*sUPERVISED machine*

*Learning*

*[Subtítulo del documento]*

*JAIME CASTRO CERNADAS*

*Machine Learning | MSc. In Computational Biology*

Índice

1. Preprocessing steps
2. Hyperparameter tuning
3. Results
4. Regression and Decission tree interpretation
5. Conclussions

# **Preprocessing**

In this dataset, the only preprocessing step applied was the removing of the “Unnamed: 0 “ feature. This feature was a row index, and it was eliminated to further clean the data and eliminate any possible interference with this unnecessary data.

No other preprocessing methods were applied, as the data had no missing values, no duplicates and no need for scaling or normalizing.

# **Pipeline**

A pipeline was created for the model selection, training-test prediction and confusion matrix generation.

## Hyperparameter tunning and training evaluation

GridSearchCV was used to identify the best parameter combination for the models. **GridSearchCV is a systematic approach to hyperparameter tuning that evaluates all possible combinations of specified hyperparameter values to determine the combination that optimizes a performance metric. It works by splitting the training data into k-folds, training the model on k-1 folds, and validating it on the remaining fold for each combination of hyperparameters. The process is repeated for all folds, and the average validation score is computed for each hyperparameter combination. The combination with the highest score is then selected as the best set of hyperparameters** (Pedregosa et al., 2011).

Due to the small-sized dataset, the Leave-One-Out Cross-Validation (LOOCV) method was chosen. **LOOCV is a cross-validation technique where each data point in the training set is used once as a validation set while the remaining points form the training set.** Since LOOCV uses each data point as a validation set exactly once, it avoids overfitting to any specific subset of the training data. This is especially important for small datasets, where traditional train-validation splits may result in an insufficient number of validation samples. Additionally, LOOCV provides a more comprehensive understanding of how the model performs across all data points, ensuring that the evaluation reflects the entire dataset's characteristics. Arlot & Celisse, 2010).

The parameter combination with the highest accuracy score was selected. **Selecting the combination with the highest accuracy is a reasonable choice because accuracy is a straightforward and widely used metric that reflects the proportion of correctly classified instances. For classification tasks where the classes are balanced, accuracy is an effective metric to assess model performance. However, for imbalanced datasets, additional metrics like precision, recall, or F1-score may also need to be considered to provide a more nuanced evaluation** (Sokolova & Lapalme, 2009).

After selection, the best models were evaluated with the training data by LOOVC crossvalidation. (Explain why) The accuracy score and a classification report was obtained (explain the metrics of the classification report)

## Test evaluation

The models were evaluated with a set of unseen data. **This is done to assess how well the model performs on data it has never encountered before. This ensures that the model has not overfitted to the training data and provides a more realistic measure of its effectiveness. Unseen data evaluation is a critical step in validating the robustness of the model and ensuring its reliability for practical applications**(Goodfellow et al., 2016).

A confusion matrix was plotted to support the results from the test data prediction. **A confusion matrix is a visualization tool that summarizes the performance of a classification model by displaying the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. It helps to understand not just the overall accuracy but also the distribution of errors among the classes.**

# **Results**

## Logistic Regression

**Logistic Regression** is a statistical and machine learning algorithm used for binary and multiclass classification tasks. It models the relationship between one or more independent variables (features) and a dependent variable (target) using a logistic function (also known as the sigmoid function) to map predicted values to probabilities. Logistic regression is widely used in scenarios where the goal is to predict the probability of a certain event, such as spam detection, disease diagnosis, or customer churn.

## Decision Tree

**Decision Tree** is a supervised machine learning algorithm used for both classification and regression tasks. It works by splitting the data into subsets based on feature values, forming a tree-like structure. Each internal node of the tree represents a decision based on a feature, each branch corresponds to the outcome of the decision, and each leaf node represents a final class or value (in regression tasks).

## Random Forest

**Random Forest** is an ensemble learning algorithm primarily used for classification and regression tasks. It builds multiple decision trees during training and aggregates their predictions to improve accuracy, robustness, and reduce the risk of overfitting compared to a single decision tree.

## K Nearest Neighbour

**K-Nearest Neighbors (KNN)** is a simple, non-parametric, and instance-based supervised learning algorithm used for classification and regression tasks. KNN predicts the output for a given query instance by considering the kk-closest training samples (neighbors) in the feature space.

## Multi-layer Perceptron

**Multilayer Perceptron (MLP) Classifier** is a type of feedforward artificial neural network that is widely used for classification tasks. It consists of an input layer, one or more hidden layers, and an output layer, where each layer is composed of interconnected nodes (neurons). MLP is trained using supervised learning and is capable of learning complex, non-linear relationships in the data.