



IA solutions

Document Name :	IA solutions
Version :	1.0
Version Date :	15/07/2022
Modified By :	Jimmy MARTEL
Creation Date :	04/07/2022
Created by :	Marius LUPORSI
Approved by :	Marius LUPORSI
Confidentiality Level :	INTERNAL

[Introduction](#)

[Humans ressources](#)

[Organisation](#)

[Tools](#)

[Data](#)

[Algorithm](#)

[Metrics](#)

[Conclusion](#)

Introduction

This document purpose is to delivering some generic off-the-shelf solutions for the IA part of our project.

Humans ressources

Name	Role
Marius LUPORSI	IA developper
Jimmy MARTEL	IA developper

Organisation

- Discord for fast communication
- Jira for sprint organisation
- Notion for documentation and organisation
- Github for host or code
- Agile method

Tools

Name	Description	Type	Price
Jupiter	Notebook	Software	Free
Python	Language	Language	Free

Data

Currently, we don't really have information on the exact data we will have available.

However, there is a process that we know we will be able to follow:

- Load the data
- Data visualization
- Preprocessing
 - Missing values, categorical variables etc ..
- Feature Engineering
 - The goal of feature engineering is simply to make your data better suited to the problem at hand.
 - You might perform feature engineering to:
 - improve a model's predictive performance
 - reduce computational or data needs
 - improve interpretability of the results

Algorithm

Our problem seems to correspond to a supervised machine learning problem, more particularly classification problem.

Here is a list of the different algorithms we will be able to try:

- Random Forest
- Neural network
- KNN
- XGBoost
- SVM

Then we will determine which one is the most efficient (see Metrics) for our problem and therefore which one we will use.

Metrics

- **Accuracy**

The simplest indicator is accuracy: it indicates the percentage of good predictions. It's a very good indicator because it's very simple to understand. The higher it is, the more accurate our model is.

$$\text{Accuracy} = \frac{\text{Vrai positif} + \text{Vrai négatif}}{\text{Total}}$$

- **Confusion matrix**

Matrix confusion is a cross table between real values and predictions. This matrix identifies 4 categories of results:

- The right predictions:
 - True positives: customers who have terminated for whom the score predicted they would terminate
 - True negatives: customers who are still subscribed and for whom the algorithm correctly predicted that they would remain subscribed
- False predictions:
 - False negatives: customers who have terminated but for whom the score wrongly predicted that they would remain subscribers

- False positives: customers who stayed subscribed when the score incorrectly predicted they would terminate


Données réelles		Données prédites par l'algorithme	
		 Résilié prédit	 Abonné prédit
	 Résilié  Abonné	Vrai positif	Faux négatif
		Faux positif	Vrai négatif

Matrice de confusion

- Classification report**

This is one of the evaluation of the classification models. It displays your model's accuracy, recall, F1 score and support. It provides a better understanding of the overall performance of our model.

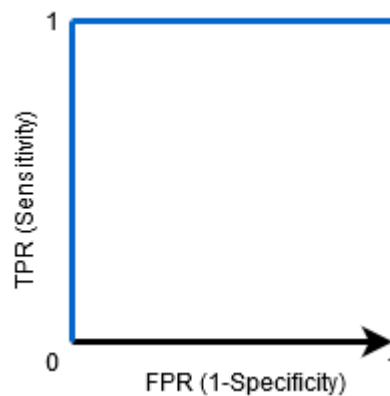
Classification report (1)

 Metrics	Aa Definition
Precision	<u>Precision is defined as the relationship between true positives and the sum of true and false positives.</u>
Recall	<u>Recall is defined as the relationship between true positives and the sum of true positives and false negatives.</u>
F1 Score	<u>F1 is the weighted harmonic mean of precision and recall. The closer the F1 score is to 1.0, the better the expected performance of the model.</u>
Support	<u>Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance review process.</u>

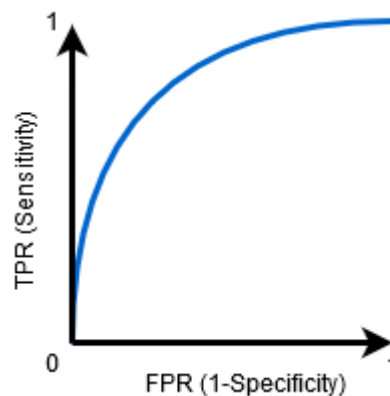
- Area under curve**

The area under the curve (AUC) is the measure of a classifier's ability to distinguish classes and serves as a summary of the ROC curve.

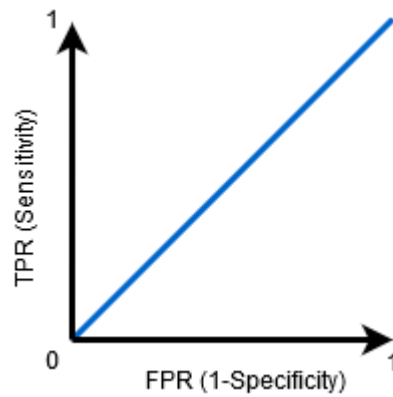
The higher the AUC, the more effective the model is in distinguishing positive from negative classes.



When $AUC = 1$, then the classifier is able to distinguish all Positive and Negative class points correctly. If the AUC had been 0, then the classifier would predict all negatives as positive, and all positives as negative.



When $0.5 < AUC < 1$, there is a good chance that the classifier can distinguish positive class values from negative class values. All this because the classifier is able to detect more numbers of true positives and negatives than false negatives and false positives.



When $AUC=0.5$, the classifier is not able to distinguish between positive and negative class points. This means that the classifier predicts a random class or a constant class for all data points.

Thus, the higher the AUC value for a classifier, the greater its ability to distinguish between positive and negative classes.

Conclusion

After different try, we create a model that predict health of grapes in function of pictures of them.

We have a success of 86% of the predictions (accuracy).