Winter22
Capstone

Computational gastronomy: Taste Prediction

Team Members:

1. Aman Srivastava MT21007

2. Mahak Sharma MT21047

3. Rishabh Kumar Pundhir MT21071

Taste Prediction

ABSTRACT

Taste is one of the essential senses of the human being for survival. It helps us intake the correct food, which is necessary for the proper functioning and maintenance of the human body. There are five basic types of taste - sweet, bitter, sour, salty and umami. This research shows various machine learning techniques that can be used to classify taste for a molecule. Two datasets were used, one for binary classification of sweet and bitter molecules and the other for multiclass classification of nine different types of molecules. Mordred and PaDEL are the two molecular descriptors used for feature generation of each molecule with various feature selection/reduction techniques such as PCA, boruta, select-K-best and correlation using 5-fold cross-validation. PaDEL turned out to be a better molecular descriptor than Mordred for both the dataset with the highest achieved precision value of 98% for bitter molecules using adaboost, 92% for sweet molecules using random forest on BitterSweet Dataset. For ChemTasteDB, the highest achieved precision is 84.4% using xgboost.

1. Introduction

A sense is a biological system used to collect information from surroundings and respond accordingly. There are five senses of human beings - 1. Sight or vision, 2. Hearing or audition, 3. Smell or olfaction, 4. Taste or gustation, and 5. Touch or tactician. Among these five senses, taste is one of the most important senses for the survival of human beings. Taste enables us to avoid harmful food and intake only proper nutrition for the correct functioning of our human body. So, It is necessary to have a taste prediction system that can be used to classify molecules based on different types of tastes like sweet, bitter, sour, salty and umami etc.

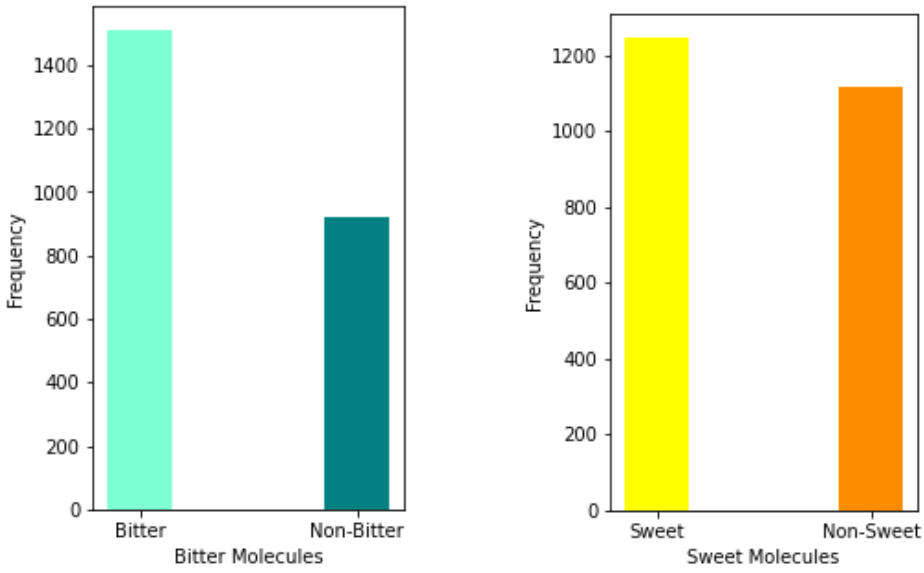
This paper aims to inspect the related work done in this field and focuses on applying various machine learning techniques to classify the taste of the molecules.
The remaining sections of this paper are ordered as follows: 2. Dataset, 3. Molecular Descriptors, 4. Literature Review, 5. Methodology, 6. Model Evaluation, 7. Results & 8. Conclusion

2. Dataset

BitterSweet and ChemTasteDB datasets are the two prominent datasets used for our study. The bittersweet dataset is collected and compiled by us, it functions upon the classification of bitter and sweet tastes. The ChemTasteDB is referred from paper [3], and it focuses on the multiclass classification of nine types of different flavors.

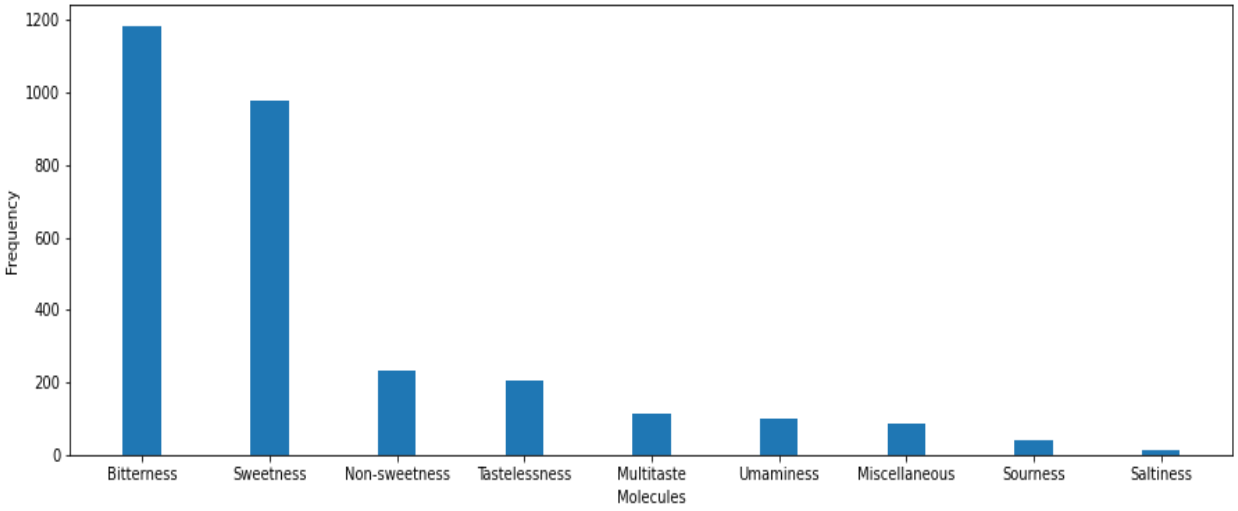
The dataset used for our training and testing of our **BitterSweet** models on split training and testing dataset with the following compositions having ‘Sweet’ as the target for Sweet models and ‘Bitter’ as the target for Bitter models. The canonical smiles column present in the dataset was used to generate the features and predict the taste of that particular molecule. For the evaluation of our models, 'Sweet' and 'Bitter' were assumed as the targets.

Class Frequency Distribution Plots For BitterSweet:



Also, we used the **ChemTasteDB** dataset, which was a single dataset with eight columns and 2944 rows, having a target column as ‘Class Taste’ containing nine unique classes, which were ‘Sweetness’, ‘Bitterness’, ‘Umaminess’, ‘Sourness’, ‘Saltiness’, ‘Multitaste’, ‘Tastelessness’, ‘Non-sweetness’, ‘Miscellaneous’.

Class Frequency Distribution Plots For ChemTasteDB:



3. Molecular descriptors

Molecular descriptors can be defined as mathematical representations of molecules' properties that are generated by algorithms. The numerical values of molecular descriptors are used to quantitatively describe the physical and chemical information of the molecules.

PaDEL and Mordred were the two molecular descriptors used for our feature generation from SMILES. PaDEL descriptor software was used for this purpose, in which a molecule.smi file (containing all the SMILES) was processed to generate all the features where only 1D and 2D features were taken into account without the Fingerprints. For the Mordred descriptor, we used the Mordred library as the prominent library for generating the features where 1D, 2D, and 3D features were extracted.

4. Literature Review:

1. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules:

This paper shows us how we can identify bitter and sweet molecules. The data is provided as bitter data consists of bitter / non-bitter molecules, and sweet data consists of sweet/non-sweet molecules. Then their features were generated using different molecular descriptors. Preprocessing of data has been done by applying PCA. After feature generation, different machine learning models have been applied like the random forest, AdaBoost with 5 -fold stratified cross-validation, and features selected using the Boruta algorithm. Then all the models were evaluated using different performance metrics such as accuracy, precision, AUC-ROC, F1, etc.

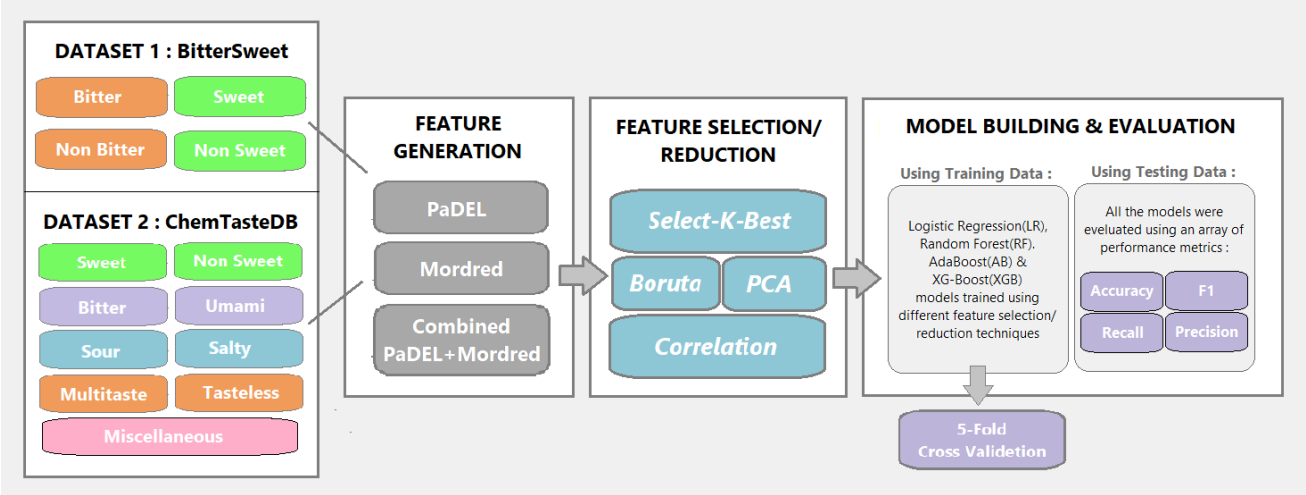
2. ChemTasteDB: A curated database of molecular tastants:

This paper has constructed a chem taste database, which includes both inorganic and organic tastants. The database consists of 2944 compounds divided into nine classes: sweet, bitter, umami, sour and salty, known as basic tastes. Some additional categories have also been included: tasteless, non-sweet, multi-taste, and miscellaneous. The information provided for each tastant by the database consists of Pub-Chem CID, CAS registry number, canonical SMILES, class taste, and references to the scientific sources from which data were retrieved. The software used for drawing the chemical structure of molecular tastants is hyperchem8, and chemical molecular structures have been checked and curated using alvaMolecule software. In-house KNIME workflow has been programmed for filtering the database,alvaDesc is used to calculate MACCS fingerprints, and t-SNE models have been calculated using MATLAB.

3. BoostSweet: Learning molecular perceptual representations of sweeteners:

According to the authors, previous machine learning investigations based on quantitative structure-activity connections have offered some molecular principles for predicting sweetness, however chemical recognition of sweetness active components can improve these models. They also discovered that the quantity of nitrogen atoms that function as hydrogen bond donors in molecules can play an important role in determining sweetness based on an examination of feature importance and dataset. Here the authors took the BitterSweet data from (Tuwani et al., 2019) paper for their study. Tree-based ML models RF, XGB, and LGBM, as well as the FCN model, were the key models employed in this work. AUROC was employed as the primary metric for hyperparameter adjustment and final validation in this study, with AUPR, F1, and NER serving as additional metrics. With an AUROC score of 0.961 and a NER score of 0.899, our LGBM soft-vote ensemble model with layered fingerprints and alvaDesc physical molecular descriptors performed the best.

5. METHODOLOGY:



A. Feature Generation -

In pattern recognition and machine learning, a feature or an independent variable represents a property or characteristic that plays a vital role in identifying an item/object and differentiates it from others. Feature generation is the process of creating one or more features with the help of existing feature(s) to improve the classification. Manually collecting the data is a costly and time-consuming task. Therefore, choosing the suitable molecular descriptors for our dataset is extremely important.

Features were generated using two molecular descriptors, PaDEL and Mordred, on both datasets. The column “SMILES” was used for generating the features from these molecular descriptors. For the BitterSweet Dataset, the PaDEL descriptor rejected five molecules for bitter data and five for sweet data, whereas Mordred denied 24 molecules for bitter data and 23 for sweet data. In the case of ChemTasteDB, the descriptors rejected no molecule. The numbers of features generated using PaDEL and Mordred molecular descriptors are 1444 and 1616, respectively, for both datasets.

Since the features generated by the descriptors contain a lot of noise, there was a need for data cleaning. For data cleaning, techniques such as removing nan, converting mixed types columns to integer/float/boolean, and replacing empty cells by 0, were applied to the generated features.

B. Feature Selection/ Reduction:

While doing feature selection, we have to get rid of the noisy features and should keep the relevant ones only. So, we have dropped the columns containing the mixed data types in the case of bitter and sweet data. In the case of padel, 2 columns, and for Mordred, 415 columns were dropped. We have also used correlation-based feature selection. First, all features that were less correlated among themselves were dropped. Then, all features’ correlation with the label was found, and those less correlated were dropped. We have kept a threshold of 0.7 so that features with a lower correlation value than this were not considered. Boruta and select k best were also used for the feature selection. The data obtained from molecular descriptors contain large numbers of columns, so we tried to reduce our data by applying feature reduction techniques.

Columns in the data set are referred to as dimensionality, and dimensionality reduction is a process where we reduce the dimensions so that the data size will be reduced but give the same essence as the original data, which will help in the training of the model as well. For feature reduction, we have applied PCA.

C. Data Cleaning

Before applying our models there was a need to clean the data as after feature generation from molecular descriptors, there was a demand to make the features consistent, for which we followed the below steps to clean the data for both BitterSweet and ChemTasteDB -

For PaDEL -

1. In bitter train replaced INFINITY (78), NAME(7) and #NAME?(51) with 0.
2. In sweet train replaced INFINITY (113) and #NAME?(51) with 0.
3. In the sweet test and train, nan values have been replaced with 0.
4. Dropped the two columns hmin and gmin due to mixed data type in them after feature generation.

For Mordred -

1. 415 columns (BitterSweet) and 245 columns (ChemTasteDB) contained the data of mixed data types.
2. Convert those columns to numeric type(integer/float/boolean).
3. Also, some of the cells were having strings like “divide / 0”, “max/min” etc.

D. Model Training & Testing

Machine Learning Models used for our study -

1. Logistic Regression - It finds out a linear relationship between explanatory variables and the target variable by giving appropriate weights to each explanatory variable.

2. Random Forest - This model is made up of various distinct decision trees which are collectively helpful in predicting the target value.

3. Adaboost - Adaptive Boosting algorithm is used to enhance the prediction classifier to improve the overall performance. Boosting is the development of several learners in a sequential order, each seeking to rectify errors made by the prior learner.

4. XGBoost - In XGBoost, weights are quite significant. All independent variables are given weights, which are subsequently input into a decision tree that predicts outcomes.

Steps involved to make data for our models -

1. We firstly trained our models on regular features generated from molecular descriptors of BitterSweet data for taste prediction.

2. Then we applied various feature extraction techniques like Correlation, PCA, SelectKbest and Boruta algorithm prior training our model.

3. After which we combined both BitterSweet's PaDEL and Mordred features to generate 3055 combined features to build a model for prediction.

4. We also trained our model on features generated from ChemTasteDB using PaDEL and Mordred molecular descriptors. Here we also applied 5 Fold cross validation methods on our generated features.

5. To get some more improvement in our prediction we also tried extending and combining both ChemTasteDB and BitterSweet features to build our final model.

6. Model Evaluation:

Evaluation metrics are used to compute the performance or quality of a machine learning model. Evaluation metric plays a vital role in comparing the results of different techniques and models in research work. A binary classification problem has four different forms of output -

True Positive(TP): The number of positive labels which are correctly predicted as positive
True Negative(TN): The number of negative labels which are correctly predicted as negative
False Positive(FP): The number of negative labels which are incorrectly predicted as positive
False Negative(FN): The number of positive labels which are correctly predicted as negative.

In this paper, we are using “accuracy”, “precision”, “recall” & “f1-score” as the evaluation metrics for testing and comparing our results.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

7. Results

DataSet -> (A) BitterSweet:

1. Sweet Data with Padel

	Precision	Recall	F1-Score	Accuracy
Logistic Regression:				
W/o feature selection/reduction	0.83	0.52	0.64	0.61
PCA(n=2)	0.68	**0.98**	0.80	0.68
PCA(n=10)	0.65	0.84	0.73	0.60
Corr on Features	0.73	0.75	0.74	0.65
Corr on Features & Y(Target)	0.72	0.75	0.73	0.64
Select K(K=200)	0.92	0.63	0.75	0.72
Random Forest				
W/o feature selection/reduction	0.89	0.72	0.79	0.75
PCA(n=2)	0.5	0.009	0.01	0.32
PCA(n=10)	0.60	0.31	0.41	0.40
Corr on Features	0.90	0.67	0.77	0.73
Corr on Features & Y(Target)	**0.92**	0.64	0.76	0.72
Select K(K=200)	0.91	0.71	0.80	0.76

AdaBoost				
W/o feature selection/reduction	0.86	0.73	0.79	0.74
PCA(n=2)	0.33	0.009	0.01	0.32
PCA(n=10)	0.56	0.32	0.41	0.37
Corr on Features	0.86	0.75	0.80	0.75
Corr on Features & Y(Target)	0.86	0.75	0.80	0.75
Select K(K=200)	0.83	0.73	0.77	0.72
XGBoost				
W/o feature selection/reduction	0.89	0.75	0.82	0.77
PCA(n=2)	0.5	0.009	0.01	0.32
PCA(n=10)	0.64	0.46	0.53	0.46
Corr on Features	0.89	0.76	**0.82**	0.78
Corr on Features & Y(Target)	0.90	0.75	0.82	0.78
Select K(K=200)	0.89	0.75	0.82	0.77

2. Bitter Data with PaDEL

	Precision	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression:				
W/o feature selection/reduction	0.74	0.57	0.64	0.61
PCA(n=2)	0.14	0.09	0.01	0.35
PCA(n=10)	0.37	0.05	0.09	0.36
Corr on Features	0.80	0.42	0.55	0.58
Corr on Features & Y(Target)	0.8	0.41	0.55	0.57
Select K(K=200)	0.70	0.48	0.57	0.56

Random Forest				
W/o feature selection/reduction	**0.93**	0.67	0.78	0.77
PCA(n=2)	0.61	1.0	0.76	0.61
PCA(n=10)	0.61	0.99	0.75	0.61
Corr on Features	0.92	0.69	0.79	0.78
Corr on Features & Y(Target)	0.91	0.69	0.78	0.77
Select K(K=200)	0.92	0.70	0.80	0.78
AdaBoost				
W/o feature selection/reduction	0.80	0.64	0.71	0.69
PCA(n=2)	0.62	0.99	0.76	0.62
PCA(n=10)	0.66	0.66	0.66	0.59
Corr on Features	0.79	0.65	0.71	0.68
Corr on Features & Y(Target)	0.79	0.65	0.71	0.68
Select K(K=200)	0.85	0.69	0.76	0.74
XGBoost				
W/o feature selection/reduction	0.85	0.75	0.80	0.77
PCA(n=2)	0.61	**0.99**	0.76	0.63
PCA(n=10)	0.58	0.61	0.60	0.49
Corr on Features	0.90	0.74	0.81	0.79
Corr on Features & Y(Target)	0.85	0.69	0.76	0.74
Select K(K=200)	0.91	0.76	**0.83**	**0.81**

3. Sweet Data With Mordred

	Precision	Recall	F1-Score	Accuracy
Logistic Regression:				
W/o feature selection/reduction	0.0	0.0	0.0	0.32
PCA(n=2)	0.66	**0.96**	**0.78**	0.64
PCA(n=10)	0.66	0.96	0.78	0.64
Corr on Features	0.0	0.0	0.0	0.32
Corr on Features & Y(Target)	0.0	0.0	0.0	0.32
Select K(K=200)	0.0	0.0	0.0	0.32
Random Forest				
W/o feature selection/reduction	0.85	0.65	0.74	0.69
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.0	0.0	0.0	0.32
Corr on Features	**0.87**	0.60	0.71	0.67
Corr on Features & Y(Target)	0.85	0.61	0.71	0.66
Select K(K=200)	0.84	0.62	0.71	0.66
AdaBoost				
W/o feature selection/reduction	0.83	0.69	0.76	0.70
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.53	0.27	0.36	0.35
Corr on Features	0.82	0.69	0.75	0.69
Corr on Features & Y(Target)	0.79	0.59	0.67	0.62

Select K(K=200)	0.80	0.69	0.75	0.68
XGBoost				
W/o feature selection/reduction	0.85	0.69	0.77	**0.71**
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.61	0.10	0.18	0.35
Corr on Features	0.85	0.66	0.75	0.69
Corr on Features & Y(Target)	0.83	0.66	0.73	0.68
Select K(K=200)	0.82	0.69	0.75	0.69

4. Bitter Data With Mordred

	Precision	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression:				
W/o feature selection/reduction	1.0	0.009	0.01	0.39
PCA(n=2)	0.48	0.51	0.5	0.36
PCA(n=10)	0.58	0.63	0.61	0.50
Corr on Features	0.61	1.0	0.76	0.61
Corr on Features & Y(Target)	0.61	1.0	0.76	0.61
Select K(K=200)	0.61	1.0	0.76	0.61

Random Forest				
W/o feature selection/reduction	0.84	**0.75**	0.79	0.76
PCA(n=2)	0.0	0.0	0.0	0.38
PCA(n=10)	0.61	0.88	0.72	0.58
Corr on Features	0.89	0.75	0.81	0.79
Corr on Features & Y(Target)	0.85	0.76	0.80	0.77
Select K(K=200)	0.85	0.76	0.80	0.77
AdaBoost				
W/o feature selection/reduction	0.87	0.72	0.79	0.76
PCA(n=2)	0.0	0.0	0.0	0.38
PCA(n=10)	0.62	0.30	0.41	0.46
Corr on Features	0.86	0.60	0.71	0.70
Corr on Features & Y(Target)	0.86	0.60	0.71	0.70
Select K(K=200)	0.80	0.73	0.76	0.72
XGBoost				
W/o feature selection/reduction	**0.88**	0.8	**0.84**	**0.81**
PCA(n=2)	0.0	0.0	0.0	0.38
PCA(n=10)	0.67	0.74	0.70	0.62
Corr on Features	0.84	0.71	0.77	0.74
Corr on Features & Y(Target)	0.88	0.71	0.78	0.76
Select K(K=200)	0.82	0.74	0.78	0.74

DataSet -> (B) BitterSweet Combined = PaDEL + Mordred:

1. Sweet Data

	Precision	Recall	F1-Score	Accuracy
Logistic Regression:				
W/o feature selection/reduction	1.0	0.009	0.01	0.33
PCA(n=2)	0.66	**0.96**	0.78	0.64
PCA(n=10)	0.67	0.89	0.76	0.64
Corr on Features	0.0	0.0	0.0	0.32
Corr on Features & Y(Target)	0.0	0.0	0.0	0.32
Select K(K=200)	0.74	0.87	0.80	0.71
Random Forest				
W/o feature selection/reduction	0.82	0.70	0.76	0.70
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.0	0.0	0.0	0.32
Corr on Features	**0.87**	0.65	0.74	0.69
Corr on Features & Y(Target)	0.84	0.67	0.75	0.69
Select K(K=200)	0.86	0.69	0.77	0.72

AdaBoost				
W/o feature selection/reduction	0.75	0.79	0.77	0.68
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.33	0.009	0.01	0.32
Corr on Features	0.78	0.72	0.75	0.68
Corr on Features & Y(Target)	0.78	0.72	0.75	0.68
Select K(K=200)	0.78	0.79	0.78	0.71
XGBoost				
W/o feature selection/reduction	0.84	0.81	**0.83**	**0.77**
PCA(n=2)	0.0	0.0	0.0	0.32
PCA(n=10)	0.0	0.0	0.0	0.32
Corr on Features	0.83	0.82	0.82	0.77
Corr on Features & Y(Target)	0.83	0.82	0.82	0.77
Select K(K=200)	0.83	0.82	0.82	0.77

2. Bitter Data

	Precision	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression:				
W/o feature selection/reduction	0.72	0.95	0.82	0.70
PCA(n=2)	0.0	0.0	0.0	0.26
PCA(n=10)	0.61	0.46	0.52	0.40
Corr on Features	0.75	**0.98**	**0.85**	0.76
Corr on Features & Y(Target)	0.75	**0.98**	**0.85**	0.76
Select K(K=200)	0.72	1.0	0.83	0.72

Random Forest				
W/o feature selection/reduction	0.95	0.66	0.78	0.73
PCA(n=2)	0.0	0.0	0.0	0.27
PCA(n=10)	0.75	0.77	0.76	0.65
Corr on Features	0.96	0.59	0.73	0.68
Corr on Features & Y(Target)	0.94	0.62	0.75	0.70
Select K(K=200)	0.95	0.68	0.79	0.74
AdaBoost				
W/o feature selection/reduction	**0.98**	0.74	0.84	**0.80**
PCA(n=2)	0.0	0.0	0.0	0.27
PCA(n=10)	0.87	0.48	0.62	0.57
Corr on Features	0.96	0.72	0.82	0.78
Corr on Features & Y(Target)	0.96	0.72	0.82	0.78
Select K(K=200)	0.93	0.68	0.79	0.73
XGBoost				
W/o feature selection/reduction	0.95	0.68	0.79	0.74
PCA(n=2)	0.0	0.0	0.0	0.27
PCA(n=10)	0.77	0.53	0.63	0.55
Corr on Features	0.96	0.69	0.81	0.76
Corr on Features & Y(Target)	0.96	0.69	0.81	0.76
Select K(K=200)	0.95	0.68	0.79	0.74

DataSet -> (C): ChemTasteDB

1. Mordred

	Precision	F1-Score	Accuracy
Logistic Regression:			
W/o feature selection/reduction	0.30	0.33	0.38
PCA(n=2)	0.23	0.24	0.35
PCA(n=10)	0.23	0.25	0.35
Corr on Features	0.25	0.25	0.36
Select K(K=200)	0.15	0.21	0.38
Random Forest			
W/o feature selection/reduction	**0.71**	0.70	0.72
PCA(n=2)	0.12	0.08	0.08
PCA(n=10)	0.27	0.24	0.38
Corr on Features	0.70	0.68	0.71
Select K(K=200)	0.69	0.69	0.71
AdaBoost			
W/o feature selection/reduction	0.46	0.47	0.51
PCA(n=2)	0.15	0.21	0.38
PCA(n=10)	0.50	0.16	0.19
Corr on Features	0.40	0.45	0.54
Select K(K=200)	0.40	0.45	0.54

XGBoost			
W/o feature selection/reduction	0.71	0.70	0.72
PCA(n=2)	0.11	0.08	0.08
PCA(n=10)	0.35	0.12	0.38
Corr on Features	0.71	**0.71**	**0.72**
Select K(K=200)	0.69	0.70	0.71

2. Padel

	Precision	**F1-Score**	**Accuracy**
Logistic Regression:			
W/o feature selection/reduction	0.40	0.34	0.45
PCA(n=2)	0.50	0.21	0.36
PCA(n=10)	0.52	0.21	0.35
Corr on Features	0.36	0.37	0.45
Select K(K=200)	0.54	0.56	0.61
Random Forest			
W/o feature selection/reduction	0.72	0.73	0.75
PCA(n=2)	0.28	0.04	0.06
PCA(n=10)	0.35	0.23	0.36
Corr on Features	0.73	0.73	0.75
Select K(K=200)	0.73	0.73	0.75

AdaBoost			
W/o feature selection/reduction	0.43	0.47	0.55
PCA(n=2)	0.22	0.03	0.03
PCA(n=10)	0.32	0.21	0.35
Corr on Features	0.38	0.34	0.41
Select K(K=200)	0.47	0.50	0.57
XGBoost			
W/o feature selection/reduction	**0.75**	**0.75**	**0.76**
PCA(n=2)	0.40	0.22	0.36
PCA(n=10)	0.30	0.11	0.39
Corr on Features	0.72	0.73	0.74
Select K(K=200)	0.74	0.74	0.75

DataSet -> (D): Extended ChemTasteDB

1. Mordred

	Precision	F1-Score	Accuracy
Logistic Regression:			
W/o feature selection/reduction	0.33	0.32	0.40
PCA(n=2)	0.44	0.19	0.27
PCA(n=10)	0.42	0.19	0.27
Corr on Features	0.42	0.32	0.40
Select K(K=200)	0.17	0.24	0.40
Random Forest			
W/o feature selection/reduction	0.75	0.74	0.75
PCA(n=2)	0.35	0.20	0.30
PCA(n=10)	0.41	0.43	0.45
Corr on Features	0.75	0.74	0.75
Select K(K=200)	0.75	0.75	0.76
AdaBoost			
W/o feature selection/reduction	0.76	0.49	0.54
PCA(n=2)	0.22	0.22	0.34
PCA(n=10)	0.45	0.42	0.42
Corr on Features	0.46	0.49	0.54
Select K(K=200)	0.46	0.49	0.54

XGBoost			
W/o feature selection/reduction	0.76	0.75	0.76
PCA(n=2)	0.26	0.19	0.29
PCA(n=10)	0.44	0.41	0.44
Corr on Features	**0.76**	**0.75**	**0.76**
Select K(K=200)	0.75	0.75	0.76

2. Padel

	Precision	**F1-Score**	**Accuracy**
Logistic Regression:			
W/o feature selection/reduction	0.49	0.47	0.53
PCA(n=2)	0.38	0.29	0.43
PCA(n=10)	0.42	0.29	0.42
Corr on Features	0.45	0.40	0.48
Select K(K=200)	0.59	0.58	0.62
Random Forest			
W/o feature selection/reduction	0.82	0.82	0.83
PCA(n=2)	0.23	0.16	0.28
PCA(n=10)	0.35	0.30	0.40
Corr on Features	0.83	0.83	0.83
Select K(K=200)	0.82	0.82	0.83

AdaBoost			
W/o feature selection/reduction	0.45	0.44	0.47
PCA(n=2)	0.22	0.13	0.20
PCA(n=10)	0.21	0.11	0.16
Corr on Features	0.47	0.48	0.51
Select K(K=200)	0.46	0.47	0.50
XGBoost			
W/o feature selection/reduction	0.83	0.83	0.84
PCA(n=2)	0.21	0.12	0.22
PCA(n=10)	0.33	0.24	0.36
Corr on Features	**0.84**	**0.84**	**0.85**
Select K(K=200)	0.83	0.83	0.84

8. Conclusion

Taste prediction is necessary for the survival of human beings. It enables us to intake proper food and nourishment for correct functioning and maintenance of our body.

After analyzing various machine learning algorithms and techniques on two different datasets i.e. BitterSweet and ChemTasteDB, we came to the conclusion that :

For BitterSweet data, we got **92% precision for Sweet data** using **Random Forest**. Also for **Bitter data we got 98% precision** using **AdaBoost** Forest after **combining the features** generated from PaDEL and Mordred molecular descriptors which shows better results than our literature review research papers.

Extending ChemTasteDB with sweet and bitter molecules of the BitterSweet dataset (Extended ChemTasteDB) has given better results than the original dataset. The highest achieved **precision, f1-score and accuracy** are **84%, 84% and 85%** respectively using **PaDEL** molecular descriptor on **XGboost** with feature selection using **correlation on features** technique.