ODOR PREDICTION
Using Machine Learning

By Shubham Sharma, Vishal Kumar and Rahul Gupta

Under the Guidance of

Prof Dr Ganesh Bagler

Abstract

In the chemical realm, machine learning and data analytics are rapidly being employed for quantitative structure property relation (QSPR) applications. One such application is the perception of odorant stimuli, as olfaction is the least worked among all other senses in the prediction domain. The usefulness of employing a data-driven strategy to predict the perceptual qualities of an odorant, specifically the odorant characters, is investigated in this work using machine learning-based algorithms and data analytics. We first analyze a Odor dataset [1] which is a repository of odor name, smiles, primary and secondary odor. We next utilize the data to train multiple machine learning algorithms [2] for olfactory character prediction, including random forest, decision tree, XGBoost, and K-nearest-neighbors, and provide the structural elements that correlate well with the odor based on the best model. Furthermore, we investigate the influence of data quality on model performance by comparing the semantic descriptors often associated with a certain odorant to how the majority of the participants perceive it. The research gives a framework for constructing odor perception models, as well as insights into odor perception and the impact of inherent bias in perception data on model performance. The algorithms and techniques described here might be utilized to anticipate novel odorant odor characteristics. In this paper , we describe a model that combines data preprocessing and scaling approaches to clean the dataset, then different classifiers are applied to the dataset and we obtained the highest accuracy of 97.01% using a random forest classifier on the test data.

Introduction

One or more volatilized chemical components, which are normally found at low quantities and may be detected by humans and animals using their sense of smell, cause an odor or scent. An odor is sometimes known as a "smell" or "scent," and it can be either pleasant or unpleasant [3]. Fragrances are also commonly added to a variety of chemical goods or formulations (body lotions, soaps, creams, detergents, and so on) to increase their sensory qualities by disguising the items' otherwise "chemical" scent.

In terms of what causes an odorant to smell the way it does, olfaction is the least known of all the senses. According to research, there may be discrepancies in how specialists and unskilled participants perceive odors [4,5,6]. In the business, there is limited consensus on how to quantify odorant qualities including quality, intensity, and resemblance. Researchers use a range of methods to obtain odor perception data, including verbal profiling, similarity scores, and sorting [7,8]. Over the years, scientists have sought to figure out how physical stimulation

affects olfactory perception. The stimulus-percept issue, on the other hand, is fraught with difficulties. In addition to the unpredictability of structure-odor connections, the extent and dimensionality of the olfactory perception space are unclear [9,10,11].

It has been proposed that, rather than only the chemical properties of the stimuli, experience variables such as memory are crucial for odour discrimination [12]. Nonetheless, as the area of machine learning has progressed, there has been an increasing interest in adopting a data-driven method to forecast structure-odor connections in recent years. Many people have attempted to show that structural properties of odorant molecules may be used to predict olfactory perception [13]. However, the methodologies employed for the predictions range substantially in terms of the data utilized, with some using perceptual data from untrained persons and others using qualitative data from experienced experts, making comparison impossible.

The goal of this work is to take untrained individuals' perceptual data and apply a machine learning-based classification strategy to predict the 68 distinct odor characteristics (OC) of odorant molecules using 11 descriptors as input features.

The study has been divided into 3 sub parts

1. Literature review
2. Proposed Methodology
3. Results and Analysis

Literature Review

This study paper covers four distinct aspects of olfaction, however our focus was on Machine-learned odor detection based on physico-chemical characteristics of volatile molecules. In this section, the author outlines how the dataset was gathered and what the various measurements of accuracy are. A classifier based on RF correctly predicted 8 out of 19 evaluated semantic descriptors ("garlic," "fish," "sweet," "fruit," "burnt," "spices," "flower," "sour"). SVM, RF, and extreme learning machines were among the machine-learning methods employed. A large psychophysical data collection was compiled from 49 people who profiled 476 structurally and perceptually distinct compounds. ELM has the highest identification accuracy (97.53%), followed by SVM (97.19%) and RF (97.19%). (92.79 percent) [13].

The goal of the study is to anticipate the olfactory features, which are mostly sweet and musky. For olfactory character prediction, they use several machine learning methods such as random forest, gradient boosting, and support vector machine. The work proposes a technique for constructing odor perception models, as well as insights into how untrained human participants perceive odorants and the impact of inherent bias in the perception data on model performance. Acid, ammonia/urinous, bakery, burned, chilly, chemical, decaying, edible, fish, floral, fruit, garlic, grass, musky, sour, spices, sweaty, sweet, warm, and wood were among the semantic descriptors. The odorant property data for 480 structurally varied chemicals is included in the dataset [14].

The "Flavors and Fragrances" library is used to create an olfactory character predicting model in this article. In the Sigma-Aldrich catalog, the descriptor's application is supplied in binary form, but in Dravnieks' sensory evaluation, it is offered on a scale of 0 to 5. For data with a binary representation. The article and its supporting information files include half of the necessary data (odor character data). The NIST database [16] has the other half (mass spectrum) of the data [15].

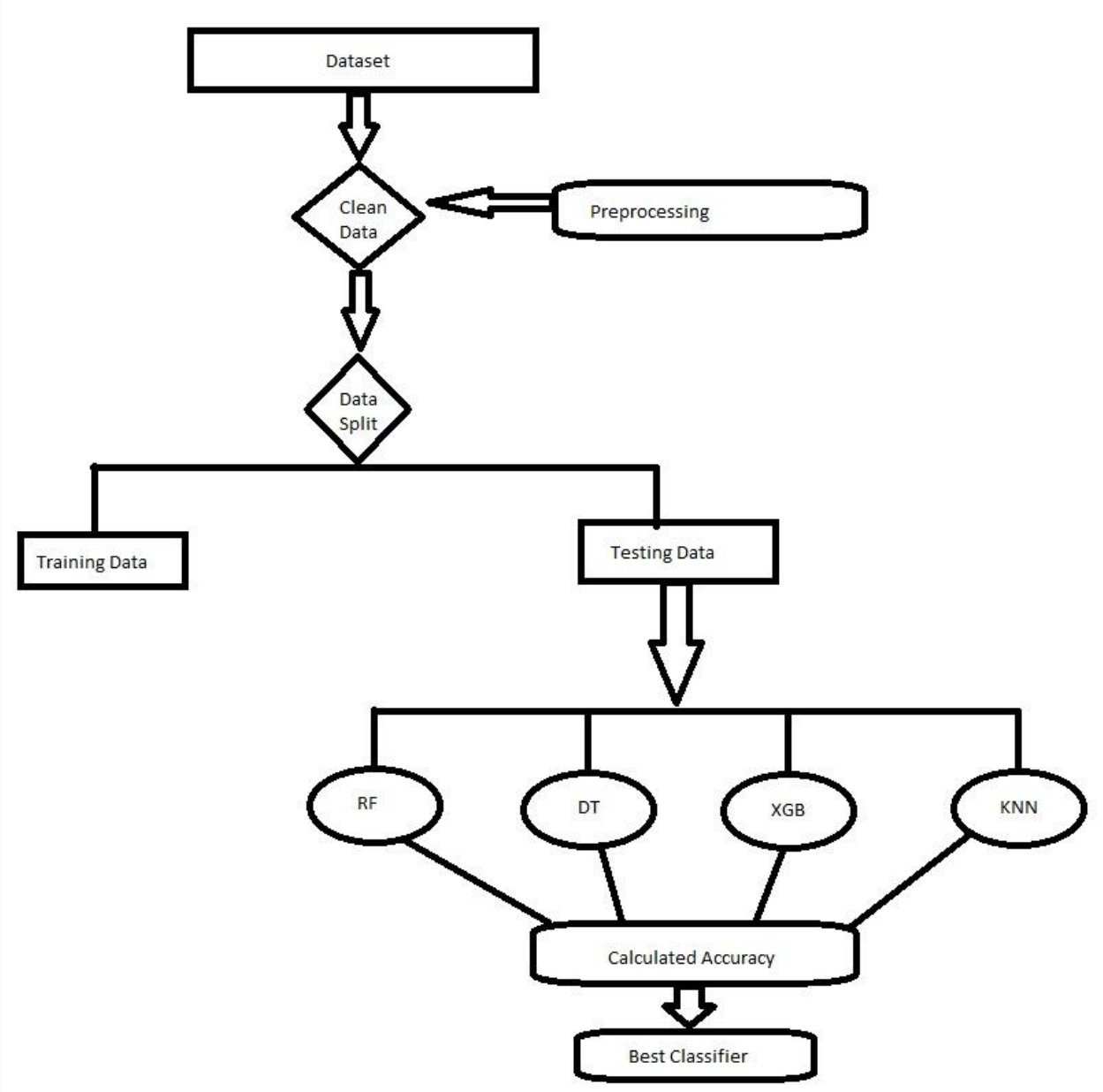
This study proposes a proof-of-concept approach for obtaining olfactory information from odorant molecules' molecular properties (MPs) using machine learning-based prediction. Molecular computation software was used to get the physicochemical properties of the odorant molecules (DRAGON). The Dragon chemo information programme was used to import SMILES strings. The MPs' characteristics were retrieved using either unsupervised (principal component analysis) or supervised (Boruta, BR) methods, and then utilized as input to machine-learning models to calibrate them. The findings revealed that SVM-calibrated models had higher accuracies than others [17].

By collecting OMs from rat olfactory bulbs and extracting feature profiles of the associated odorant molecules, this study investigated the relationship between order maps (OMs) and molecular parameters (MPs) of odorants. The OdorMapDB provided 178 pictures of glomerular (network of tiny blood veins) activity in the olfactory bulb that corresponded to odorants. In 2D space, all odorants were mapped, and comparable odorants were aggregated in t-SNE space. The models were calibrated using olfactory information or molecular data to see how well they could identify functional groupings. The OM-PCA19 ELM (extreme learning machine) offers a lot of potential for identifying functional categories [18].

Proposed Methodology

The proposed approach worked on a odor dataset [1] and predicted the odor of the given molecule. First, we have to analyze the dataset; then, we apply preprocessing on the dataset; after that, we create dummy variables (One Hot Encoding) for categorical values, then In an 80:20 ratio, we split the data into training and testing. After splitting the dataset, we apply various machine learning classifiers.

The proposed approach's framework is depicted in Figure 1.



Fig(1): Proposed Methodology

1. Dataset Description

We have worked on the Odor Dataset [1] which we have obtained from the Olfaction base section of IIIT Allahabad. The dataset consists of 3686 unique molecule information, this dataset contains 5 attributes namely Primary Odor, Sub-Odor, CAS-id, chemical name and smiles.

2. Dataset Preprocessing

We used the padelpy library of python and generated 1875 descriptors. Then we used Pearson correlation with Primary Odor as a target variable and generated 13 descriptors which were highly correlated with the target variable. After that we used correlation among the 13 descriptors and used the 0.99 threshold to generate 11 descriptors that were used for final prediction.

```
Index(['nS', 'AATS8i', 'AATSC1i', 'MATS1i', 'SM1_Dzs', 'SpMin6_Bhm',
       'SpMin7_Bhm', 'minHBa', 'gmin', 'MAXDN', 'ETA_Psi_1', 'Primary Odor'],
      dtype='object')
12
```

Fig(2): Dataset with 11 descriptors and 1 target variable

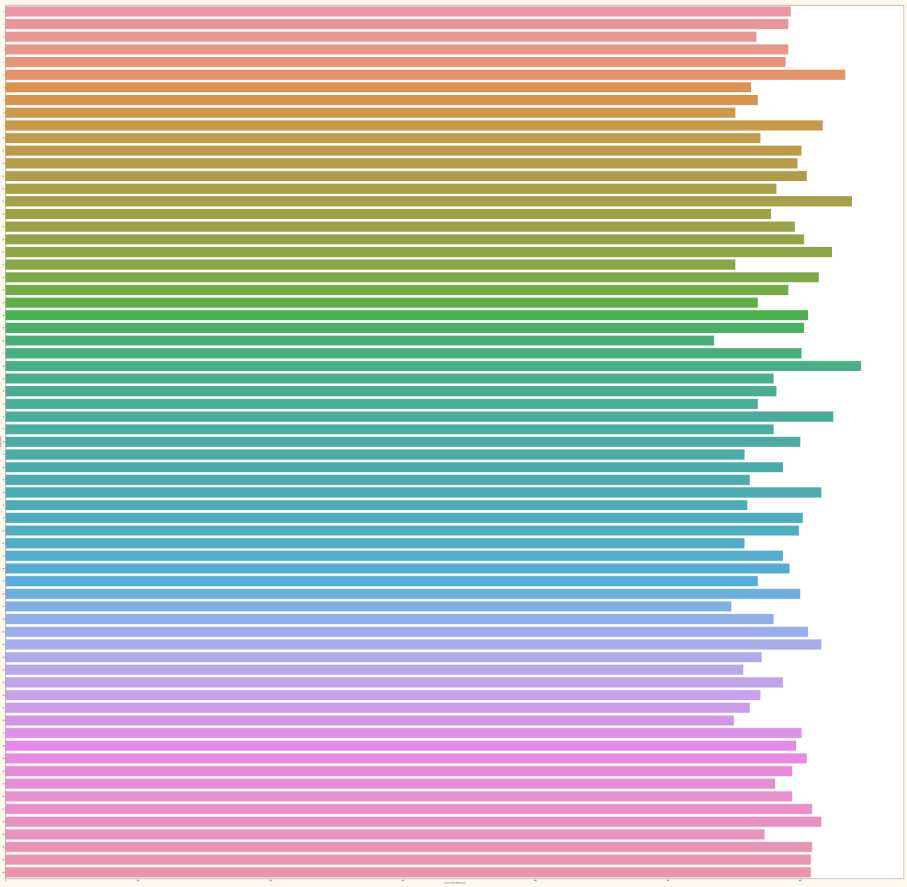
```
Score:  [1.76127945e+04 4.86215913e+05 2.08792612e+03 1.59196886e+03
 1.96011348e+04 2.65091540e+03 3.63092270e+03 1.51829871e+04
 5.45063048e+04 5.37428012e+03 1.49251532e+02]
Columns:  Index(['nS', 'AATS8i', 'AATSC1i', 'MATS1i', 'SM1_Dzs', 'SpMin6_Bhm',
       'SpMin7_Bhm', 'minHBa', 'gmin', 'MAXDN', 'ETA_Psi_1'],
      dtype='object')
```

Fig(3): Chi2 Feature Selection with Scores

```
Score:  [334.98792417 268.66128205 306.4886707  260.35426046 326.39462235
 340.64812168 290.05154283 214.11588818 419.89170005 242.54792843
 347.71306518]
Columns:  Index(['nS', 'AATS8i', 'AATSC1i', 'MATS1i', 'SM1_Dzs', 'SpMin6_Bhm',
       'SpMin7_Bhm', 'minHBa', 'gmin', 'MAXDN', 'ETA_Psi_1'],
      dtype='object')
```

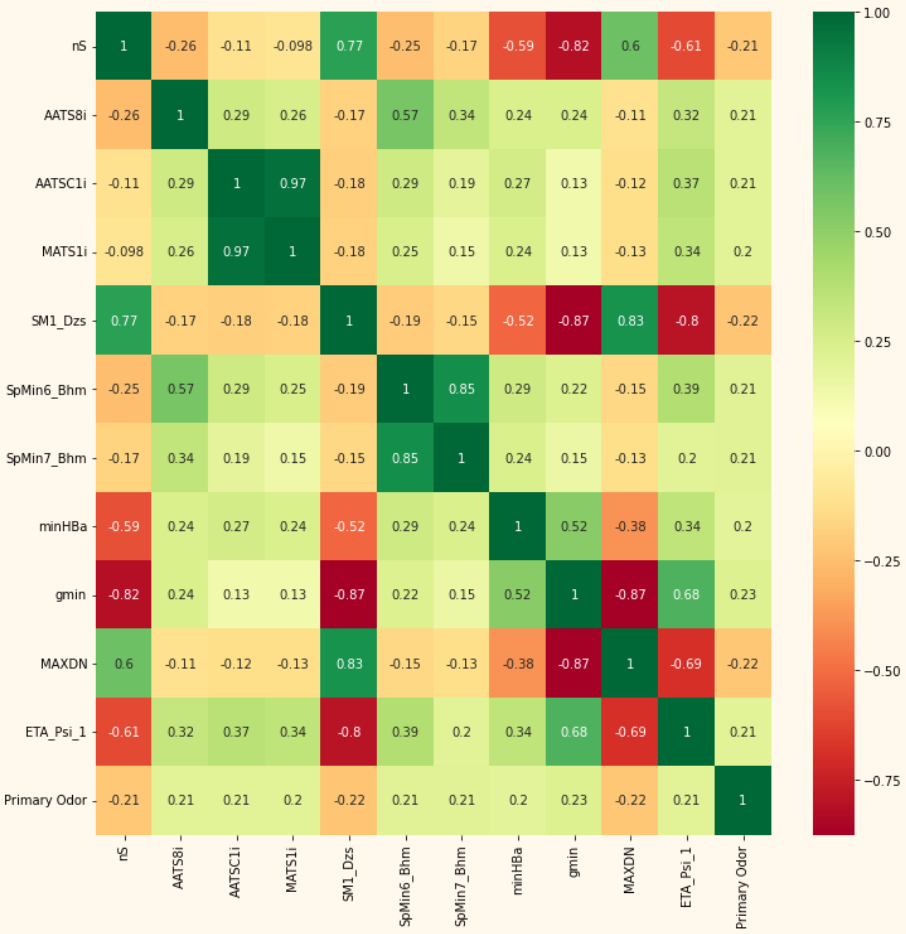
Fig(4): F_classif Feature Selection with Scores

After this we used resampling of classes/weights such that the 68 unique odor characteristics (OC) had equal distribution of count. Which made the size of the dataset as 40641 rows.

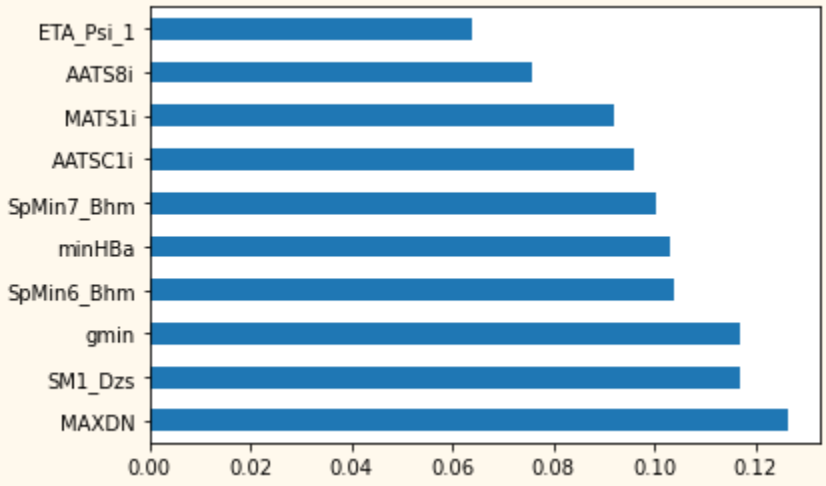


Fig(5): Dataset Primary Odor count of each Target class

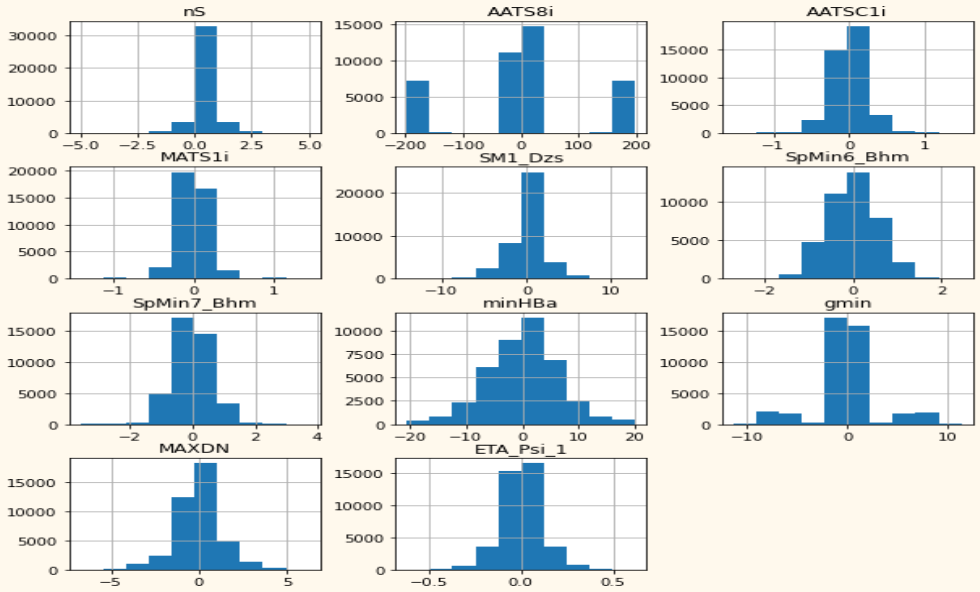
In the preprocessing part of the dataset we have also worked on the **Robust scaler** (Use statistics that are resistant to outliers to scale features. The median is removed, and the data is scaled according to the quantile range (defaults to IQR: Interquartile Range). The interquartile range (IQR) is the distance between the first and third quartiles (25th and 3rd quantiles) (75th quantile).) and **Quantile Transformer** (Quantile information is used to transform features. The characteristics are transformed into a uniform or normal distribution using this procedure.).



Fig(6): Distribution of correlation between all the 11 features and target feature



Fig(7): Most important features according to the Target



Fig(8): Features Distribution According to count and scale

3. Machine Learning Classifiers

As we get the 11 features and 1 target variable, we divide the dataset in two parts X and Y with X.Shape=(27852,11) and Y.Shape=(27852,). Such that X contains all the 11 features and Y contains only the target variable.

Then we use the Train_Test_Split function to split the dataset into Train and Test with 80:20, such that the Training dataset contains 80% of the data and the testing dataset contains 20% of the data.

Then We apply different machine learning classifiers on the dataset, such as Random Forest, Decision Tree, K-Nearest Neighbor and XGBoost [2]. The accuracy of the classifiers was calculated using the confusion matrix. A classifier that the highest accuracy bags can identify as the best classifier.

Random Forest -The RF is created via supervised machine learning. Random forest is a user-friendly and adaptable algorithm. Random forest is a method that, in most cases, produces good results without the use of hyper-parameter tweaking. It creates a forest out of a collection of decision trees.

Decision tree - The DT is a supervised machine learning approach for segmenting data based on specific features. It splits the dataset into smaller and smaller subgroups as it develops the tree. There are two sorts of nodes in the tree: decision nodes and leaf

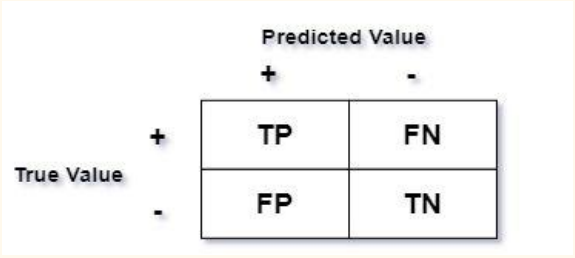
nodes. In a DT, decision nodes indicate outcomes, whereas leaf nodes are where the data is split.

K-Nearest Neighbor(KNN) - K-NN is a supervised learning approach that may be used for both regression and classification, however classification is the most prevalent use. The K-NN technique implies that new and old data are similar, and it assigns new data to the category that is closest to the general categories. During the training phase, the K-NN approach saves the data, and when fresh data is received, it is classified into a class that is highly close to the most recent data.

XGBoost(XGB)- XGBoost is a distributed gradient boosting toolkit that has been tuned for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to create machine learning algorithms. XGBoost is a parallel tree boosting (also known as GBDT, GBM) algorithm that solves a variety of data science issues quickly and accurately.

Evaluation Parameters - For the assessment, we utilized the confusion matrix, accuracy, sensitivity, specificity, f1-score, Matthews correlation coefficient (MCC) score, and Cohen kappa score. The confusion matrix is a table-like structure with true and projected values. In the below we describe confusion matrix and evaluation parameters formulas :-

Confusion Matrix -



$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

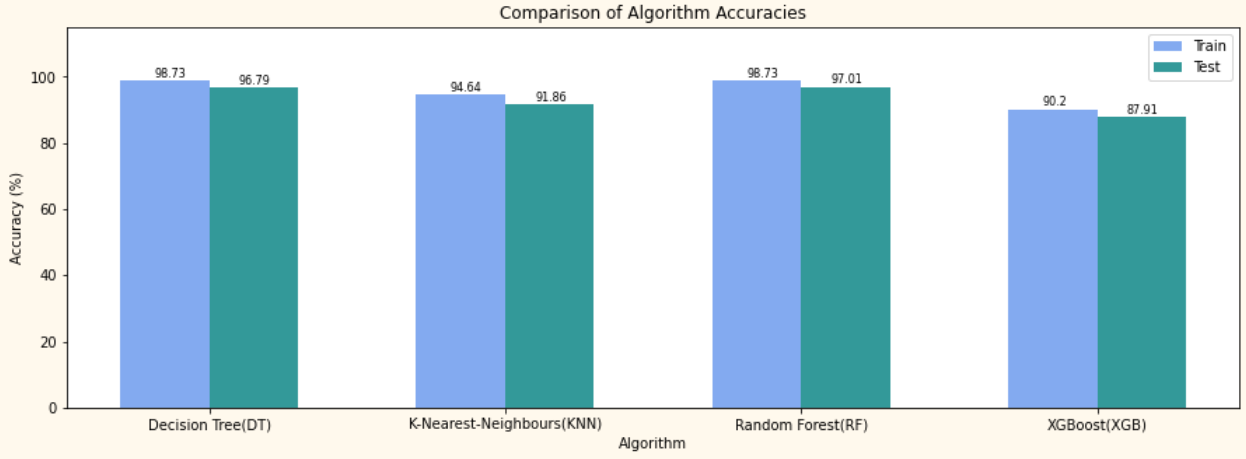
$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

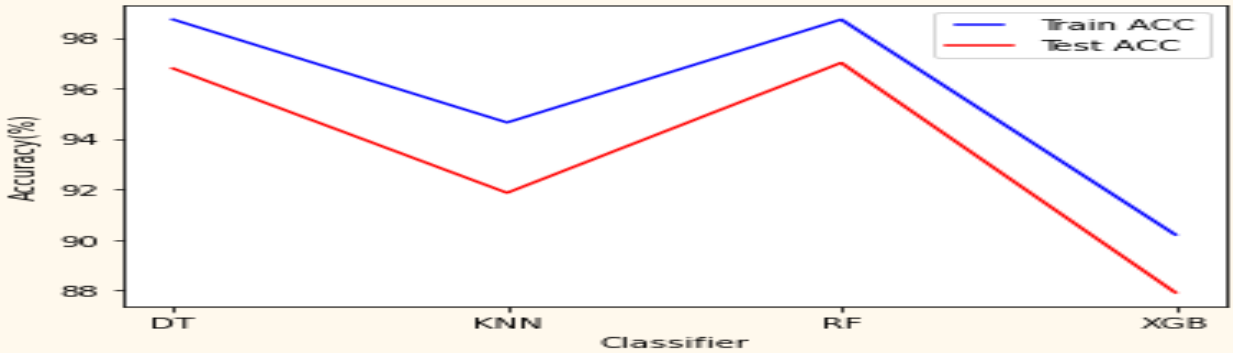
$$\text{F1-Score} = \frac{True\ Positive}{True\ Positive + 1/2(False\ Positive + False\ Negative)}$$

Results and Analysis

For the sake of our research, we divided our dataset 80:20 for training and testing. We can see that our work achieves the best accuracy of 97.01% by using the random forest algorithm. We have used different ML classifiers for our work and the algorithms used are RF,k-NN, DT, and XGB[2]. By the use of Decision Tree, we got an accuracy of 96.79%. When we use the K- Nearest neighbor algorithm, then we got an accuracy of 91.86%. The accuracy of the XGB classifier is 87.91%.

Classifier	Accuracy	Specificity	Sensitivity	F1-Score	MCC	Kappa
DT	96.79	96.79	96.79	96.79	0.97	0.97
KNN	91.86	91.86	91.86	91.86	0.92	0.92
RF	97.01	97.01	97.01	97.01	0.97	0.97
XGB	87.91	87.91	87.91	87.91	0.88	0.88





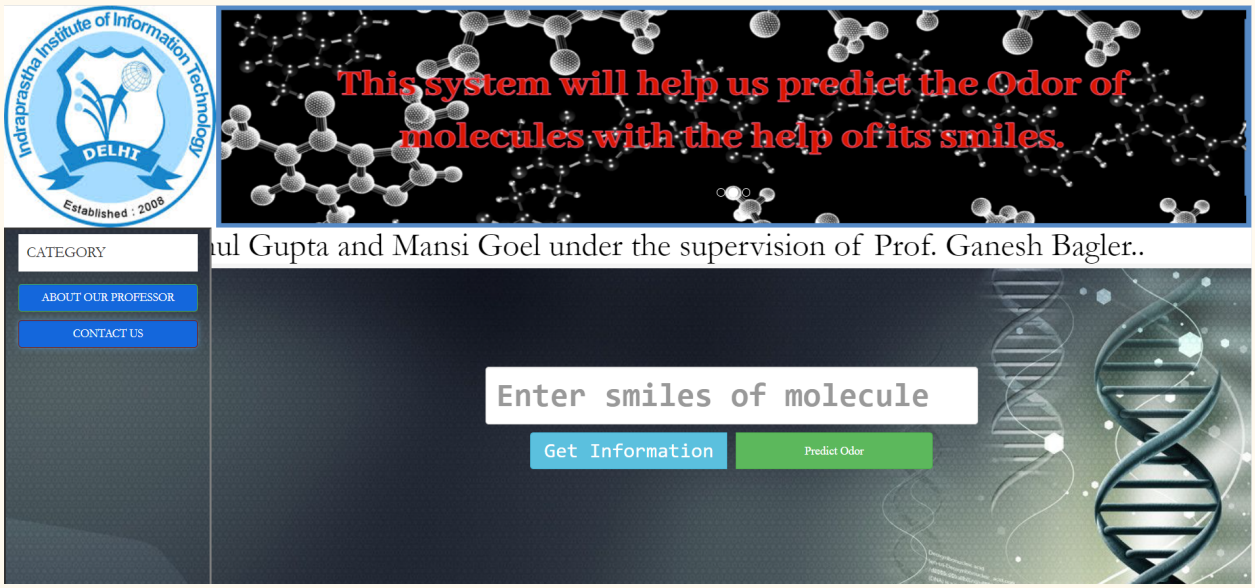
About Odor Prediction WebPage

We have used different tool/technologies for building the webpage :

- HTML for displaying our web page on a web browser.
- CSS used for presentation of our webpage and managing styles etc.
- JavaScript used for creating dynamic and interactive web page.
- Bootstrap for developing responsive websites..

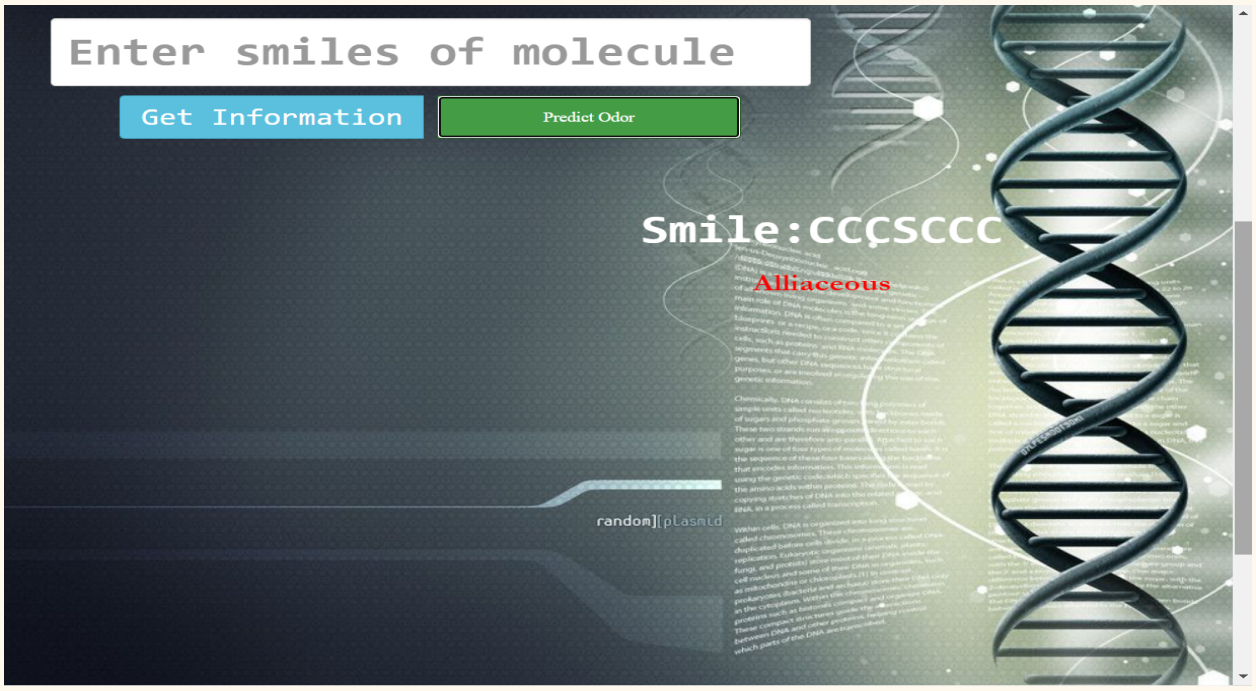
Odor Prediction WebPage

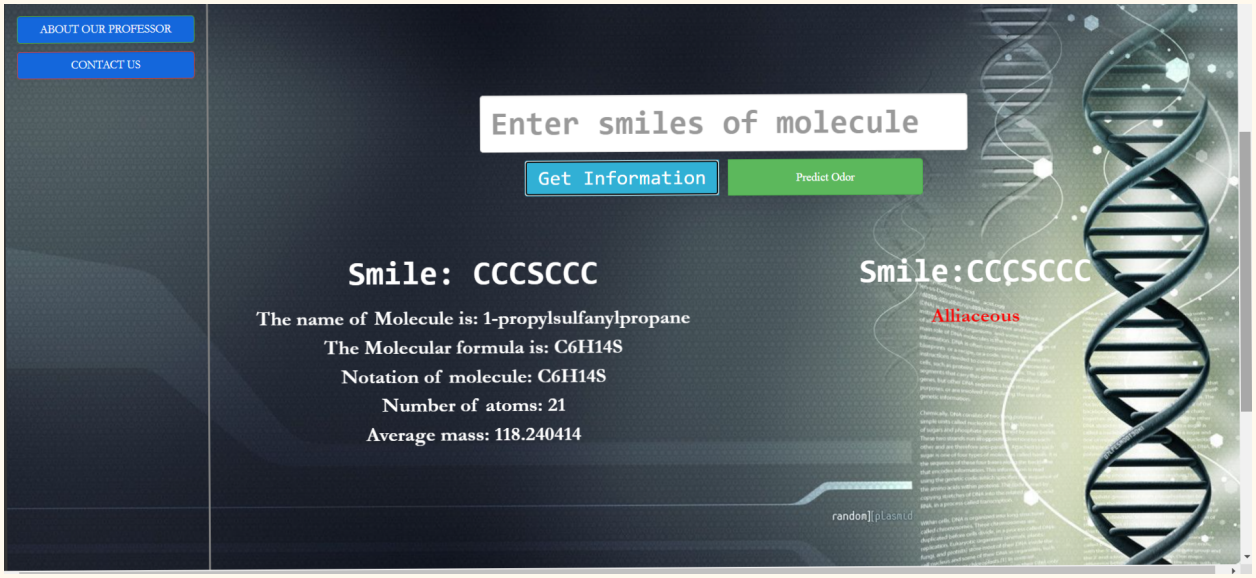
a. Our Front view of webpage



b. Enter the smile of the molecule in the text box to get molecular information of that particular smile by clicking on Get Information Button.

c. To Predict the Odor of a molecule through smile as input, click on the Predict odor button.





References

1) https://olfab.iiita.ac.in/olfactionbase/
2) https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/
3) https://en.wikipedia.org/wiki/Odor
4) Ohloff, G. *et al.* Stereochemistry-odor relationships in enantiomeric ambergris fragrances. *Helv. Chim. Acta* 63, 1932–1946 (1980).
5) Wise, P. M., Olsson, M. J. & Cain, W. S. Quantification of odor quality. *Chem. Senses* 25, 429–443 (2000).
6) Engen, T. Remembering odors and their names. *Am. Sci.* 75, 497–503 (1987).
7) Chastrette, M. *Classification of odors and structure-odor relationships in Olfaction, taste, cognition 100–116* (Cambridge University Press, Cambridge, 2002).
8) Yoshida, M. Studies of psychometric classification of odors (5). *Jpn. Psychol. Res.* 6, 145–154 (1964).
9) Mamlouk, A. M. & Martinetz, T. On the dimensions of the olfactory perception space. *Neurocomputing* 58, 1019–1025 (2004).
10) Sell, C. On the unpredictability of odor. *Angew. Chem. Int. Ed.* 45, 6254–6261 (2006).
11) Bentley, R. The nose as a stereochemist Enantiomers and odor. *Chem. Rev.* 106, 4099–4112 (2006).
12) Keller, A. *et al.* Predicting human olfactory perception from chemical features of odor molecules. *Science* 355, 820–826 (2017).
13) Lötsch J, Kringel D, Hummel T. Machine Learning in Human Olfactory Research. Chem Senses. 2019 Jan 1;44(1):11-22. doi: 10.1093/chemse/bjy067. PMID: 30371751; PMCID: PMC6295796.
14) Chacko, R., Jain, D., Patwardhan, M. *et al.* Data based predictive models for odor perception. *Sci Rep* 10, 17136 (2020). https://doi.org/10.1038/s41598-020-73978-1
15) Nozaki Y, Nakamoto T (2018) Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. PLoS ONE 13(6): e0198475. https://doi.org/10.1371/journal.pone.0198475
16) http://webbook.nist.gov/chemistry/
17) Liang Shang, Chuanjun Liu, Yoichi Tomiura, and Kenshi Hayashi,Analytical Chemistry 2017 *89* (22), 11999-12005, DOI: 10.1021/acs.analchem.7b02389
18) Liang Shang, Chuanjun Liu, Yoichi Tomiura, Kenshi Hayashi, Odorant clustering based on molecular parameter-feature extraction and imaging analysis of olfactory bulb odor maps, Sensors and Actuators B: Chemical, Volume 255, Part 1, 2018, Pages 508-518, ISSN 0925-4005 https://doi.org/10.1016/j.snb.2017.08.024.