Feature selection details :

In our analysis, we chose different feature selection methods for our models based on their performance during the validation phase. We chose Naive Bayes features using the chi-square method, and we selected the Random Forest features using a random search. However, for a fair comparison between models, we ran three experiments using both sets of features for both models. The final results are shown in the following table.

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
NB+chi-square	**0.61**	**0.60**	**0.66**	**0.63**
NB+random search	0.56	0.55	0.59	0.57
NB +all features	0.53	0.53	0.54	0.53
RF + Chi-square	0.56	0.56	0.56	0.56
RF+random search	**0.64**	**0.64**	**0.61**	**0.63**
RF + all features	0.64	0.67	0.53	0.59

In analyzing the performance of the Naive Bayes model, we compared the results of using the chi-square method for feature selection to two other methods: using all features in the dataset, and using a set of features that were randomly selected for the random forest model. The results on the validation phase showed that the chi-square method led to the best performance metrics of the three methods. Based on this, we made the decision to use the chi-square method for feature selection in our implementation of the Naive Bayes model. Table 1 shows the final results on the test set for each set of features in three experiences. The results show the same pattern as we observed in the validation phase.

Based on table 1, we can see that the features that we select using a random grid search for the random forest model give us better performance metrics in the validation phase. The steps that led us to choose these features are as follow:

- When the Random Forest model used all available features in the dataset initially, the recall score was quite low.
- In an effort to improve performance, we also tried using the chi-square method for feature selection, but the results showed that this did not improve the model's performance.
- As a next step, we employed a random search method to identify the best features for our model [1]. This approach is computationally expensive, but since our dataset is relatively small, containing only nine features, it was possible to implement. The features identified by the random search were pH, hardness, solids, chloramine, and sulfate.
- We then used grid search cross-validation to find the optimal hyperparameters, resulting in the final trained Random Forest model.

Glossary:

- SMOTH

The SMOTE technique is used to deal with datasets that have an imbalanced target label. This approach increases the number of data points in the minority class by creating synthetic data points based on the existing data in the minority class. It does this by clustering blends of the data points that are near each other in the feature space[2].

- Chi-square feature selection:

The chi-square test is a technique used to measure the relationship between two qualitative features. It can be used to determine if a particular variable is useful for predicting the target class. The test calculates a ratio of the relationship between a feature and the target class. The larger this ratio, the more useful the feature is for predicting the target class. By selecting the features with the highest relationship to the target class and training a machine learning model on them, we can improve the model's ability to predict the target class [3].

- Grid Search:

Grid search is a method for hyperparameter tuning, which involves training a model on every possible combination of the list of hyperparameters that have been defined. The main aim is to find the best combination of hyperparameters in order to achieve the best performance metrics for the machine learning model[4].

- Random grid search:

Random grid search is a method for hyperparameter tuning that involves training the model on a randomly chosen combination of predefined hyperparameters. It is less computationally expensive compared to grid search, but may not lead to the best possible combination of hyperparameters[4].

- K-fold cross-validation

K-fold cross-validation is a technique that allows us to see how well a machine learning model generalizes to new data. This method divides the dataset into k folds, trains the model on k-1 folds, and uses the last fold for validation. This process is repeated until all folds have been used as test sets. The average performance metric across all k iterations is used as the performance of the model.

- maximum posterior or MAP

The maximum posterior or MAP decision rule is a decision-making method that involves making a decision based on the highest posterior probability with respect to the available data[5].

- Precision:

Precision is a performance metric for a classification. It is calculated by dividing the number of true positives by the total number of positive predictions made by the machine learning model.

- Recall:

Recall is a performance metric for a classification task. It measures the ability of the classifier to recognize all relevant instances in the dataset. It is calculated by dividing the number of true positive observations by the sum of the false negatives and true positive observations.

- F1

F1 score is a performance metric for a classification task. It is a harmonic mean of precision and recall, giving equal weight to both precision and recall. It is a good metric to use when we want to balance precision and recall.

1. Zheng, A., *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. 2015: O'Reilly Media.
2. Genesis. *SMOTE (Synthetic Minority Oversampling Technique)*. 2018; Available from: https://www.fromthegenesis.com/smote-synthetic-minority-oversampling-technique/.
3. gajawada, s.k. *Chi-Square Test for Feature Selection in Machine learning*. 2019; Available from: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.
4. Gupta, L., *Comparison of Hyperparameter Tuning algorithms: Grid search, Random search, Bayesian optimization.* 2020.
5. hejiawei. *Maximum A Posterior (MAP) Rule*. 2014; Available from: https://jiaweihe.wordpress.com/2014/04/19/maximum-a-posterior-map-rule/#:~:text=In%20this%20sense%2C%20the%20decision%20rule%20is%20optimal,which%20minimizes%20the%20probability%20of%20the%20decision%20error.