Project Report- Group 31

Henrique Catarino-56278 (13 hours); Vasco Maria-56374 (13 hours); Miguel Nunes-56338 (13 hours); Bruno Raimundo-56322 (13 hours)

Introduction and Goals:

We were given an augmented and edited version of the QSAR Biodegradation Dataset, the goal of the project is, with the Dataset, to provide the best classification models with the methods covered in class, test different hyperparameters for this model and identify the dataset's most significant features.

The first part of this project is to process the data by first using several imputation methods to fill up the missing data and, afterwards, use several methods to normalize the dataset. Afterwards, using an evaluation model, we choose the best combination of imputation/normalization methods.

After choosing the best combination we eliminated the less impactful variable using feature selection and dimensionality reduction methods, which we later also decided which one was the best for this project.

Data Processing:

First we separated the "Biodegradable" variable from the main Dataset to a separate variable, and divided the data in 80/20 proportion for testing, being the 80% the training sample and the 20% the testing sample.

With the data separated we used 2 imputation methods to add the missing values, we used Univariate Imputation and KNN Imputation using 4 neighbours, since we tried multiple values and 4 gave us the best results.

Now that we had imputed datasets for each one, we applied 4 normalization methods: Standard Scaler, MinMax Scaler, PowerTransformer and Normalizer.

To choose one, we decided to evaluate these 8 datasets using KFold cross validation since it offers a more solid and reliable estimate of the model's performance, lowering the chance of overfitting and giving a more accurate assessment of its real performance on unknown data.

We used KFold to separate the data from each normalized dataset in 5 equal parts and then use the logistic regression, DecisionTreeClassifier and RandomForestClassifier as classifiers for each dataset. Afterwards we evaluated the accuracy, precision and f1 score for each and decided that the best normalized dataset is the one normalized by the Standard Scaler method and imputed by the KNN Imputation.

Variable Selection:

After selecting the correct normalized dataset we excluded the 5 less impactful variables. The first method we used was feature selection using Pearson's correlation, which consists of calculating Pearson's correlation matrix to calculate the index of the worst 5 variables to be removed.

The second and third was sequential selector with the frontward and backward direction to calculate the top 36 most impactful variables and remove the remaining 5.

Finally, we used principal component analysis to calculate the variance ratio and from there do the same procedures and the other ones.

The remaining step was to identify which of these method was the best one to identify the worst 5 variables. We did this by using the logistic regression, DecisionTreeClassifier and RandomForestClassifier classifiers to measure the one method with the best variance and according to the results we concluded that the best method to use is the PCA method to eliminate the 5 columns.

With all of this done we implemented the imputed method, the normalizer method and the variable selection method to the full dataset and appended again the "Biodegradable" variable from the DataSet.

Model Results:

To discover the best model we, again, split the last section's dataset into training and testing sets, with the "Biodegradable" variable as y and the rest as x and then performed Cross-Validation on the training set with each of the models. We chose, as the best, the one with the highest cross-validation accuracy, the SVM Model.

Model	Cross-Validation Accuracy
SVM	0.9638
KNN	0.9548
Logistic Regression	0.9521
Decision Tree	0.9485
Naive Bayes	0.9417

Hyperparameter Tuning:

For this best model, SVM, we tuned three variables, “kernel” (values: ‘linear’, ‘rbf’), “C” (values: 1, 10, 100, 1000) and “gamma” (values: 1e-1, 1e-3, 1e-5, 1e-7), using a grid search with cross-validation method, which estimated the best combination of these values is the rbf kernel with a C of 10 and a gamma of 1e-1.

Conclusion:

With this project we got a better understanding of the several processes involved in data processing and classification. We not only learned how the methods for the several processes work, such as implementation and normalization, how to apply them and how to choose the best one.

We observed from the test results that the differences between them aren’t a lot. As such any of the other methods, for example for normalization, would be perfectly viable for the project and that different test methods could also give different results as we used previously simple validation before changing to cross validation, and the best normalization method changed with it.

With this we conclude that the best classification model to help predict biodegradability, of the ones we tested, is an SVM with a C of 10 and a gamma of 1e-1.

At the end we also discovered that the dataset’s most significant features are nH, NssssC, nCb, nCp and F03.