|  |
| --- |
| Dorset college |
| AI Mini Project |
| Report |

|  |
| --- |
| MULLER Julie  <https://github.com/JulieMuller/AI-mini-project> |

Table des matières

[Introduction 2](#_Toc151740230)

[SVM 2](#_Toc151740231)

[Improving the model 2](#_Toc151740232)

[Cleaning the data 2](#_Toc151740233)

[Imputation 2](#_Toc151740234)

[Outlier detection 2](#_Toc151740235)

[Feature selection 2](#_Toc151740236)

[Using random forest 3](#_Toc151740237)

[Using correlation coefficient 3](#_Toc151740238)

[Conclusion 3](#_Toc151740239)

[Hyperparameters 3](#_Toc151740240)

[Final model training 4](#_Toc151740241)

[Introduction 5](#_Toc151740242)

[Random forest 5](#_Toc151740243)

[Logistic regression 5](#_Toc151740244)

[Improving the model 5](#_Toc151740245)

[Cleaning the data 5](#_Toc151740246)

[One hot encoding 6](#_Toc151740247)

[Scaling the data 6](#_Toc151740248)

[Final model training 6](#_Toc151740249)

Iris

# Introduction

The goal is to classify images of flowers based on their features such as the petal length, petal width, sepal length, and sepal width. To do that, we are going to use classification learning. We now have the choice between 2 algorithms: decision tree algorithm or Support Vector Machines (SVM) algorithm. In our case, we will try both to see if one of them is more fitting to our problem.

# SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. SVM aims to find a hyperplane that best separates data points into different classes while maximizing the margin, the distance between the hyperplane and the nearest data points of each class.

They can handle non-linear relationships by mapping the input data into a higher-dimensional space using kernel functions, allowing it to find complex decision boundaries.

The critical data points that define the decision boundary are called support vectors, and SVM relies on them to determine the optimal hyperplane for classification.

# Improving the model

## Cleaning the data

Cleaning the data is very important to limit the errors and inconsistencies. Quality data reduces noise and enables the algorithm to focus on relevant patterns and relationships. Ultimately leading to more precise predictions.

### Imputation

First action was to remove nan or null values.

### Outlier detection

An outlier data is an observation in a dataset that significantly deviates from the overall pattern, potentially indicating errors, anomalies, or unique phenomena. Removing those observations prevents skewed learning, ensuring more accurate and representative generalization on unseen data.

## Feature selection

The goal is to identify and choose the most relevant botanical characteristics from our dataset (sepal\_length, sepal\_width, petal\_length, petal\_width, species) to improve model efficiency of the model.

|  |  |
| --- | --- |
| Une image contenant texte, capture d’écran, diagramme, ligne  Description générée automatiquement |  |
| Figure 1.1 | Figure 1.2 |

### Using random forest

One way of doing this is to use random forest feature selection. We first create a random forest classifier, then we train it and use the feature\_importances\_ property that outputs the impurity-based feature importances. The higher the value, the more important the feature. According to figure 1.1, the 2 most relevant parameters are petal\_lenght and sepal\_width.

### Using correlation coefficient

The correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where:

* 1 indicates a perfect positive linear relationship.
* 0 indicates no linear relationship between the variables.
* -1 indicates a perfect negative linear relationship.

We can see on figure 1.2 that we have a strong positive linear relationship between petal\_lenght and petal\_width and a strong negative linear relationship between petal\_lenght and sepal\_width.

### Conclusion

We can see the conclusion of both methods are intersecting thus we will choose as our 2 parameters petal\_lenght and petal\_width.

### Hyperparameters

Hyperparameter tuning can be performed using techniques like Grid Search. The GridSearchCV function in Python works by exhaustively searching a predefined hyperparameter grid to find the combination of hyperparameter values that maximizes or minimizes a specified scoring metric through cross-validation.

## Final model training

|  |  |
| --- | --- |
|  |  |

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

The most accurate kernel to use for our SVM algorithm on our dataset is a linear kernel.

Churn predictions

# Introduction

We want to predict whether a customer will churn (cancel) their subscription to a service based on their usage patterns, demographics, and other features of the dataset.

To get an optimal result, we will try 3 different machine learning algorithms: SVM, random forest and logistic regression.

## Random forest

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputs the mean prediction (regression) of the individual trees.

Random Forest randomly selects subsets of the training data through bootstrapped sampling, creating multiple diverse subsets for training individual decision trees.

At each node of the decision tree, a random subset of features is considered for splitting, introducing randomness and reducing correlation between trees.

The predictions from each decision tree are aggregated to produce the final prediction. This ensemble approach improves robustness, generalization, and reduces overfitting compared to a single decision tree.

## Logistic regression

Logistic Regression works by modeling the probability of a binary outcome using the logistic function, which transforms a linear combination of input features into values between 0 and 1.

It estimates coefficients through optimization techniques such as gradient descent to maximize the likelihood of the observed outcomes.

The predicted probability is then thresholded, commonly at 0.5, to classify instances into one of the two classes, making Logistic Regression suitable for binary classification tasks.

# Improving the model

## Cleaning the data

Cleaning the data is very important to limit the errors and inconsistencies. Quality data reduces noise and enables the algorithm to focus on relevant patterns and relationships. Ultimately leading to more precise predictions. (Just as for the Iris dataset, we check for outliers and null values)

### One hot encoding

One hot encoding is a technique in which categorical variables are converted into a binary matrix, representing each category as a unique binary column, with only one column having a value of 1 for each observation.

One hot encoding is essential with logistic regression as it transforms categorical variables into a binary format, allowing the algorithm to effectively handle the data.

### Scaling the data

Scaling is crucial before training a machine learning model as it ensures that features with different scales contribute equally to the model, preventing larger magnitude features from dominating the learning process. Scaling also helps gradient-based optimization algorithms converge faster and improves the model's sensitivity to variations in all input features, enhancing overall performance.

## Final model training

Once again, we use the feature\_importances\_ property of a random forest classifier trained on our data.

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

|  |  |
| --- | --- |
| Une image contenant texte, capture d’écran, Rectangle, diagramme  Description générée automatiquement | We can see that the algorithm with the highest accuracy is the random forest.  It would also be interesting to look at what model takes the longer to train. |