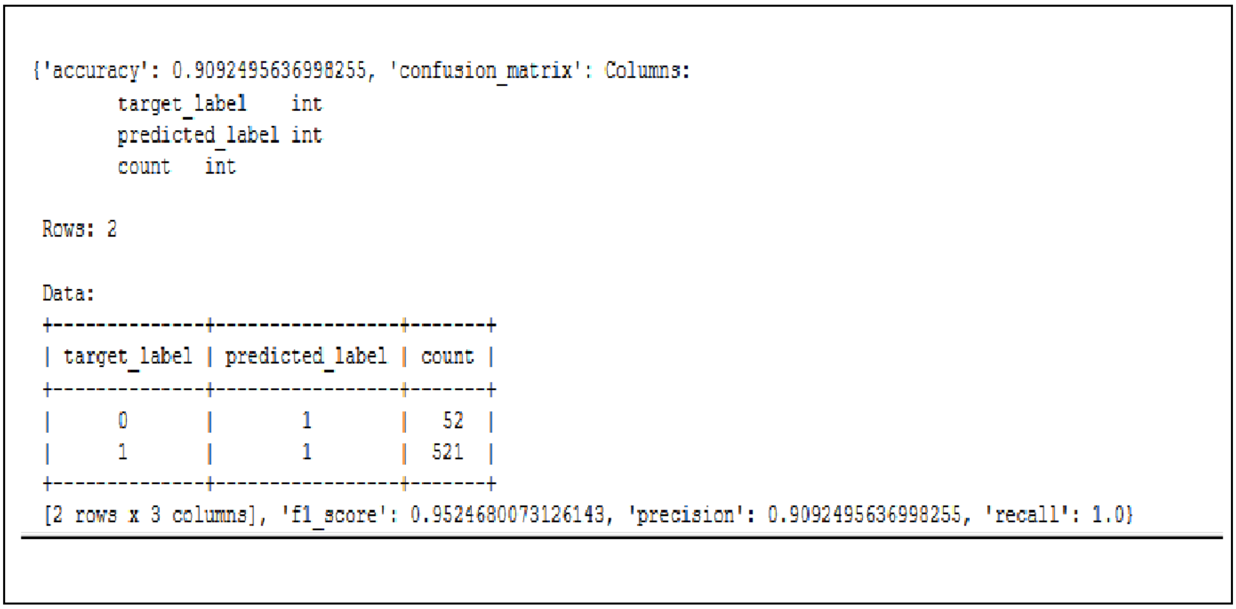
The root node in the decision tree 3 is the touch. So, by using the majority voting the root that can be considered as the factor that shows its impact in black tea is touch.

4.4 Support Vector Machine

The Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate.

Advantages:

 It works well with clear margin of separation
 It is effective in high dimensional spaces.
 It is effective in cases where number of dimensions is greater than the number of samples.
 It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient



```
{'accuracy': 0.9092495636998255, 'confusion_matrix': Columns:
        target_label    int
        predicted_label int
        count   int

 Rows: 2

 Data:
 +--------------+-----------------+-------+
 | target_label | predicted_label | count |
 +--------------+-----------------+-------+
 |      0       |        1        |   52  |
 |      1       |        1        |  521  |
 +--------------+-----------------+-------+
 [2 rows x 3 columns], 'f1_score': 0.9524680073126143, 'precision': 0.9092495636998255, 'recall': 1.0}
```

Fig7: Results of Support Vector Machine(SVM)

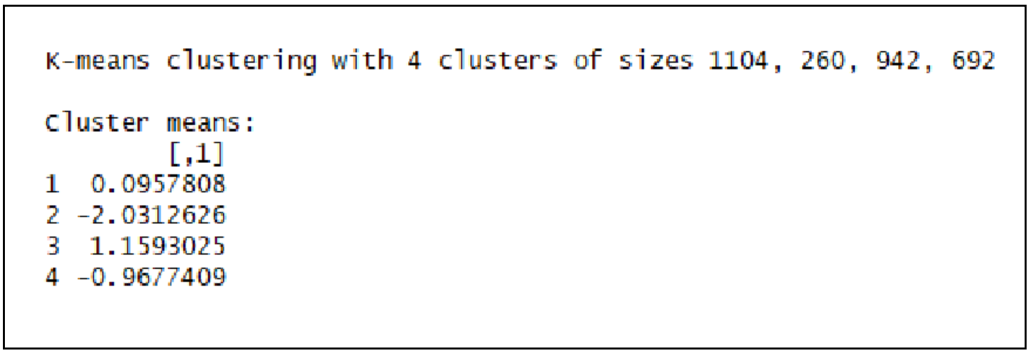
In SVM classification technique is used on the tea quality dataset and generated the desired results with an accuracy of nearly 91% and this is done under the condition of having the tea quality more than 1 i.e. the SVM models applies only for sets that consist tea quality normal or above.

4.5 K-Means Clustering:

1. *K*-means clustering is a type of unsupervised learning, is used when having unlabeled data (i.e., data without defined categories or groups).
2. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.
3. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity.
4. It is for partitioning where each cluster center is represented by the mean value of objects in cluster
5. Input: Data set and number of clusters
6. Output: A set of k clusters

Randomly selects k objects as initial clusters for the remaining objects, an object is assigned to cluster to which it is most similar based on equidistance between cluster object and cluster centers k-means algorithm improves by computing new mean values or centroids all the objects are again reassigned using the updated meansThe *K*-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters *K* and the data set.The data set is a collection of features for each data point. The algorithms starts with initial estimates for the *K* centroids, which can either be randomly generated or randomly selected from the data set.The algorithm then iterates between two steps:

a. Data assignment step
b. Centroid update step



```
K-means clustering with 4 clusters of sizes 1104, 260, 942, 692

Cluster means:
        [,1]
1  0.0957808
2 -2.0312626
3  1.1593025
4 -0.9677409
```

Fig 8: K-means Clustering

The data of the tea quality analysis is clustered into 4 different groups because tea quality there are 4 different types of responses recorded in the data. So the K-Means clustering is used for prediction of tea quality.

5. SUMMARY :

- All the machine learning methods used are mostly based on the classification technique. The methods use are supervised learning method.
- In the decision tree the output are generated according the entropy and information gain and the results that obtained are under condition and the non-condition is having the root node as the touch.

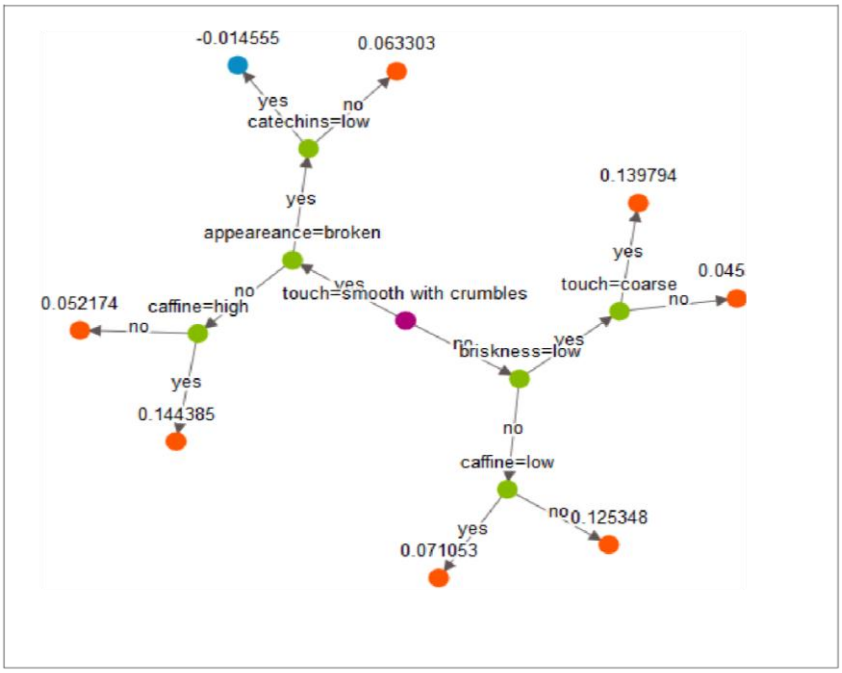


Fig 9: Results of decision tree without condition

- With the help of naïve bayes we determined the priori probabilities of quality of tea.

Table 2: probabilities of naive bayes

Tea quality	Probability
Very good	0.08956646
Good	0.23725584
Normal	0.361017151
Not good	0.31300619

- With the help of support vector machine (svm) we can do both classification and regression, but here we are doing only the classification technique and we get the accuracy of 90%. This accuracy comes under the specific condition like tea quality having greater or equal to normal.
- With the help of random forest we get several decision trees if most of the root node consists of repeated node then we can finalize that attribute as the root node. Here for the black we get touch as the root node.
- In the clustering algorithms there is an accuracy around 84%-86%, which shows classification techniques are much better for analyzing and for forecasting.

6. CONCLUSIONS

By implementing of Random Forest and Decision Tree on the data set of green tea, we got the 'Aroma' as a root node. In Random Forest the root node is elected by the number of majority voting concept. It shows that the feature 'Aroma' puts a major impact on the quality of tea. In aroma we are having nine major attribute and twenty-four dependent attributes the attribute that show much impact is the light fresh on the green tea. In the black tea the attribute that shows its impact is physical appearance by touch of the leaf. In the above information about quality analysis of tea we have taken a dataset with 9 attributes and 3000 rows.

In Random Forest the root node is elected by the number of majority voting concept. The classification technique in the SVM resulted with an accuracy nearly 91%. In the clustering algorithms there is an accuracy around 84%-86%, which shows classification techniques are much better for analyzing and for forecasting.

In future we can perform the prediction by using the sentiment analysis like having the review of the people who has tasted the tea, instead of using the normal data. By using this method it gives more précised analysis on the quality of tea when compared to the normal data that we have done at present.

Table 3: Table of accuracy

Machine Learning Algorithms	Accuracy
Decision tree	92%
Support vector machine	90%
Naïve bayes	88%
Random forest (confusion matrix)	90%
Clustering	85%

REFERENCES

[1] Michio Sugeno, Takahiro Yasukawa, A Fuzzy-Logic-Based Approach to Qualitative Modeling lEEE transactions on fuzzy system VOL. I, NO. **I**

[2] Shrimali Ronak Baghubhai, Ganga Sahay Meena, Vijay Kumar Gupta, Yogesh Khetra, Raghu HV and Ritika Puri, Sensorial and chemical changes in buffalo milk Kheer Mohan during Storage Indian J Dairy Sci 69(1), 2016

[3] G.Sindumathi, S. Amutha, Processing and quality evaluation of coconut based jam IOSR Journal Of Environmental Science, Toxicology And Food Technology.

[4] D. B. MacDougall, Principles of colour measurement for food.

[5] Chakraborty Debjani & Shrilekha Das & H. Das, Aggregatioof sensory data using fuzzy logic for sensory quality evaluation of food, J Food Sci Technology.

[6] Yuerong Liang, Jianliang Lu, Lingyun Zhang, Shan Wu and Ying Wu, Estimation of tea quality by infusion colour difference analysis, Journal of the Science of Food and Agriculture, 2005.

[7] Tea quality : https://www.teaclass.com/lesson_0210.html

[8] online references 1:- https://en.wikipedia.org/wiki/Perceptron

[9] online reference2:https://en.wikipedia.org/wiki/ID3_algorithm

[10] onlinereference3:-https://en.wikipedia.org/wiki/K-means_clustering.

[11] Online

[12] onlinereference4:http://lib3.dss.go.th/fulltext/Journal/J.Sci.Food%20and%20Agri/200585/n o.2/2005v85no2p286-292pdf.pdf.

[13] online reference:- https://sevencups.com/learn-about-tea/how-to-judge-tea/

[14] online reference http://www.fao.org/fileadmin/templates/est/COMM_MARKETS_MONITORING/Tea/Documents/Andrew_Scott_ISO_Dehli_13_May_10.pdf

[15] online reference 7: http://oregonstate.edu/instruction/bot101/Patterson/lab_pdf/Lab_6-teas_coffees_Chocolate_and_Peppers.pdf