1. Describe the research goal

The objective of this study is to develop a nonparametric model that can **recognize various chemical substances with high accuracy by analyzing data collected from a colorimetric sensor array that has been exposed to those substances**. To achieve this, the color change patterns exhibited by dyes in the sensor array upon exposure to various chemicals are collected in the form of high-dimensional vectors, and machine learning algorithms are applied to interpret these complex response patterns effectively. With this approach, the study aims to evaluate the potential of electronic nose technology for chemical detection and examine its real-world applicability across a wide range of fields, including environmental monitoring, industrial safety, and medical diagnostics.

2. Construct a nonparametric model

Before deciding on which models to apply, we should first consider the structure of the dataset and the type of information it holds. The dataset consists of high-dimensional numerical data measuring color variation in a series of several dyes, which react in a complicated way depending on chemical structure and substance properties. Each row corresponds to an experimental sample, where the values of color change of 36 different dyes are quantified in the order of red (R), green (G), and blue (B) channels. Thus, each sample can be represented as a 108-dimensional feature vector. Additionally, one thing to know is, since more than one measurement was taken for each chemical substance, there are several samples with different experiment numbers - for example, HCHO_1, HCHO_2 - which are assumed to be in the same class in this study.

Rather than having a single feature that clearly distinguishes one class from another, this dataset contains patterns that emerge only through combinations and interactions among multiple features. For instance, one chemical may only be identifiable when the R channel value of one dye increases while the G channel value of another dye decreases within a specific range. These types of nonlinear feature interactions are critical for classification. In such a high-dimensional and complex context, linear models - which rely on weighted sums of individual features - may lack the expressive power to capture the underlying decision boundaries. Therefore, models with the capacity to flexibly learn nonlinear relationships and complex decision surfaces are required for effective classification.

Based on the characteristics of the dataset, **K-Nearest Neighbors (KNN)** and **Random Forest** were selected as the two classification models for this study.

KNN classifies each sample by referencing the entire training data, identifying the K nearest neighbors, and assigning the most common class label among them. Since KNN makes predictions by directly relying on distance-based comparisons at inference time, this might be particularly suitable for the current dataset, where multiple observations within the same chemical substances have variability and where clear, consistent division boundaries are hard to define. In such cases, where class separation is challenging to capture using a fixed functional form, KNN's utilization of local patterns allows it to capture complex and irregular boundaries extremely well.

Random Forest is an ensemble method that builds many decision trees and aggregates their predictions to generate final outputs. Each tree is trained on a random subset of data and features, which allows the model to learn various combinations and interactions of features. This might be especially helpful for the dataset like this, where classification depends on complex, nonlinear combinations between multiple features and not on a few dominant variables. In addition, Random Forest provides feature importance scores, giving insight into which dyes or color channels were most predictive of classification performance. This interpretability will enhance both the reliability and practical interpretability of the results.

Lastly, both KNN and Random Forest are easy to implement and have a small number of hyperparameters to optimize compared to other nonparametric models. So for these reasons, KNN and Random Forest were selected as balanced alternatives offering solid predictive performance, interpretability, and practical usability for the objectives of this study.

KNN was applied with K=3, a relatively small value chosen to allow the model to respond sensitively to local structure in the data. In this type of dataset, where some classes have few samples and there is no strong global separation, a small K can work well for capturing localized patterns without over-generalizing. The model classifies each test sample by looking at its three nearest neighbors in the training set and selecting the most common label among them.

Random Forest was configured with 100 decision trees, a well-accepted default and standard which is strong performing but has fair computational costs. By the random subsampling of data and features between trees, Random Forest will be able to capture complex nonlinear feature interactions which are difficult for more basic models to describe.

To compare the performance of the two nonparametric models, both were trained and tested 50 times on a 70/30 train-test split. The accuracy and macro-averaged F1 scores of each run were taken and used as the final performance score. The results are presented below.

Performance of KNN, Random Forest on chemical classification task

Model	Accuracy (avg.)	Macro F1 (avg.)
KNN	0.9835555555555555	0.988978996510885
Random Forest	0.98	0.9809490076732448

This result shows that both models performed really well, with high accuracy above 0.98. KNN was slightly better than Random Forest on both accuracy and macro F1-score, indicating that KNN made strong and consistent predictions for all classes of chemicals, with a good precision-recall trade-off. This implies that its local decision making process was particularly effective at capturing class-level differences within this dataset.

Random Forest also generated stable results with a mean accuracy of 0.98 and macro F1-score of 0.981, confirming its ability to handle high-dimensional, nonlinear data. However, its slightly lower macro F1-score compared to KNN indicates that its class-level performance might have been marginally less balanced.

In summary, both models are appropriate for this classification task, but KNN demonstrated a slight advantage in having uniform performances across different chemical classes.

3. Discuss the advantage and disadvantage of proposed nonparametric model compared to corresponding parametric models

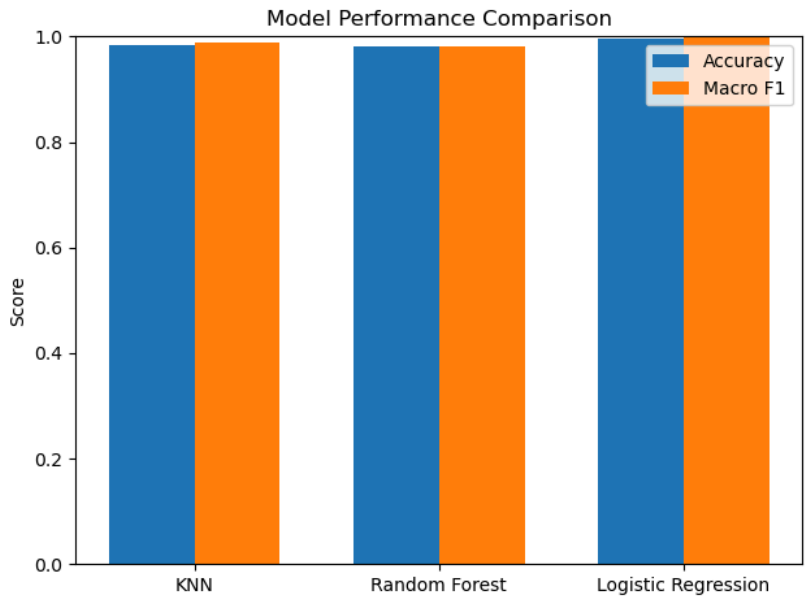
In order to assess the performance of the proposed nonparametric models, the study here compares their performance to a representative parametric model - **Logistic Regression**. Logistic Regression was selected because it is one of the most common and best understood classifiers. It involves a fixed structure model where the linear combination of input features is passed through the sigmoid function to provide an estimate of class probabilities. This makes it a reference point to compare with the more flexible, structure-free nature of nonparametric models like KNN and Random Forest.

The Logistic Regression model was configured with max_iter=5000. The increase bound on iterations is used to handle potential convergence issues that normally occur when using the default value for complex, high-dimensional data.

In theory, Logistic Regression is limited by its linear decision boundary assumption, which restricts its ability to model complex and nonlinear interaction patterns between input features and class labels. This constraint is relevant in the context of colorimetric sensor data, where chemical substance classification depends on subtle and potentially nonlinear interaction between 108 high-dimensional RGB features. In terms of these theoretic considerations, Logistic Regression was expected to be doing worse than the nonparametric models - KNN has no assumption of decision boundary shape and resorts to local proximity in training data, and Random Forest constructs a decision tree ensemble that can approximate highly nonlinear functions through recursive splitting. However, as can be seen from the table below, the results contradicted the above hypothesis and showed an unexpected outcome.

Performance of KNN, Random Forest, and Logistic Regression on chemical classification task

Model	Accuracy (avg.)	Macro F1 (avg.)
KNN	0.9835555555555555	0.988978996510885
Random Forest	0.98	0.9809490076732448
Logistic Regression	0.9964444444444445	0.9972026143790851



Logistic Regression performed the highest accuracy and macro F1-score, outperforming KNN and Random Forest. This was unexpected as aforementioned, given the dataset's high dimensionality and non-linear inter-feature interactions, which seemed more appropriate for models that do not assume linear separability.

There are several reasons why this might be so. First, the feature space itself may have been more linearly separable than initially expected. The RGB based input features, although high-dimensional, may follow a structure that aligns well with linear boundaries - especially if feature and label relationships are largely additive or monotonic. Second, since all chemical classes were equally represented with exactly 7 samples each, there was no class imbalance to introduce learning bias or affect evaluation metrics like macro F1-score. This might allow Logistic Regression to give equal treatment to all classes while optimizing. Finally, it is also possible that the relatively few samples per class simplified the classification problem overall, reducing the complexity that would require more flexible decision boundaries. So, these arguments suggest that under certain conditions, even a linear model can perform better than more complex algorithms, emphasizing the necessity of empirically validating theoretical assumptions.

Anyway, these results demonstrate that nonparametric models are just as good even when parametric models like Logistic Regression perform unexpectedly well. Their flexibility allows them to handle the huge diversity of data structures, especially where there is a high feature interaction or unknown ones.

So, if I were to summarize the strengths and weaknesses of nonparametric models, they would be as follows.

One of the significant advantages of nonparametric models is that they are very flexible - they do not assume any form about a specific functional form and can mold their structure directly from the training data. This enables them to learn complex patterns more effectively and achieve high accuracy in classification problems where the interactions among variables are nonlinear and cannot be readily captured by simpler methods.

Nonparametric models, however, are not without their own limitations. One of the largest drawbacks is their computational complexity and cost. For instance, KNN must calculate distances between a new observation and all training observations for prediction, which can significantly slow down inference time, especially when working with big data. Random Forest, also, must train and maintain a large number of decision trees, which takes a lot of computational power during both training and inference.

Interpretability could also become an issue with these nonparametric models as complexity increases. In KNN, for example, model accuracy can degrade in high-dimensional space due to the “curse of dimensionality” where increased feature dimensions make distances between points become less meaningful - reducing the effectiveness of distance-based classification. Unless there is proper feature selection or dimensionality reduction, KNN might perform poorly in such spaces. In the case of Random Forest, although the single decision trees are understandable, the ensemble itself is less intuitive to understand. It is hard to understand how the model as a whole makes a particular prediction, so it is less transparent and harder to explain than simpler models.