Université SAAD DAHLEB – Blida 1

Faculté des Sciences

Département d'informatique

Mémoire présente pour l'obtention du master

Spécialité : Ingénierie de logiciel

Thème

Etude et Prototypage d'une Machine Learning pour la catégorisation des réclamations clients dans l'application RIGHTNOW by BRENCO

PAR

KHETTAR Hanane Soumia

HAOUCHINE Silia

Devant un jury composé de :

- GUESSOUM Dalila — Promoteur
- BOUDJEMA Sofiane — Encadreur
- CHERIGUENE Soraya — Président
- HAMOUDA Mohamed — Examinateur

Année académique 2021/2022

Remerciement

Nos premiers et plus importants remerciements vont à Dieu Tout-Puissant pour nous avoir accordé la santé, la patience et les connaissances nécessaires pour compléter ce mémoire.

Deuxièmement, nous tenons à remercier Mme GUESSOUM Dalila et M. BOUDJEMA Sofiane pour les conseils et les encouragements qu'ils nous ont prodigués tout au long de notre parcours d'étudiants et de stagiaires.

Enfin, nous réservons un merci spécial à nos merveilleuses familles :

- Nos parents qui nous ont apporté leur amour et leur soutien inconditionnel tout au long de la vie, que nous ayons réussi ou perdu, ainsi que pour leur patience, leurs encouragements et leurs prières. Nous espérons que tout ce que nous avons fait vous rend heureux et en bonne santé.
- Nos merveilleux frères et sœurs pour leur amour, leurs encouragements et leur soutien émotionnel et moral.
- Tous nos professeurs que ce soit du Primaires, du moyen, du secondaire ou de L’enseignement supérieur.

Dédicace

To my dear Father, Mohamed Tahar, who left us too soon,

To my dear Mother, Malika, may God grant her a long life,

To my sisters, To my brother and the whole family one by one, young and old,

To my friends,

To my deceased teacher BENMAHIEDDINE El-Aid,

To everyone who helped me from near or far,

I dedicate this work to you.

KHETTAR Hanane Soumia

Dédicace

Je dédie ce modeste travail à :

Mes chers parents Nacer et Ratiba ·Aucun hommage ne pourrait être à la hauteur de l'amour Dont ils ne cessent de me combler· Que dieu leur procure bonne santé et longue vie·

Celui que j'aime beaucoup , qui a été toujours à mes côtés et qui m'a soutenu tout au long de ce projet : mon Mari Rafik Que dieu te protège pour moi ·

Mes beaux-parents : Mourad et Lila qui m'ont encouragé au cours de ce mémoire· Que dieu les préserve·

Ma chère grand-mère Que j'aime beaucoup que dieu te protège pour nous ·

Mes Chères Sœurs : Wassila ,Chahinez et ma petite Maroua · Pour leurs soutiens moraux et leurs soutien précieux tout au long de mes Études·

Mes chères belle sœurs : Sarah et Assia pour leur soutiens ·

Mes chers Beaux-frères : Omar , Akram et Nabil ·

Mes petites nièces et petits neveux : Mounira , Taline , Marame ,lina , Adem,Mohamed,Iyad et Anis ·

Mon binôme Hanane, je te souhaite beaucoup de sucées, mes aimables amis et sœurs de cœur : Yassmine , Nawel , Roumaissa , Selssabile, Maroua ,Wafia , Naila ,Manel, Imene, Ichrak ,Baya , Je vous aime ·

Toutes mes cousins, mes oncles et mes tantes et toutes les personnes de ma grande famille·

Sans oublier tous les professeurs que ce soit du Primaires, du moyen, du secondaire ou de L'enseignement supérieur·

Et surtout à moi-même HAOUCHINE Silia·

HAOUCHINE Silia

Résumé

Les réclamations des clients sont l'une des paramètres les plus critiques qui déterminent la dynamique du marché du développement de produits. Dans ce sens, l'analyse des réclamations liées aux produits aide les vendeurs à identifier les caractéristiques de qualité et l'orientation client. De nombreuses études ont été menées sur la conception de systèmes de Machine Learning (ML) pour traiter les causes d'insatisfaction des clients.

La société SETRAM - Société d'exploitation des tramways, chargée de l'exploitation et de la maintenance des tramways algériens - un des clients de RightNow By Brenco, veut trouver une solution pour son problème: mauvais choix des catégories de réclamations envoyées par ses clients à travers les formulaires rend ses statistiques inexactes, et elle se trouve obligée de recatégoriser manuellement ces réclamations afin d'obtenir des statistiques précises et ça prend beaucoup de temps et d'effort.

Ce sujet vise à construire un modèle de Machine Learning qui donne une prédiction correcte que possible de catégories de réclamations des clients.

En conséquence, nous avons utilisé l'apprentissage supervisé par classification multi classe à l'aide du SVM et l'approche OVO (One Vs One). Nous avons obtenu un « Accuracy Score» à 0.834 % qui est une valeur très satisfaisante, par contre les résultats de « Précision », « Recall », « F1 Score », « Specificiy » et « Accuracy» sont tout entre 70% et 98% et c'est un très bon signe.

Selon ces résultats obtenus, nous pouvons dire que cette étude semble prometteuse pour de futures études sur les systèmes de catégorisation des réclamations.

Nous avons également développé une simple application web afin de tester la prédiction du modèle construit.

Mots clés : Réclamation, Catégorisation, Machine Learning, Classification multi classe, Modèle, TF IDF, SVM, OVO, RBF.

Abstract

Customer complaints are one of the most critical parameters that determine the dynamics of the product development market. In this sense, the analysis of complaints related to products helps sellers to identify quality characteristics and customer orientation. Many studies have been conducted on the design of Machine Learning (ML) systems to address the causes of customer dissatisfaction.

The company SETRAM - Société d'exploitation des tramways in charge of the operation and maintenance of Algerian trams - one of the customers of RightNow By Brenco, wants to find a solution for its problem: bad choice of the categories of complaints sent by its customers to through forms makes her stats inaccurate, and she finds herself having to manually re-categorize these claims in order to get accurate stats and that takes a lot of time and effort.

This topic aims to build a Machine Learning model that gives a correct prediction as possible of categories of customer complaints.

As a result, we used supervised learning by multi-class classification using the SVM and the OVO (One Vs One) approach. We obtained an "Accuracy Score" of 0.834 % which is a very satisfying value, on the other hand the results of "Precision", "Recall ", "F1 Score", "Specificity" and "Accuracy" are all between 70% and 98% and it is a very good sign..

According to these obtained results, we can say that this study seems promising for future studies on complaint categorization systems.

We have also developed a simple web application to test the prediction of the built model.

Key words: Complaint, Categorization, Machine Learning, Multi-class classification, Model, TF IDF, SVM, OVO, RBF.

الملخص

تعد شكاوى العملاء من أهم العوامل التي تحدد ديناميكيات سوق تطوير المنتجات. وبهذا المعنى ، فإن تحليل الشكاوى المتعلقة بالمنتجات يساعد البائعين على تحديد خصائص الجودة وتوجه العملاء. تم إجراء العديد من الدراسات حول تصميم أنظمة التعلم الآلي لمعالجة أسباب عدم رضا العملاء.

شركة -SETRAM - Société d'exploitation des tramways المسؤولة عن تشغيل وصيانة الترام الجزائري - أحد عملاء RightNow By Brenco ، تريد إيجاد حل لمشكلتها: الاختيار السيئ لفئات الشكاوى المرسلة من قبل عملائها من خلال النماذج تجعل إحصائياتها غير دقيقة ، وتجد نفسها مضطرة إلى إعادة تصنيف هذه الادعاءات يدويًا من أجل الحصول على إحصائيات دقيقة وهذا يستغرق الكثير من الوقت والجهد.

يهدف هذا الموضوع إلى بناء نموذج التعلم الآلي الذي يعطي تنبؤًا صحيحًا قدر الإمكان لفئات شكاوى العملاء.

نتيجة لذلك ، استخدمنا التعلم الخاضع للإشراف من خلال التصنيف متعدد الفئات باستخدام طريقة دعم ألات المتجهات و طريقة واحد مقابل واحد . تحصلنا على قيمة "Accuracy Score" و قد بلغت 0.834٪ وهي قيمة مرضية للغاية، ومن ناحية أخرى نتائج "Accauracy" ، "Précision" ، " Recall " و" Specificity "و" F1 Score " تتراوح بين 70٪ و 98٪ وهذا مؤشر جيد جدا.

وفقًا لهذه النتائج التي تم الحصول عليها ، يمكننا القول أن هذه الدراسة تبدو واعدة للدراسات المستقبلية حول أنظمة تصنيف الشكاوى.

لقد قمنا أيضًا بتطوير تطبيق ويب بسيط لاختبار التنبؤ بالنموذج المبني.

الكلمات المفتاحية: شكوى ، تصنيف ، تعلم الآلة ، تصنيف متعدد الفئات ، نموذج، تردد المصطلح - معكوس تردد الوثيقة، دعم آلات المتجهات، واحد مقابل واحد، وظيفة الأساس الشعاعي.

Sommaires

Liste des figures

Liste des tableaux

Liste d’acronymes

ML: Machine Learning

IA: Intelligence Artificielle

TF IDF: Term Frequency-Inverse Document Frequency

SVM: Support Vector Machines

RBF: Radial basis function

OVO: One Vs One

OVA: One Vs All

NLP: Natural Language Processing

TN: True Negative

TP: True Positive

FN: False Negative

FP: False Positive

VS: Visual Studio

HTML: Hypertext Markup Language

Introduction Général

1. Contexte générale et problématique:

RightNow By Brenco est l'une des solutions de la société Brenco Engineering & Consulting, elle est une application professionnelle mobile et web en marque blanche basée sur le cloud qui permet de créer des formulaires selon la structure data à travers des champs riches et variés (Texte, Image, Audio ...), récupérer la data avec le mode Hors-Ligne à tout moment et sur n'importe quel appareil. Aussi, elle permet de suivre et analyser les progrès du processus global, d'établir des stratégies basées sur les données et de prendre des décisions précises.

Un des clients de RightNow – L'entreprise SETRAM – a rencontré un problème qui a affecté ses statistiques et la prise de décision.

Ce problème s'agit de mauvais choix des catégories de réclamations envoyées par les clients de SETRAM à travers les formulaires, par exemple le client catégorise sa réclamation à propos de « ne pas porter le masque » comme un problème de respect mais SETRAM la considère un problème de comportement, ou certains clients la catégorisent comme un problème de respect et d'autre comme un problème de comportement, ce qui rend les statistiques de l'entreprise inexacte.

Donc, SETRAM se trouve obligée de recatégoriser manuellement ces réclamations afin d'obtenir des statistiques précises, mais cela prend beaucoup de temps et d'effort.

2. Objectif:

Ce travail vise à construire un modèle de Machine Learning qui donne une prédiction correcte que possible de catégories de réclamations des clients.

En conséquence, les réclamations doivent être classées en catégories suivantes :

- Gestion : S'il y a une erreur administrative.
- Technique : Lorsqu'une erreur technique survient à la station.
- Comportement : S'il y a un problème avec le comportement des agents.
- Respect : En cas de non-respect des règles de l'entreprise soit par les agents ou les clients.
- Sécurité : S'il y a quelques choses qui met les clients en danger.
- Métier : Lorsqu'il y a un problème dans la profession des agents en termes de qualité et de fabrication.

3. Structure du travail:

Ce travail est structuré de la manière suivante :

- Chapitre 01: Machine Learning

  Il s'agit d'aborder les définitions d' Intelligence Artificielle (IA) , de Machine Learning (ML) et leurs différentes méthodes et modèles en détail, aussi son importance, avantages et inconvénients.

- Chapitre 02: Spécification de l'approche

  Au cours de ce chapitre, nous présenterons en détail l'application RightNow By Brenco, ses objectifs et les étapes que nous suivrons pour construire le modèle.

- Chapitre 03: Implémentation

  Dans lequel nous présenterons les outils et les langages utilisés, nous discuterons les résultats obtenus, nous évaluerons la performance du modèle, puis nous parlerons sur un travail similaire au notre dans la dernière section.

Chapitre 01 : Machine Learning

1. Introduction

Machine Learning (ML) est un domaine de recherche à l'intersection de l'intelligence artificielle (IA), de mathématique et de l'informatique (Figure 1). Pendant ces dernier temps, l'utilisation des méthodes de ML est devenue largement répandue dans la vie quotidienne. Qu'il soit question des plats à commander ou les produits à acheter, de la reconnaissance de nos amis sur nos photos, de nombreux sites Web et appareils modernes sont dotés d'algorithmes de ML.**[1]**

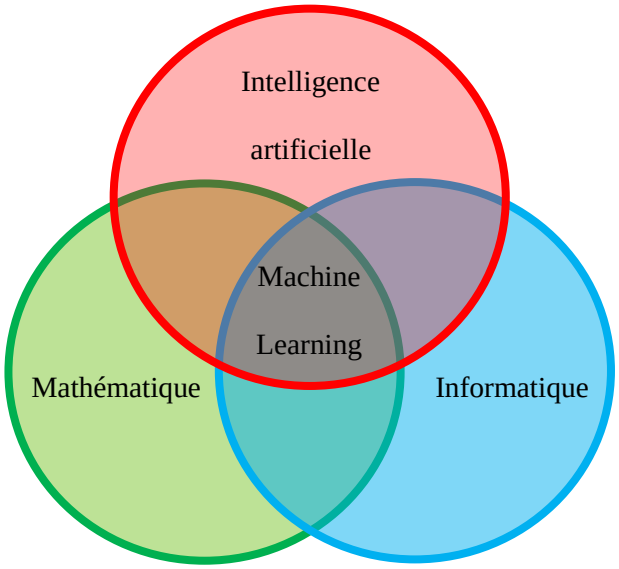


Figure 1. Machine Learning

Dans ce chapitre, nous allons aborder les définitions de l'intelligence artificielle (IL) et de Machine Learning, leurs différentes méthodes et modèles en détail (équations mathématiques et statistiques, algorithmes…etc.), puis nous continuerons avec l'importance, les avantages et les inconvénients de ML.

2. L'Intelligence artificielle

L'Intelligence Artificielle est une science existe depuis 30 ans, son but est de reconstruire par des moyens artificiels des raisonnements et des actions intelligents.

L'IA est une science expérimentale :

- Expériences sur des ordinateurs permettant de tester et de raffiner les modèles exprimés dans les programmes sur de nombreux exemples.
- Observations de l'être humain afin de découvrir ces modèles et de mieux comprendre comment fonctionne l'intelligence humaine. **[2]**

Par ailleurs, les principaux domaines fonctionnels de l'IA témoignent de son immense potentiel, des innombrables réalisations auxquelles elle est susceptible de donner lieu. **[3]**

3. Application d'IA

Nous trouvons l'IA dans différents domaines par exemple :

- **La reconnaissance des formes :**

  La reconnaissance d'images est des domaines du codage et du traitement des signaux, ils sont connus sous le nom de reconnaissance des formes, il s'agit d'un passage obligatoire pour un système autonome afin du traitement de l'information.

- **Les mathématiques et la démonstration automatique de théorèmes :**

  Les premiers domaines où les chercheurs ont travaillé furent les mathématiques qui font figure de bastion de l'intelligence humaine, sont d'excellents sujets tests, les premiers programmes en démonstration automatique travaillaient d'abord sur des théories limitées.

  Les jeux:

  Les jeux furent et sont encore un sujet de prédilection en IA. Les difficultés rencontrées à propos des jeux eurent été les mêmes qu'en mathématiques et que dans beaucoup d'autres domaines. Elles ont toujours trait à la masse considérable de connaissances que l'homme a su assembler à travers les âges.

- **La résolution de problèms:**

  Il s'agit de poser, d'analyser et de représenter des situations concrètes, tout autant que de les résoudre. C'est la capacité d'invention, qu'il s'agisse de problèmes de la vie de tous les jours, de recherche opérationnelle ou de mathématiques.

- **Compréhension du language naturel:**

  Dans cette partie, les chercheurs s'intéressent à l'analyse et à la génération des textes, à leur représentation interne et à la mise à jour des connaissances nécessaires à leur compréhension : connaissances syntaxiques, sémantiques, pragmatiques. **[2]**

Il s'agit d'approfondir des domaines fonctionnels déjà connus, comme des systèmes experts, la planification, l'optimisation ou de la robotique, et aussi de développer de nouveaux domaines tels que le Machine Learning, le Natural Language processing ou encore le speech (Figure 2). **[3]**

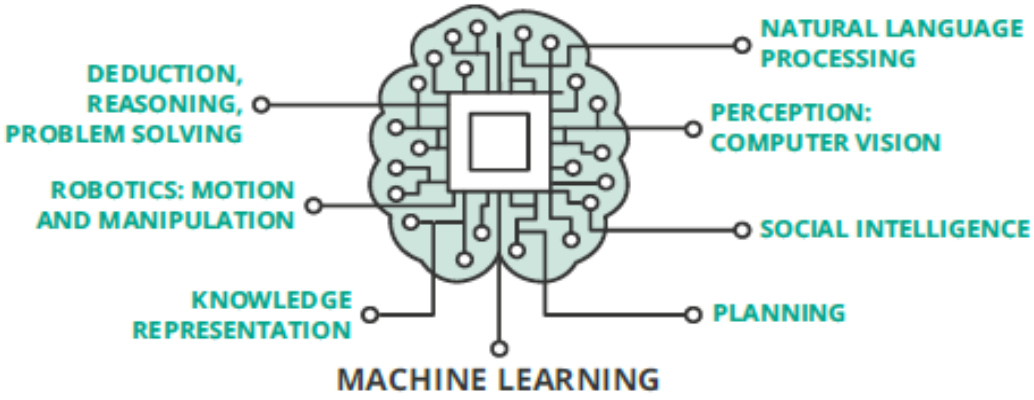


Figure 2. Composantes d'IA **[4]**

4. Machine Learning

Avant de parler sur le ML, nous devons parler sur la programmation traditionnelle qui nous implique d'écrire des règles (exprimées dans un langage de programmation), qu'utilisent des données pour nous donnent des réponses (Figure 3) **[5]**

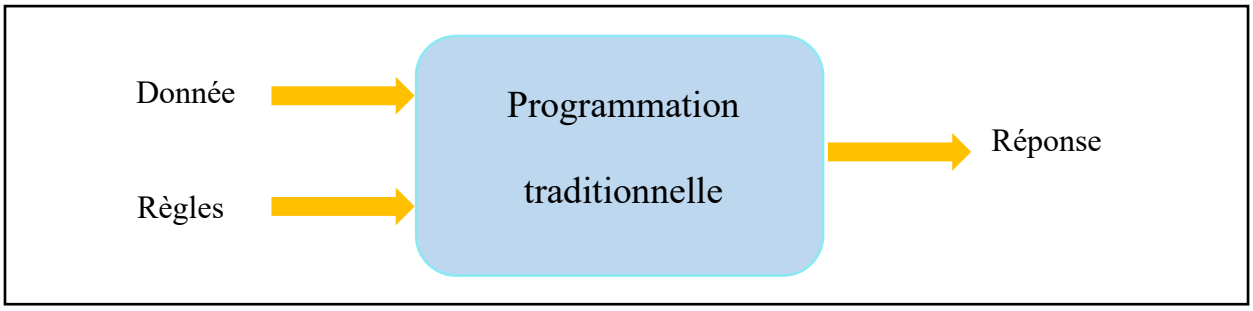


Figure 3. Le flux de programmation traditionnel

Par exemple, nous voulons créer un algorithme pour classer la réclamation d'un client pour deux cas:

Cas 01 :

Figure 4. Envoyer une réclamation (Cas 01)

L'algorithme du 1er Cas:

```
reclamation= Reclamation.getText().toString();
if (technique.isChecked())
    {problème= "Problème Technique" ;}
else if (respect.isChecked())
    {problème= "Problème de Respect";}
else if (gestion.isChecked())
    {problème= "Problème de Gestion";}
else if (metier.isChecked())
    {problème= "Problème de Métier";}
else if (comportement.isChecked())
    {problème= "Problème de Comportement";}
else if (securite.isChecked())
    {problème= "Problème de Sécurité";}
```

Cas 02 :

Figure 5. Envoyer une réclamation (Cas 02)

L'algorithme du 2ème Cas:

reclamation = Reclamation.getText().toString() ;

Dans le 1er cas, l'utilisateur doit choisir le type de son problème - même s'il y a plus de 20 types - contrairement au 2ème cas, il écrit la réclamation seulement.

Notre capacité à classer la réclamation de ce dernier à l'aide des règles traditionnelles est impossible.

C'est là que le Machine Learning entre en jeu.

Tout simplement, Machine Learning est un domaine d'informatique qui rend la machine capable de penser et de prendre des décisions par elle-même sans la donner une solution ou un code spécifique, grâce à son apprentissage par les expériences des humains et lorsqu'elle est exposé à des nouvelles données elle apprend, change et se développe par elle-même sans que nous ayons besoin de changer le code à chaque fois (Figure 6).

Donc, ce qui se passe, c'est qu'au lieu d'écrire le code à chaque fois pour un nouveau problème, nous fournissons les données à l'algorithme ML qui va construire la logique et fournir des résultats fondés sur les données. Initialement, les résultats obtenus ne sont pas corrects à 100%, mais avec le temps, la précision des algorithmes ML devient plus élevée. **[5]**

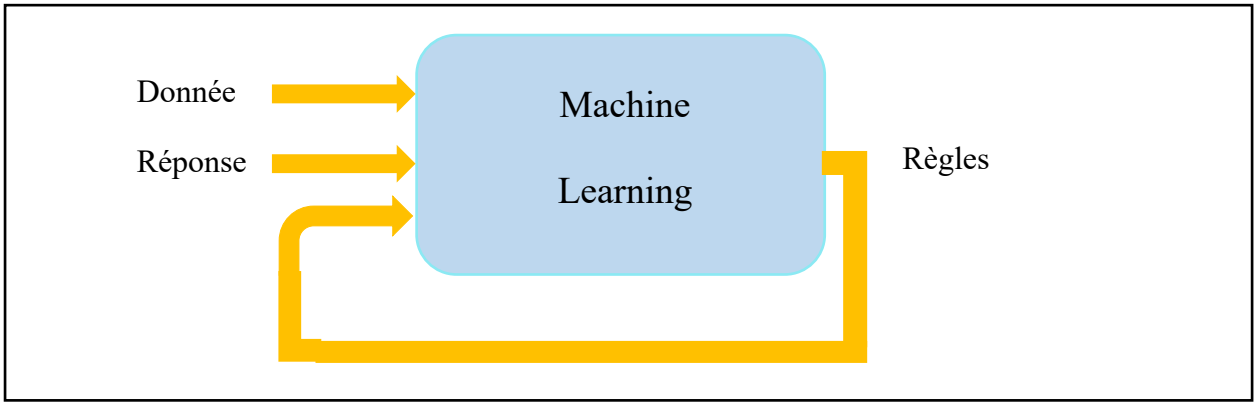


Figure 6. Le flux de Machine Learning

Donc pour classer la réclamation de client, nous allons collecter des données sur cette réclamation puis nous obtiendrons comme résultats « Technique », « De Respect », « de Gestion » (Figure 7)

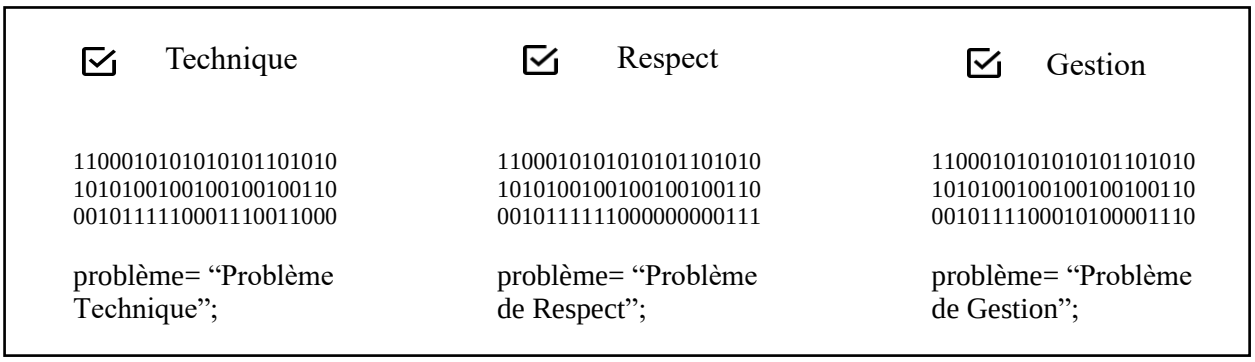


Figure 7. Du codage au ML: collecte et étiquetage des données

5. Les méthodes de Machine Learning

Il existe différents types de systèmes de Machine Learning (Figure 8) :

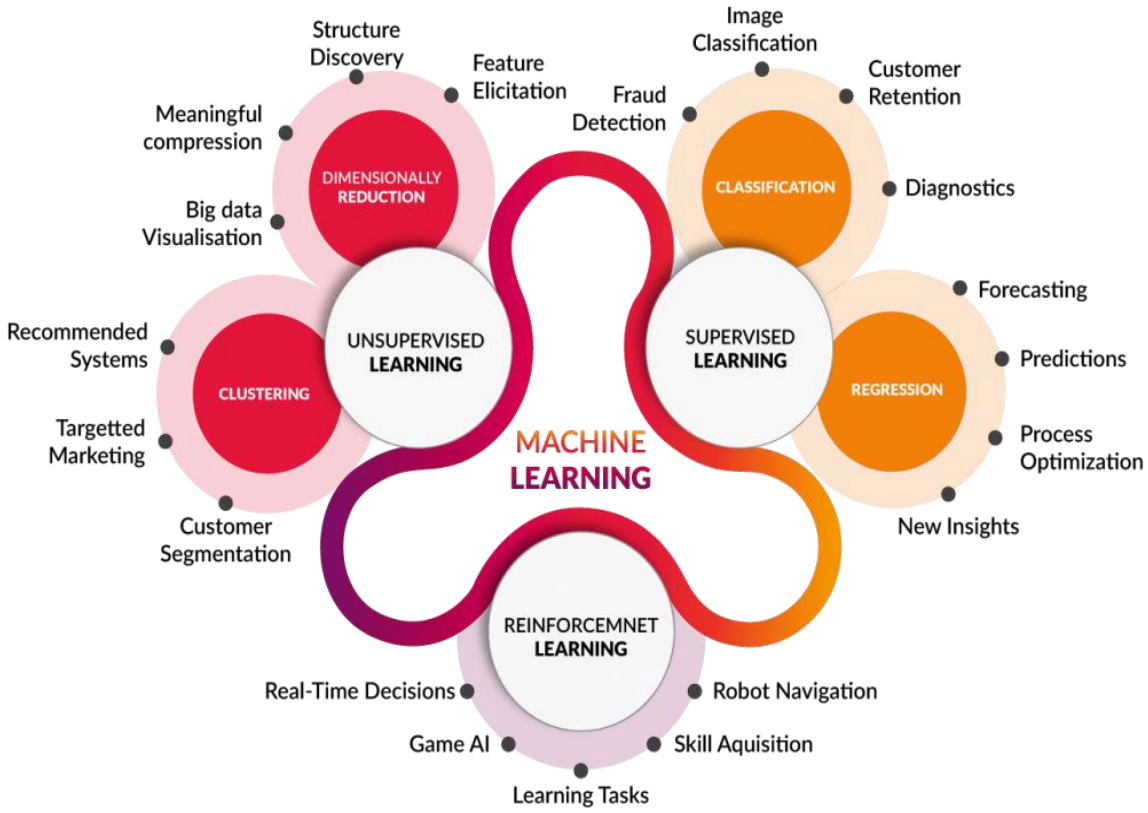


Figure 8. Les méthodes de Machine Learning **[6]**

5.1 Apprentissage supervisé

Les algorithmes d'apprentissage supervisé sont des modèles prédictifs développés fondés sur les entrées et les sorties.

L'objectif d'un algorithme d'apprentissage supervisé est d'utiliser le dataset pour produire un modèle qui prend en entrée un vecteur de caractéristiques X et produit des informations permettant de déduire la sortie Y de ce vecteur de caractéristiques. **[7]**

Pour décrire le problème d'apprentissage supervisé de manière un peu plus, on prend une fonction H : X → Y de sorte que H(X) soit un bon prédicteur pour la valeur correspondante de Y. **[8]**

Cette fonction H est appelée une hypothèse, le processus est donc comme ceci (Figure 9) :

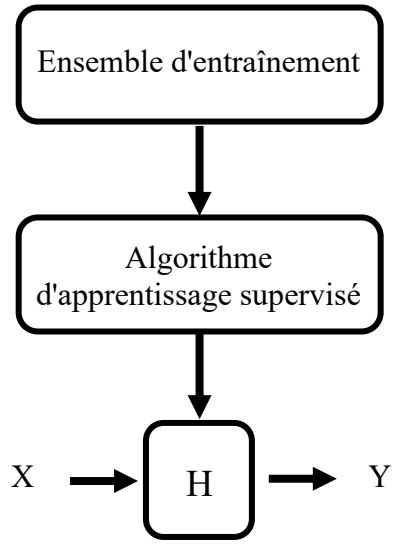


Figure 9. Le processus de la fonction H **[8]**

L'apprentissage supervisé est divisé en deux types :

5.1.1 Apprentissage supervisé par classification

Son idée est d'avoir des données liées les unes aux autres et vouloir les diviser en groupes par une ligne ou une courbe **[8]**
Il est divisé en deux types :

5.1.1.1 Classification binaire

Nous parlons de la classification binaire si le problème n'a que deux classes possibles, par exemple nous voulons savoir si la réclamation de client est à cause d'un problème technique ou un problème de respect (Figure 10) :

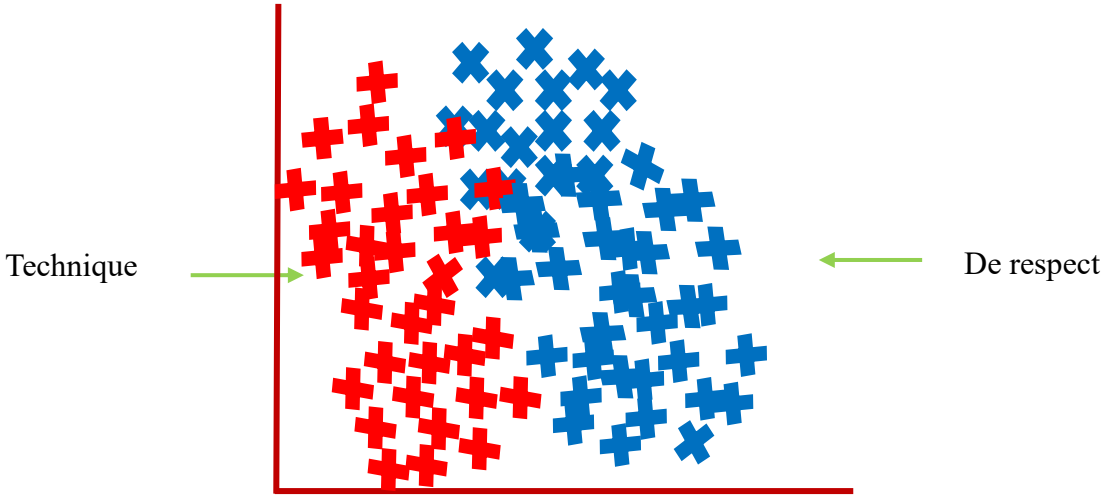


Figure 10. Classer la réclamation de client

5.1.1.2 Classification multi classes

Nous allons maintenant aborder la classification des données lorsque nous avons plus de deux catégories, ça veut dit qu'au lieu de Y = {0,1} nous allons étendre notre définition pour que Y = {0,1...N} [8] (Figure 11).

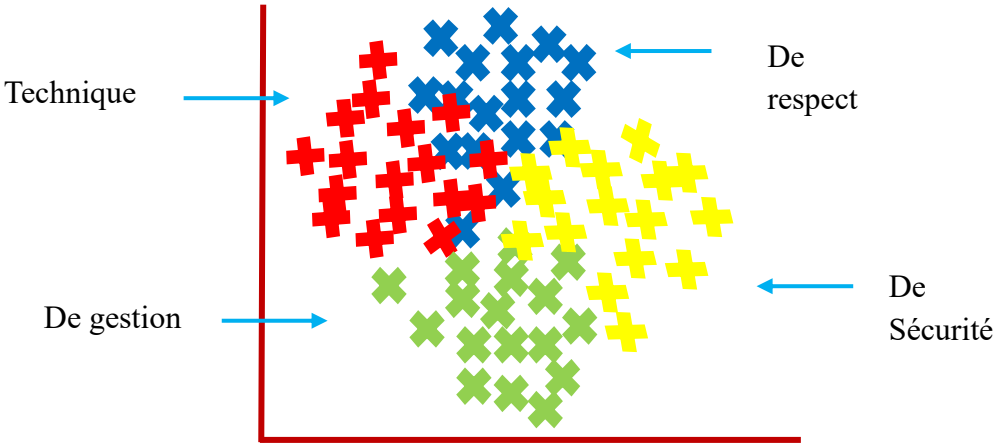


Figure 11. Classification multi classes

L'apprentissage supervisé n'est pas un processus simple, nous devons créer des modèles plus précis basés sur notre dataset.

5.1.2 Apprentissage supervisé par régression

Il s'agit de prendre en entrée une collection d'exemples étiquetés et produit un modèle qui peut prendre en entrée un exemple non étiqueté et produire une cible. **[7]**

L'apprentissage supervisé par régression est l'un des algorithmes de ML, les plus simples et les plus populaires, c'est une méthode statistique utilisée pour l'analyse prédictive. Il fait des prédictions pour des variables continues/réelles ou numériques telles que les ventes, le salaire, l'âge, le prix du produit, etc. (Figure 12). **[9]**

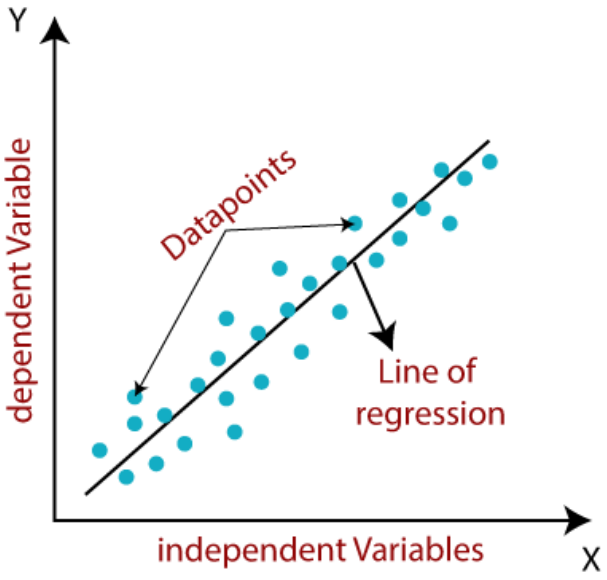


Figure 12. Fonctionnement de l'apprentissage supervisé par régression**[9]**

5.2 Apprentissage non supervisé

L'apprentissage non supervisé est un type de Machine Learning dans lequel les algorithmes reçoivent des données qui ne contiennent aucune étiquette ou instruction explicite sur ce qu'il faut en faire. L'objectif est que l'algorithme d'apprentissage trouve une structure dans les données d'entrée (Figure 13).

Donc l'apprentissage non supervisé est une sorte d'auto-apprentissage où l'algorithme peut trouver des modèles précédemment cachés dans les ensembles de données non étiquetés et donner la sortie requise sans aucune interférence. L'identification de ces modèles cachés aide au regroupement, à l'association et à la détection d'anomalies et d'erreurs dans les données. **[10]**

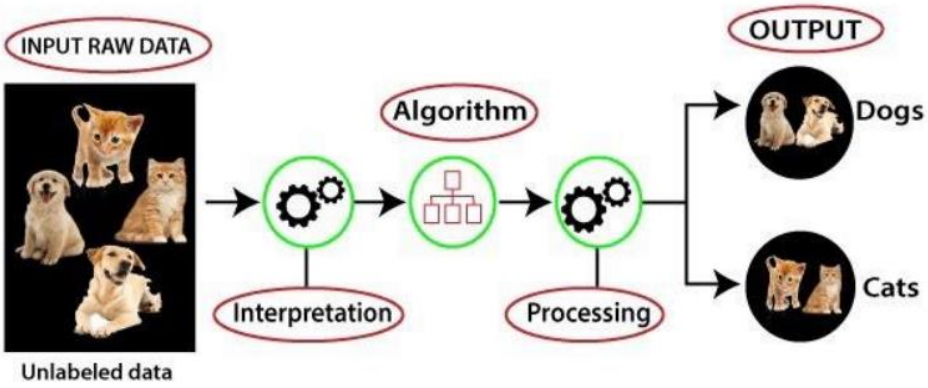


Figure 13. Illustration de l'apprentissage non supervisé **[11]**

5.3 Apprentissage semi supervisé

L'apprentissage semi-supervisé est une combinaison d'apprentissage supervisé et non supervisé. Il utilise une petite quantité de données étiquetées et une grande quantité de données non étiquetées et cela offre les avantages de l'apprentissage non supervisé et supervisé en évitant les difficultés liées à la recherche d'une grande quantité de données étiquetées, alors nous pouvons former un modèle pour étiqueter les données sans avoir à utiliser autant de données d'apprentissage étiquetées (Figure 14). **[12]**

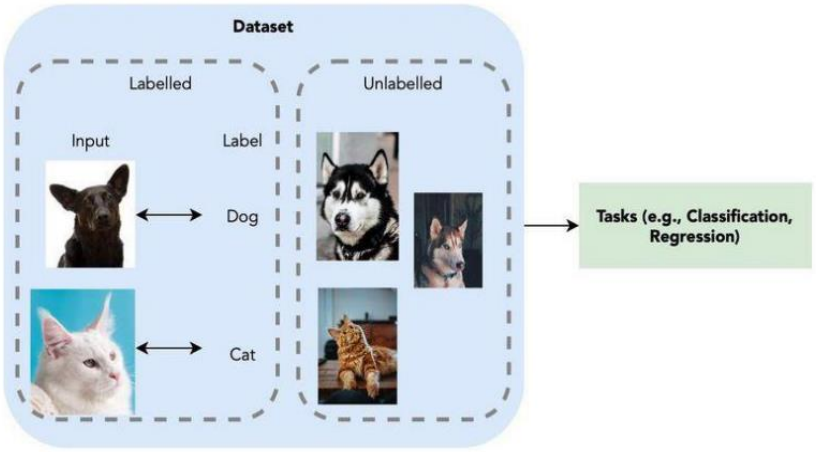


Figure 14. Illustration de l'apprentissage semi-supervisée **[12]**

5.4 Apprentissage par renforcement

Il s'agit de prendre des mesures appropriées pour maximiser la récompense dans une situation particulière, il est utilisé par divers logiciels et machines pour trouver le meilleur

chemin ou comportement possible qu'il devrait suivre dans une situation spécifique. L'apprentissage par renforcement diffère de l'apprentissage supervisé car dans ce dernier, les données de formation contiennent la clé de réponse, de sorte que le modèle est lui-même formé avec la bonne réponse, par contre dans l'apprentissage par renforcement, il n'y a pas de réponse, mais l'agent de renforcement décide quoi faire pour effectuer la tâche donnée (Figure 15) **[13]**.

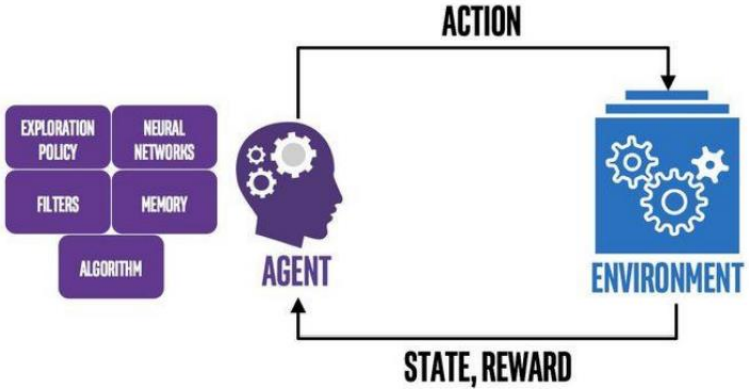


Figure 15. Illustration de l'apprentissage par renforcement **[13]**

6. Définition de modèle de ML

Un modèle de ML est l'expression d'un algorithme qui parcourt des montagnes de données pour trouver des modèles ou faire des prédictions (Figure 16). Alimentés par les données, les modèles de ML sont les moteurs mathématiques de l'IA. **[14]**

Le processus d'exécution d'un algorithme de ML sur le dataset (appelé training data) et d'optimisation de l'algorithme pour trouver certains modèles ou sorties est appelé model training.

La fonction résultante avec des règles et des structures de données est appelée « The Trained Machine Learning Model » **[15]**

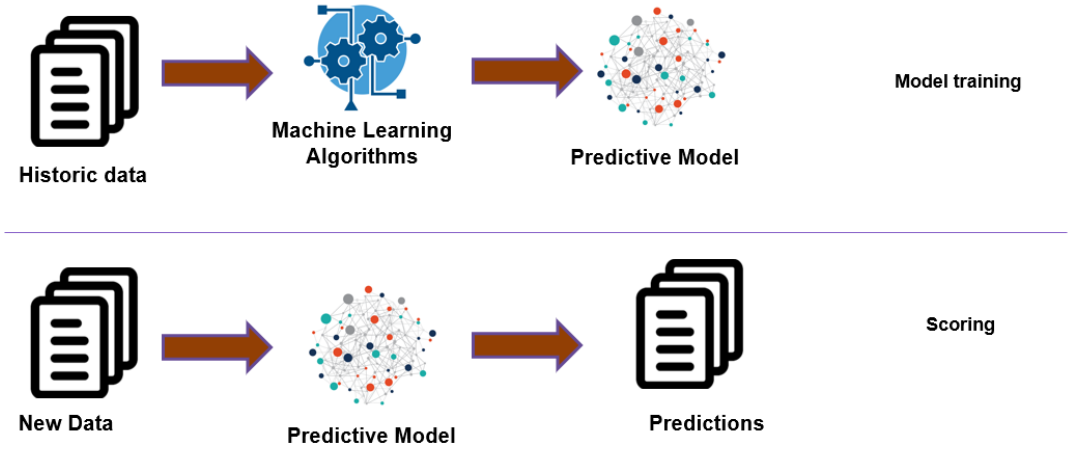


Figure 16. Objectif du modèle de ML **[16]**

6.1 Les différents modèles de ML

Pour chaque méthode de ML, il existe de nombreux modèles :

6.1.1 Modèle d'apprentissage supervisé :

Chaque type de l'apprentissage supervisé a ses modèles :

6.1.1.1 Modèle d'apprentissage supervisé par classification:

Pour l'apprentissage supervisé par classification, il existe trois modèles qui sont les plus communs dans ML :

1. **Régression Logistique**

Ce type de modèle utilise la fonction sigmoid afin de trouver la frontière de décision (la ligne qui sépare la zone où Y = 0 et où Y = 1) **[8]**, elle est créée par notre fonction d'hypothèse :

$$H(X) = g(\theta^T X)$$

Avec :

- $g(z) = \frac{1}{1+e^{-z}}$, est la fonction sigmoïde (Figure 17)

- $0 \leq H(X) \leq 1$
- X: sont les entrées et sont appelées caractéristiques (symbolisées par n).
- θ: sont les coefficients des caractéristiques, on les trouve utilisant la fonction de cout J.
- $X = \begin{bmatrix} X0 \\ X1 \\ X2 \\ . \\ . \\ . \\ Xn \end{bmatrix}$
- $\theta = \begin{bmatrix} \theta 0 \\ \theta 1 \\ \theta 2 \\ . \\ . \\ . \\ \theta n \end{bmatrix}$
- $\theta^T X = \theta_0 X_0 + \theta_1 X_{1+\cdots+} \theta_n X_n$
- $X_0 = 1$
- $\theta^T X$ est une équation de 1[er] degré ou 2[ème] degré, 3[ème] degré… ça dépend le problème.

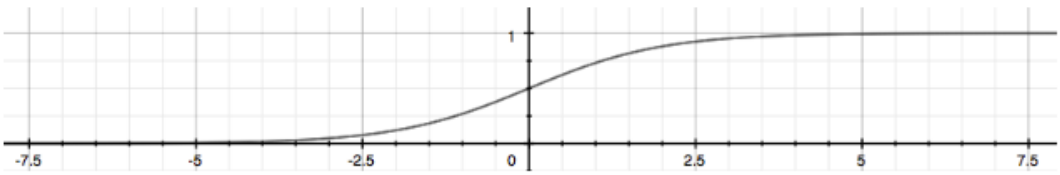


Figure 17. La fonction sigmoïde **[8]**

1.1 Fonction de coût J

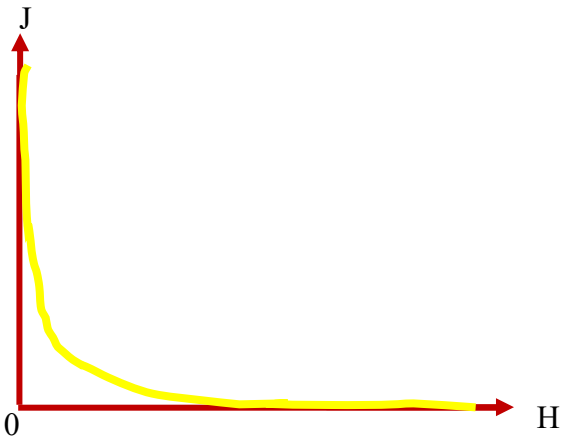
Nous pouvons mesurer la précision de notre fonction d'hypothèse en utilisant la fonction de coût **[8]** pour trouver les valeurs de θ qui minimisent J le moins possible, donc :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} COST\left(H(X^i), Y^i\right)$$

Avec:

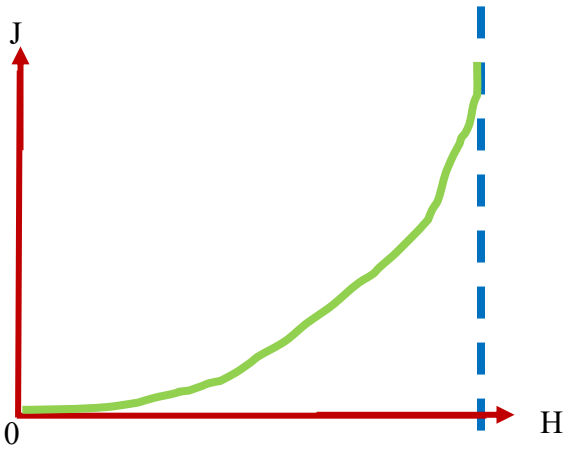
- m : nombre de lignes de vecteur de caractéristiques X.
- $COST\ (H(X), Y) = -\log\big(H(X)\big), si\ Y = 1$
- $COST\ (H(X), Y) = -\log\big(1 - H(X)\big), si\ Y = 0$

Si Y= 1 :



- $H \to 0 \Rightarrow J \to +\infty$
- $H \to 1 \Rightarrow J \to 0$
- $H = 1 \Rightarrow J = 0$

Si Y=0 :



- $H \to 1 \Rightarrow J \to +\infty$
- $H \to 0 \Rightarrow J \to 0$
- $H = 0 \Rightarrow J = 0$

Nous supposons des valeurs quelconques de θ puis nous utilisons un des algorithmes d'optimisation afin que nous aide d'obtenir les meilleures valeurs de θ.

1.2 Les algorithmes d'optimisation

Son objectif est de réduire les valeurs de θ jusqu'à ce que nous trouverons la valeur la plus basse de J (Figure 18). **[8]**

Il existe beaucoup d'algorithme d'optimisation par exemple :

- Stochastic gradient descent (HDD).
- Atom optimization.
- RMS PROP.
- Convex optimization.
- Gradient Descent.
- Conjugate gradient.
- BFGS

- L-BFGS **[8]**

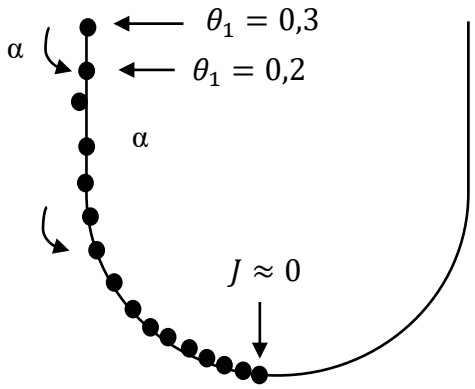


Figure 18. Le rôle d'algorithme d'optimisation

Avec:

- α: le taux d'apprentissage, Un α plus petit entraînerait un pas plus petit, plus de précision et plus de temps, et un α plus grand entraînerait un pas plus grand, moins de précision et moins de temps c'est pour ça il faut que nous choisirons la meilleure valeur pour α. **[8]**

Une fois que nous aurons une information sur une nouvelle réclamation, nous multiplierons les caractéristiques X aux meilleures valeurs de θ et le résultat sera 1 (technique) ou 0 (de respect).

Dans le cas dans la classification multi classe, nous allons choisir essentiellement une classe, puis regrouperons toutes les autres en une seule seconde classe. Nous le faisons à plusieurs reprises, en appliquant une régression logistique binaire à chaque cas, puis en utilisant l'hypothèse qui a renvoyé la valeur la plus élevée comme notre prédiction. **[8]**

2. Support Vector Machines (SVM)

SVM est une technique de classification supervisée qui peut devenir très compliquée mais plutôt intuitive au niveau le plus fondamental.

Le SVM trouvera un hyperplan ou une frontière qui maximisera la marge entre les différentes classes de données, son objectif est le même que l'objectif de la régression logistique : trouver la bonne ligne d'ajustement.

Pour trouver cette dernière, nous devons trouver les points les plus proches (vecteurs de support) de la ligne des classes, après on va calculer la distance (la marge) entre la ligne et les vecteurs de support, notre objectif est de maximiser cette marge.

L'hyperplan pour lequel la marge est maximale est l'hyperplan optimal. (Figure 19) **[17]**

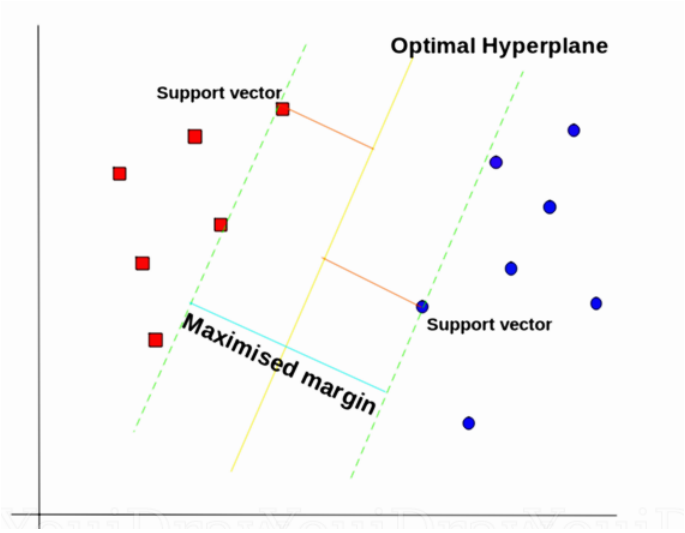


Figure 19. Hyperplan optimal utilisant l'algorithme SVM **[17]**

2.1 Fonction de coût J

Ici, la fonction de cout J est différente de celle de la régression logistique :

$$C \sum_{i=1}^{m} \left[Y^i \, cost_1\left(\theta^T X^i\right) + \left(1 - Y^i\right) cost_0\left(\theta^T X^i\right) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 \quad \textbf{[8]}$$

Avec :

- C : est une constante qui contrôle le compromis entre une frontière de décision lisse et la classification correcte des points d'entraînement. Un C plus petit entraînerait le underfitting, et un C plus grand entraînerait le overfitting, c'est pour ça il faut que nous choisissions la meilleure valeur pour C. **[8]**
- $cost_1$: le slob lorsque Y=1 (Figure 20) :

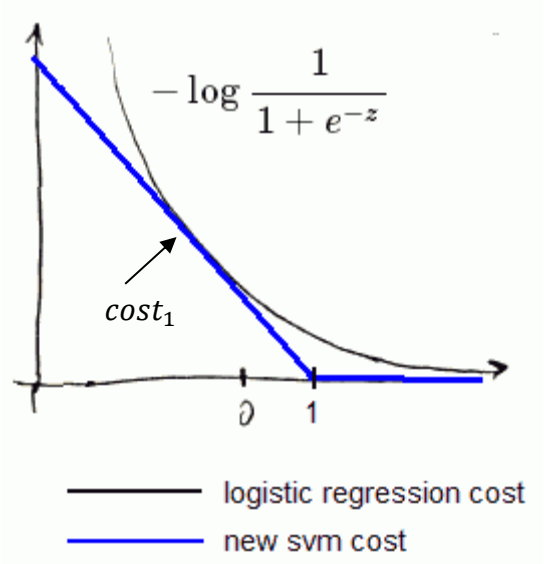


Figure 20. La valeur de cost1 **[8]**

- $cost_0$: le slob lorsque Y=0 (Figure 21) :

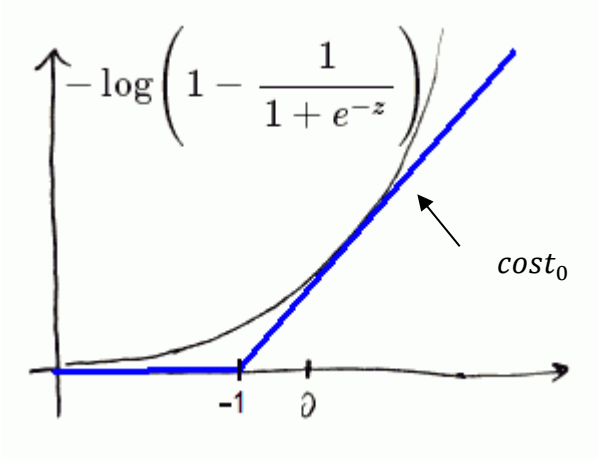


Figure 21. La valeur de cost0 **[8]**

Dans le SVM :

- La classification est effectuée si $\theta\, TX \geq 1$ et $\theta\, TX \leq -1$.

- La ligne séparée peut être linéaire ou non linéaire (Figure 22) :

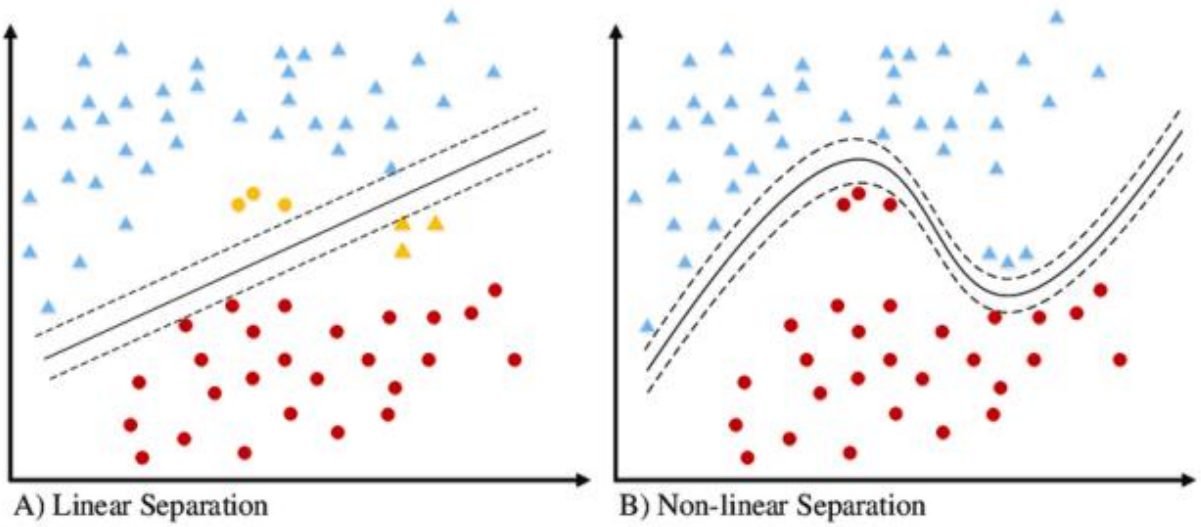


Figure 22. Le SVM linéaire et non linéaire **[18]**

Les algorithmes SVM utilisent un ensemble des fonctions mathématiques définies comme « kernels ».

2.2 Kernels

L'objectif du kernels est de prendre des données en entrée et de les transformer sous la forme requise, elle renvoie le produit scalaire entre deux points dans un espace de caractéristiques approprié **[19]**. Voici quelques types de kernels :

2.2.1 Polynomial kernel :

Il est populaire dans le traitement d'image, son équation est :

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

Où d est le degré du polynôme. **[19]**

2.2.2 Linear Kernel :

est utilisé lorsque les données sont séparables linéairement, c'est-à-dire qu'elles peuvent être séparées à l'aide d'une seule ligne **[20]**, son équation est :

$$k(X, Y) = X^T Y$$ **[21]**

2.2.3 Sigmoid :

cette fonction équivaut à un modèle perceptron à deux couches du réseau neuronal, qui est utilisé comme fonction d'activation pour les neurones artificiels **[22]**, son équation est :

$$K(X,Y) = \tan(\delta X^T Y + C)$$ **[21]**

2.2.4 RBF :

Radial basis function (RBF) est un noyau à usage général utilisé lorsqu'il n'y a aucune connaissance préalable des données, son équation est :

$$k(X,Y) = e(-\frac{\|X-Y\|^2}{2\delta^2})$$ **[21]**

Où σ définit la portée de l'influence d'un seul exemple de formation. S'il a une valeur faible, cela signifie que chaque point a une portée éloignée et, inversement, une valeur élevée de gamma signifie que chaque point a une portée proche. **[17]**

SVM ne prend pas la classification multi classes en charge, c'est pour ça nous utilisons des approches pour effectuer une multi-classification sur les énoncés de problème à l'aide de SVM.

2.3 Approche un contre un (OVO)

C'est la plus simple et la plus ancienne des approches, elle décompose notre problème de classification multi classe en sous-problèmes (problèmes de classification binaire) afin d'obtenir des classificateurs binaires pour chaque paire de classes.

Dans l'approche OVO (One Vs One), nous essayons de trouver l'hyperplan qui sépare toutes les deux classes, en négligeant les points de la troisième classe (Figure 23).

Le problème majeur avec cette approche est que nous devons former trop de SVM. **[23]**

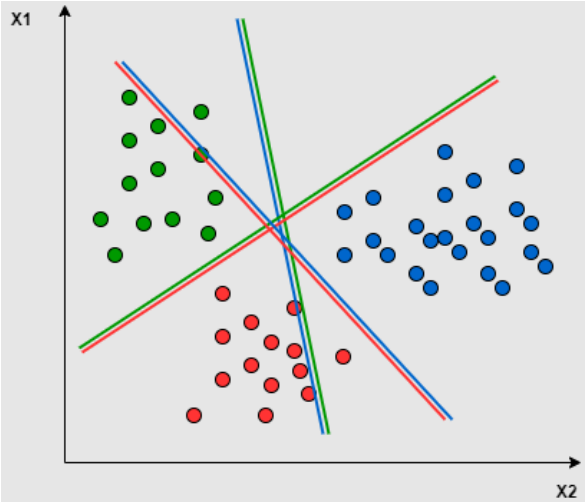


Figure 23. Image montrant 3 classes différentes séparer par le OVO **[23]**

2.4 Approche un contre tous (OVA)

Dans l'approche OVA (One Vs All), nous essayons de trouver un hyperplan pour séparer les classes, cela signifie que la séparation prend en compte tous les points et les divise ensuite en deux groupes dans lesquels il y a un groupe pour les points d'une classe et l'autre groupe pour tous les autres points (Figure 24). **[23]**

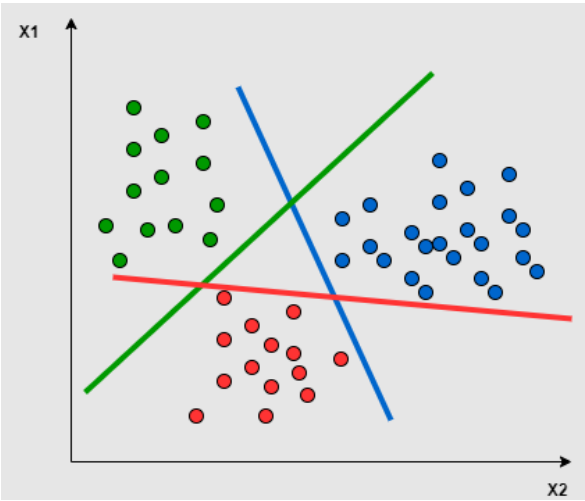


Figure 24. Image montrant 3 classes différentes séparer par le OVA **[23]**

Dans cette méthode, si on a N classes donc nous apprenons N SVM :

- SVM numéro -1 apprend "class_output = 1" vs "class_output ≠ 1"

- SVM numéro -2 apprend "class_output = 2" vs "class_output ≠ 2"
- Le numéro SVM -N apprend "class_output = N" vs "class_output ≠ N"

Il y a quelques difficultés à former ces SVM :

3. **Trop de calcul :** Pour mettre en œuvre la stratégie OVA, nous avons besoin de plus de points d'entraînement, ce qui augmente notre calcul.

4. **Les problèmes deviennent déséquilibrés :** la grande différence de nombre de points entre les classes rend notre problème déséquilibré. **[23]**

3. Naïve Bayes

Est un autre classificateur populaire utilisé en Data Science, son idée est inspirée du théorème de Bayes **[16]** :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ **[24]**

6.1.1.2 Modèle d'apprentissage supervisé par régression:

Il existe différents modèles pour l'apprentissage supervisé par régression tel que :

- Arbre de régression.
- Régression linéaire et non linéaire.
- Régression linéaire bayésienne.
- Régression polynomiale.

6.1.2 Modèle d'apprentissage non supervisé :

Les modèles d'apprentissage non supervisé peuvent effectuer des tâches plus complexes que les modèles d'apprentissage supervisé, voici les principales tâches qui utilisent cette approche :

- Regroupement
- Association
- Réduction de la dimensionnalité **[10]**

7. NLP

Natural Language Processing (NLP) représente une très grande partie de l'IA, il est la capacité d'un programme informatique à comprendre le langage humain tel qu'il est parlé et écrit. **[45]**

Son objectif est de prendre des données du monde réel, les traiter et leur donner un sens d'une manière qu'un ordinateur peut comprendre. **[45]**

Le NLP comporte une phase principale: le prétraitement des données qui consiste à préparer, à nettoyer et à transformer les données textuelles en donnée numérique pour que les machines puissent les analyser. **[45]**

Pour transformer ces données textuelles en donnée numérique nous utilisons différentes approches par exemple :

- TFIDF
- Word2vec
- One Hot Encoding **[46]**

Le NLP joue un rôle essentiel dans la technologie et la façon dont les humains interagissent avec elle. Il est utilisé dans de nombreuses applications du monde réel dans les sphères commerciales et grand public, notamment les chabots, la cybersécurité, les moteurs de recherche et l'analyse de données volumineuses. Bien qu'elle ne soit pas sans défis, le NLP devrait continuer à être une partie importante de l'industrie et de la vie quotidienne. **[45]**

8. Problèmes dans Machine Learning

1. Problème d'Underfitting :

Ce problème se produit lorsque la forme de notre fonction d'hypothèse H correspond mal à la tendance des données. Il est généralement causé par une fonction trop simple ou utilisant beaucoup de caractéristiques. **[8]**

2. Problème d'Overfitting:

Ce problème est causé par une fonction d'hypothèse qui s'adapte aux données disponibles mais ne se généralise pas bien pour prédire de nouvelles données. Cela est fréquemment causé par une fonction compliquée qui crée de plusieurs courbes et angles sans lien avec les données. **[8]**

Il existe trois options principales pour résoudre le problème d'Overfitting :

- **1ère option :** supprimer les caractéristiques inutilisées, par exemple : pour accepter un étudiant dans l'université, nous devons connaître son âge, mais sa taille n'est pas une caractéristique importante, donc nous la supprimons. **[8]**

- **2ème option :** changer le coefficient des caractéristiques qui ont un effet et une importance minimes, par les multipliant par un grand nombre pour minimiser la fonction de coût.
  Par exemple :

$$H(X) = \theta_0 X_0 + \theta_1 X_1 + 10000\ \theta_2 X_2$$

  X2 représente le climat qui a un effet sur l'étude de l'étudiant mais ce n'est pas une caractéristique importante (il a un petit effet), donc nous le multiplions à 10000. **[8]**

- **3ème option :** appliquer la régularisation, son idée est d'utiliser un facteur de régularisation λ en le multipliant par la somme du carré de tous les θ sauf θ_0, puis nous l'ajoutons à la fonction de coût J.
  Nous supposons λ puis nous faisons un teste, si le « overfitting » n'a pas disparu, nous devons changer sa valeur et tester à nouveau. **[8]**

9. Avantages et inconvénients de Machine Learning

1- Avantages de Machine Learning :

- **Identifie facilement les tendances et les modèles :**
  Le Machine Learning peut analyser de grandes données et découvrir beaucoup de choses qui ne seraient pas apparente pour les humains.
- **Aucune intervention humaine nécessaire :**
  Avec Machine Learning, l'humain n'a pas à surveiller son projet à chaque étape du processus, en reconnaissant les spams…
- **Amélioration continue :**

Les algorithmes de Machine Learning continuent de s'améliorer en finesse et en performance en gagnant de l'expérience, ce qui leur permet de prendre de meilleures décisions.

- **Traitement de données multidimensionnelles et multi variétés :**

  Les algorithmes de Machine Learning gèrent des données multidimensionnelles et multi variétés, et ils peuvent le faire dans un environnement inefficace ou incertain.

2- Inconvénients de Machine Learning :

Tous ces avantages ne rendent pas le Machine Learning parfait, les facteurs suivants servent à le limiter :

- **L'acquisition des données :**

  Machine Learning nécessite de grands ensembles de données à encadrer, et ceux-ci doivent être d'excellente qualité, complets et équitables.

- **Temps et ressources :**

  Machine Learning nécessite suffisamment de temps pour laisser les algorithmes se développer et apprendre, il a également besoin d'énormes ressources pour fonctionner.

- **Interprétation des résultats :**

  Un autre défi majeur est la capacité à interpréter exactement les résultats générés par les algorithmes.

- **Haute sensibilité aux erreurs :**

  Machine Learning est indépendant mais très prédisposé aux erreurs et il faut un certain temps pour reconnaître la source du problème, et encore plus pour le corriger. **[25]**

10. Importance de Machine Learning

Machine Learning est devenu un différenciateur concurrentiel important pour de nombreuses entreprises parce qu'elle donne aux entreprises une vue des tendances du comportement des clients et des modèles opérationnels commerciaux, tout en soutenant le développement de nouveaux produits. De nombreuses entreprises leaders d'aujourd'hui, telles que Facebook, Google et Uber, font de Machine Learning un élément central de leurs opérations.

Avec l'évolution constante du domaine, il y a eu une augmentation subséquente des utilisations, des exigences et de l'importance de Machine Learning.

L'interprétation est effectuée à l'aide d'ensembles automatiques de méthodes génériques qui ont remplacé les techniques statistiques traditionnelles. **[26]**

11. Conclusion

Au cours de ce chapitre, nous avons présenté brièvement l'intelligence artificielle et ses applications, ensuit nous avons bien expliqué en détail le terme « Machine Learning », ses méthodes et ses modèles, son importance, ses avantages et inconvénients. Dans le chapitre suivant, nous parlerons du l'application RightNow By Brenco en détails, puis nous passerons à expliquer l'objectif de ce travail et ses étapes.

Chapitre 02 : Spécification de l'approche

1. Introduction

Aujourd'hui, avec l'évolution technologique, les entreprises s'efforcent d'automatiser et de réaligner leurs opérations en permanence pour apporter aux clients ce qu'ils veulent plus rapidement et plus efficacement que jamais auparavant. **[27]**

Brenco Engineering & Consulting est l'une de ces entreprises qu'a fait preuve d'un engagement fort dans le cadre du développement de l'écosystème des start-up algériennes, à travers la détection et le soutien d'entrepreneurs dans les industries du numérique et par créer des solutions informatisées afin d'aider les gens. **[28]**

RightNow By Brenco est l'une des solutions de la société Brenco, elle est une solution cloud (Mobile App & SaaS) en marque blanche qui nous permettra d'offrir un nouvel user expérience UX aux utilisateurs. **[29]**

Au cours de ce chapitre, nous présentons en détail l'application RightNow By Brenco, ses objectifs et les étapes que nous suivons pour la construction du modèle.

2. RightNow By Brenco

RightNow By Brenco est une application professionnelle mobile et web en marque blanche basée sur le cloud qui permet de collecter des données partout, à tout moment et sur n'importe quel appareil.

Elle permet aux directions/institutions d'avoir une vision à 360° de leurs clients/citoyens grâce à des algorithmes intelligents, d'établir des stratégies basées sur les données et de prendre des décisions précises. **[29]**

Figure 25. Logo et Slogan de l'application RightNow By Brenco **[29]**

RightNow permet de :

1. Collecter de données sans limite et en temps réel, quel que soit le lieu.
2. Suivre et analyser les progrès du processus global.
3. Identifier les perspectives, les lacunes et les opportunités de l'entreprise.
4. Garantir l'assurance qualité et la conformité.
5. Fournir des informations de manière efficace et sécurisée.
6. Automatiser les flux de travail et optimiser les opérations.
7. Récupérer la data avec le mode Hors-Ligne : l'application RIGHTNOW permet de récolter la data sans accès internet puis de synchroniser ultérieurement.
8. Créer des formulaires à l'infini selon le besoin en Data : créer des formulaires selon la structure data à travers des champs riches et variés (Image, Audio, texte, QCM...)
9. Gérer différentes équipes à travers la logique de groupes : gérer les équipes et les taches assignés en regroupant les formulaires par équipe.
10. Notifier les collaborateurs sur l'application : notifier les équipes terrain en temps réel à travers du push notification.
11. Smart algorithme pour décider en fonction des datas reçues : préconfigurer des orientation/directives en fonction des réponses aux formulaires avec notre algorithme conditionnel.
12. Géolocaliser la data à travers d'une map interactive : suivre la remontée d'information sur terrain à travers la géolocalisation des devices sur une map interactive.
13. Afficher le logo à travers une solution en marque blanche : brander la solution aux couleurs à travers notre offre en marque blanche.
14. Déploiement de la solution en SaaS ou On-Premise : déployer la solution sur les serveurs propriétaires ou profiter d'un hébergement en SaaS.
15. Une solution sur tous les support (mobile et web) : Profiter de la solution web et mobile de RightNow pour envoyer la data à partir de tous les devices. **[29]**

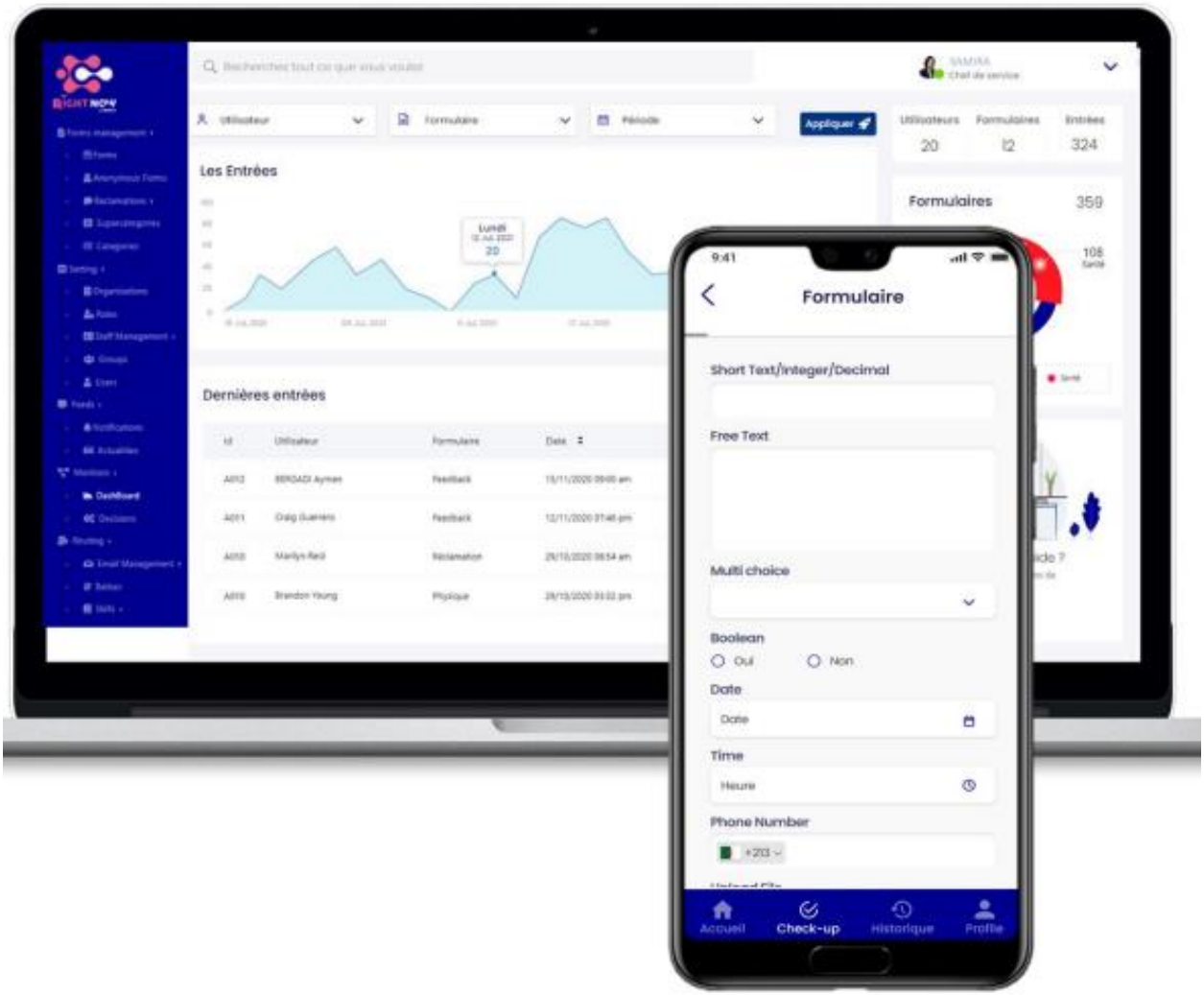


Figure 26. Capture de l'application mobile et web RightNow **[29]**

3. Catégorisation automatique des réclamations clients

Comme nous avons mentionnés précédemment, la société SETRAM, un des clients de RightNow By Brenco, a un problème des mauvaises catégorisations de réclamations par les clients, par exemple le client catégorise sa réclamation à propos de « ne pas porter le masque » comme un problème de respect mais SETRAM la considère un problème de comportement, ou certains clients la catégorisent comme un problème de respect et d'autre comme un problème de comportement, donc SETRAM se trouve obligée de recatégoriser manuellement ces réclamations afin d'obtenir des statistiques précises et ça prend beaucoup de temps et d'effort.

Notre objectif dans ce travail est de construire un modèle de ML qui permettra à la société SETRAM de catégoriser d'une manière automatique ses réclamations reçues.

Ces réclamations reçues, sont recueillies par un formulaire créé par SETRAM, représentent le dataset (l'ensemble des entrées) que nous utilisons pour la construction du modèle.

En conséquence, ces réclamations doivent être classées en six catégories que nous avons défini et qui représentent l'ensemble des sorties de notre modèle :

- Gestion : S'il y a une erreur administrative.
- Technique : Lorsqu'une erreur technique survient à la station.
- Comportement : S'il y a un problème avec le comportement des agents.
- Respect : En cas de non-respect des règles de l'entreprise soit par les agents ou les clients.
- Sécurité : S'il y a quelques choses qui met les clients en danger.
- Métier : Lorsqu'il y a un problème dans la profession des agents en termes de qualité et de fabrication.

Pour la construction de notre modèle de ML, nous utilisons l'apprentissage supervisé par classification multi classes à l'aide du SVM et l'approche OVO, que nous avons bien expliqué dans le Chapitre 01.

La Figure 27 suivante explique le flux général de la tâche de catégorisation des réclamations à l'aide du modèle construit:

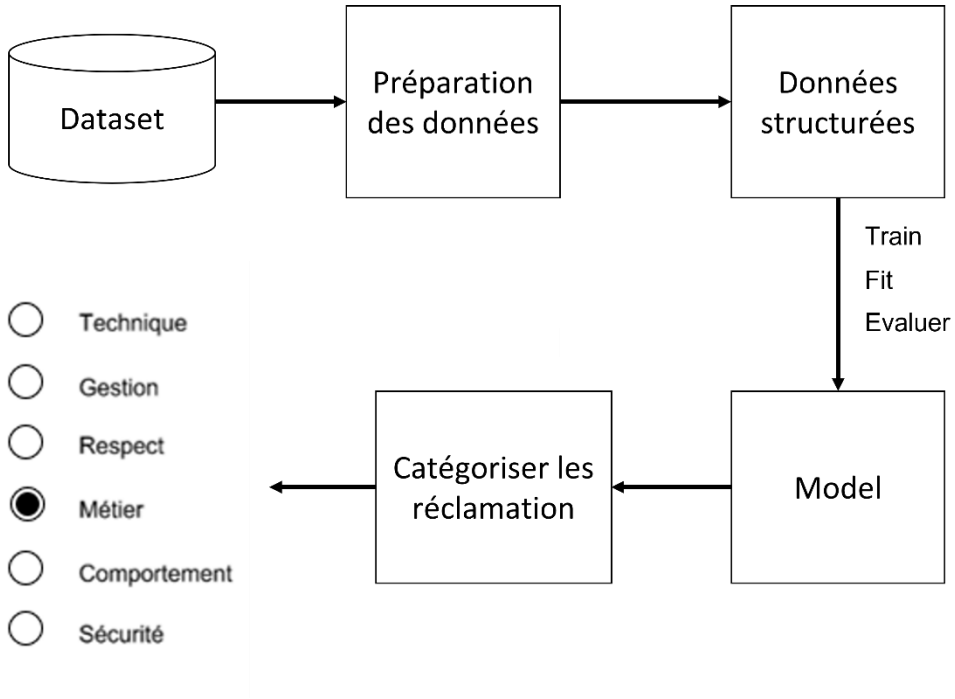


Figure 27. Le flux général de la tâche de catégorisation des réclamations à l'aide du modèle de ML

Afin de construire ce modèle, nous devons suivre ces étapes :

3.1 Etape 01 : Préparation de données

La préparation des données est une condition préalable à la conception de modèles et d'analyses de ML précis. **[30]**

Dans notre cas, elle contienne les étapes suivantes :

3.1.1 Dataset

Comme nous avons mentionnés, nous utilisons les réclamations reçues d'un formulaire que la société SETRAM a créé comme un dataset (Figure 28) :

Submittin	Device Us	Form Com	Upload Cc	Submissic	Submissic	A.Fiche si	A.Fiche si	A.Fiche si	A.Fiche si	A.Fiche si	Réclamati	Catégorie
Android_(	EN03	2022-01-0	2022-01-0	47302810	cf8cfb6b-a	1	1	2	1	2022-01-0	Le kioosq	gestion
Android_(	EN03	2022-01-0	2022-01-0	47302833	138b54ff-1	1	1	2	2	2022-01-0	L'agent ne	gestion
Android_(	EN03	2022-01-0	2022-01-0	47302914	939fa496-1	7	6	8	6	2022-01-0	L'agence c	gestion
Android_(	EN03	2022-01-0	2022-01-0	47302973	714d6909-7	7	6	8	8	2022-01-0	Kiosque 0	gestion
Android_(	EN03	2022-01-0	2022-01-0	47303025	7706b9fa-	12	10	12	10	2022-01-0	L'agent n'a	gestion
Android_(	EN03	2022-01-0	2022-01-0	47303045	8437549c-	12	10	12	12	2022-01-0	Il dans cet	gestion
Android_(	EN03	2022-01-0	2022-01-0	47304410	89216a98-	13	11	13	11	2022-01-0	Il ya 2 age	respect
Android_(	EN03	2022-01-0	2022-01-0	47305643	1ceacf10-1	13	11	13	13	2022-01-0	Il ya 3 age	respect
Android_(	EN03	2022-01-0	2022-01-0	47314239	7ba3d87e-	3	2	3	2	2022-01-0	Valider de	technique
Android_(	EN03	2022-01-0	2022-01-0	47314297	32df75f9-(	3	2	3	3	2022-01-0	Il ya un dé	technique
Android_(	EN03	2022-01-0	2022-01-0	47314443	09257ede-	6	5	7	5	2022-01-0	agent à ve	respect
Android_(	EN03	2022-01-0	2022-01-0	47314502	4c31b2f7-(	6	5	7	7	2022-01-0	Les agents	respect
Android_(	EN03	2022-01-0	2022-01-0	47314819	3d585bfe-	11	9	15	9	2022-01-0	Les3 agen	respect
Android_(	EN03	2022-01-0	2022-01-0	47314920	5441ff01-2	11	9	15	15	2022-01-0	Circulatio	gestion
Android_(	EN03	2022-01-0	2022-01-0	47315984	cc475135-(	22	22	20	20	2022-01-0	Voir le tra	respect
Android_(	EN03	2022-01-0	2022-01-0	47316056	f77a2d2d-	22	20	22	22	2022-01-0	L'absence	gestion
Android_(	EN03	2022-01-0	2022-01-0	47316272	1e01c75c-	23	23	21	21	2022-01-0	L'absence	gestion
Android_(	EN03	2022-01-0	2022-01-0	47316416	a152b70c-	23	21	23	23	2022-01-0	Le kioosq	gestion
Android_(	EN03	2022-01-0	2022-01-0	47316936	0e5858d2-	24	24	26	24	2022-01-0	L'agent ne	gestion
Android_(	EN03	2022-01-0	2022-01-0	47317012	1552dc68-	24	24	26	26	2022-01-0	L'agence c	gestion
Android_(	EN03	2022-01-0	2022-01-0	47352040	15c64246-	31	25	20	25	2022-01-0	Kiosque 0	gestion

Figure 28. Dataset

Cet dataset est composé de 643 lignes et 14 colonnes :

1. Submitting Device : Le dispositif de soumission.
2. Device User : Utilisateur de dispositif
3. Form Completed : Date et heure de remplir le formulaire
4. Upload Completed : Date et heure du téléchargement du formulaire
5. Submission ID : ID de soumission.
6. Submission ID generated by device : ID de soumission généré par l'appareil.
7. A.Fiche signalétique : A-1.Numéro parcours : On donne à chaque parcours un certain numéro pour les différencier.
8. A.Fiche signalétique : A-2. Station Départ : Le numéro de la station de départ.
9. A.Fiche signalétique : A-3. Station Destination : Le numéro de la station de destination.
10. A.Fiche signalétique : A-4. Station évaluée : Le numéro de station évalué par le client.
11. A.Fiche signalétique : A-7.Date : Le jour où l'incident pour lequel le client a déposé la plainte s'est produit.
12. Reclamation : La réclamation du client.
13. Catégorie : Type de problème de chaque réclamation.

Nous avons considéré dans notre travail seulement les colonnes qui ont une relation avec notre objectif : la colonne « Réclamation » (L'ensemble des entrés X) et la colonne « Catégories » (L'ensemble des sorties Y).

3.1.2 Equilibrage des données

On parle de déséquilibrage de données lorsque le dataset contient des catégories asymétriques :

1- Les catégories qui constituent une grande partie de l'ensemble de données sont appelées catégories majoritaires.
2- Les catégories qui constituent un pourcentage plus faible sont issus des catégories minoritaires.

Le déséquilibre des données fait que le modèle de formation passe la plupart de son temps sur les catégories majoritaires et n'apprend pas assez les catégories minoritaires.

Au début, nous avons que trois catégories (sorties) qui sont « Gestion », « Respect » et « Technique », la classe majoritaire était « la catégorie Gestion », et les deux autres catégories étaient des classes minoritaires (Figure 29) :

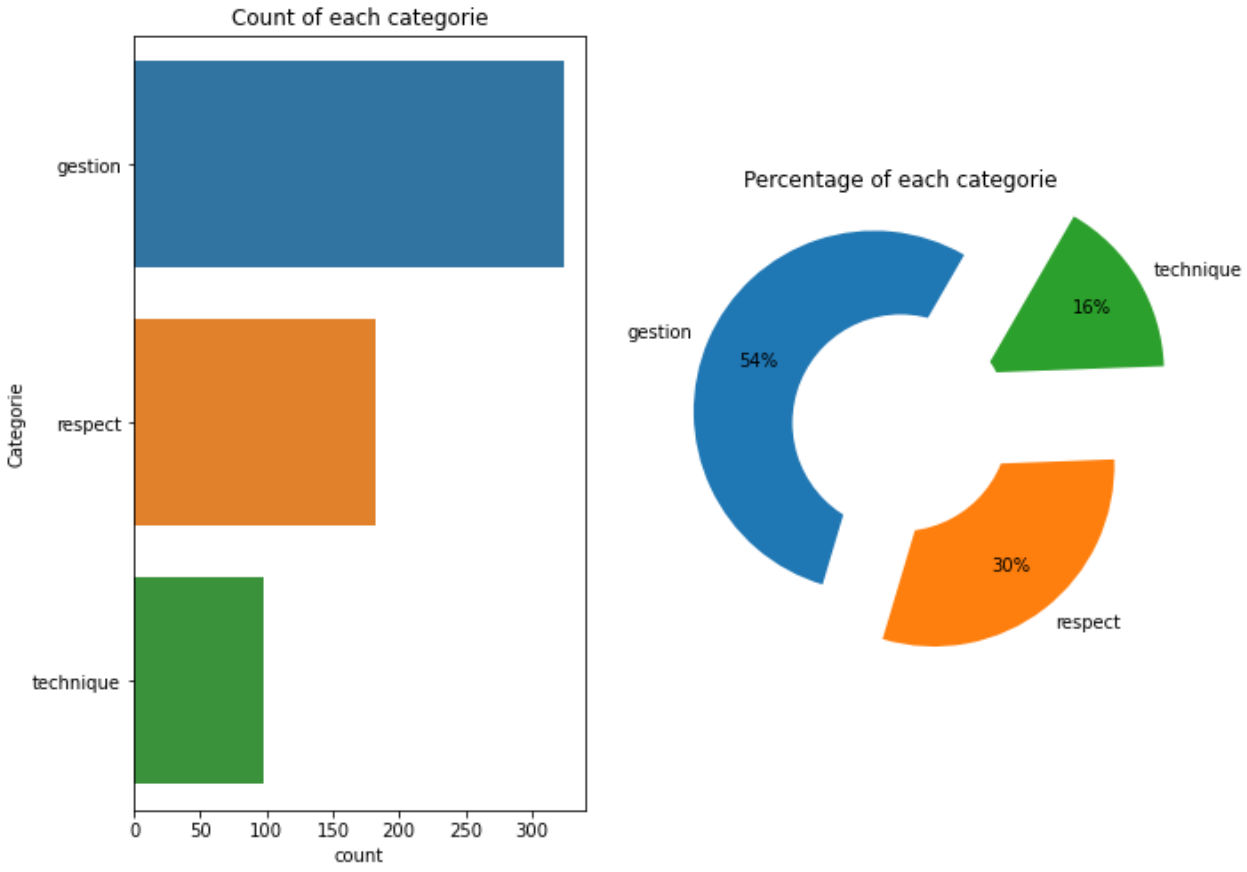


Figure 29. Pourcentage des catégories

Nous avons remarqué que la catégorie "Gestion" peut être divisée en trois catégories supplémentaires comme suit:

1. Nous considérons la catégorie des réclamations comme **une catégorie de Gestion** s'il y a une erreur administrative, par exemple « L'agence commerciale est fermée »
2. Nous considérons la catégorie des réclamations comme **une catégorie de Métier** lorsqu'il y a un problème dans la profession des agents en termes de qualité et de fabrication, par exemple « Des informations incorrectes sur la carte d'abonnement ».
3. Nous considérons la catégorie des réclamations comme **une catégorie de Sécurité** s'il y a quelques choses qui met les clients en danger, par exemple « Le fil électrique est trop près de l'eau ».

Et la catégorie "Respect" peut être divisée en deux catégories :

4. Nous considérons la catégorie des réclamations comme **une catégorie de Respect** en cas de non-respect des règles de l'entreprise soit par les agents ou les clients.

5. Nous considérons la catégorie des réclamations comme **une catégorie de Comportement** lorsqu'il y a un problème avec le comportement des agents, par exemple « L'agent ne porte pas le masque».

Aussi, Nous considérons la catégorie des réclamations comme **une catégorie Technique** lorsqu'il y a un problème technique, par exemple « La machine ne fonctionne pas».

Après cette division, nous obtenons trois nouvelles catégories (Sécurité, Métier et Comportement), donc nous obtenons un dataset équilibrée contient des catégories symétriques comme montre la Figure 30 :

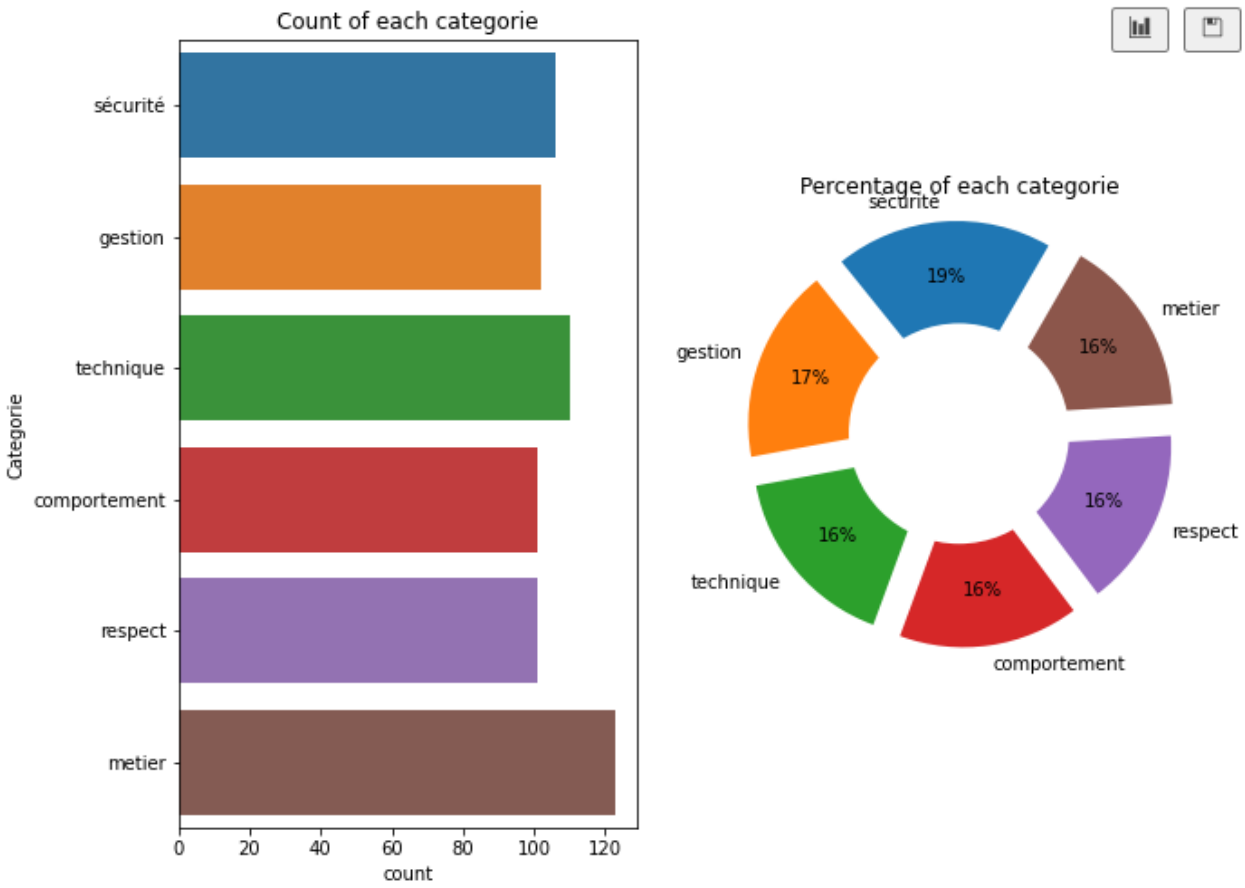


Figure 30. Pourcentage des nouvelles catégories

3.1.3 Nettoyage des données

Cette étape est l'une des étapes de NLP, Il s'agit de trouver les valeurs nulles, les valeurs manquantes et les données dupliquées dans chaque ligne de le dataset, si certaines lignes ont des données manquantes dans plusieurs colonnes importantes, nous pouvons les supprimer.

3.1.4 Fractionnement des données

Dans le cas d'apprentissage supervisé, il ne faut pas entrainer et tester le modèle sur les mêmes données. Le système doit être tester sur des données qu'il n'a pas encore rencontrées pour tester s'il a bien généralisé à partir des données qu'il a vu déjà (Figure 31).

Donc, on a besoin de diviser notre ensemble de données sur deux sous-ensembles :

1. **Training Data :** Données d'entrainement avec une majorité des échantillons (70-80%)
2. **Test Data :** Données de test avec une minorité des échantillons (30-20%) **[31]**

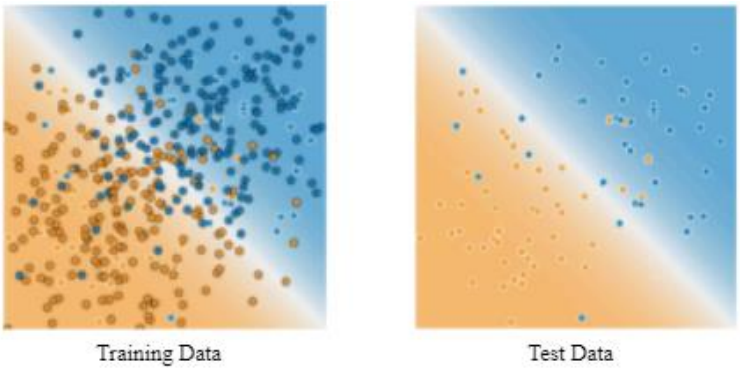


Figure 31. Exemple de fractionnement de données **[31]**

3.1.5 Mise en forme et transformation des données

Comme nous avons mentionné, les algorithmes de Machine Learning ne peuvent pas fonctionner directement avec des données catégorielles, donc nous devons les convertir en données numériques en utilisant l'approche **TFIDF**.

3.1.5.1 Données catégorielles

Les données catégorielles sont des variables qui contiennent des valeurs d'étiquettes plutôt que des valeurs numériques.

Par exemple, la colonne « Réclamation » contient des phrases (Table 1) :

Réclamation
La porte de l'armoire électrique ouverte
Présente des fiches sur la façade latérale du kiosque
Verre de DAT fissuré
Le kiosque était fermé â 7h 09
L'un des DAT n'était pas fonctionnelle
Ram 122 le valideur n'était pas fonctionnelle
Un valideur de la ram 119 wagon 4 n'était pas fonctionnelle
Présent autocollant sur l'armoire électrique
L'absence de l'agent de kiosque

Table 1. La colonne « Réclamation »

3.1.5.2 TF IDF

Il s'agit d'un algorithme très commun pour transformer le texte en une représentation significative des nombres qui est utilisé pour ajuster l'algorithme de la machine pour la prédiction, sa formule est :

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

Avec :

- d : document.
- t : fréquence des termes
- TF (t,d) : $\frac{Nombre\ de\ fois\ que\ t\ apparait\ dand\ d}{Nombre\ Total\ de\ t\ dans\ d}$
- IDF (t) : $log\ \frac{N}{1+df}$ **[31]**

3.2 Etape 02 : Choisir le bon modèle de Machine Learning

Nous avons déjà expliqué dans le Chapitre 01 les modèles utilisés dans l'apprentissage supervisé par classification. Pour notre problème nous n'avons pas de probabilités donc nous éliminons le Naïve Bayas, le choix doit se faire entre la régression logistique et le SVM.

3.2.1 Utilisation de SVM par rapport à la régression logistique

Dans ce projet, nous avons testé les deux modèles -la régression logistique et le SVM - et nous avons trouvé que le SVM est le bon choix pour notre problème.

3.2.2 Kernels

Il s'agit de choisir le bon kernel qui fait la bonne transformation des données sous la forme requise et donne des bons résultats.

Après les tests que nous avons réalisés, nous avons trouvé que choisir un noyau RBF en donnant les valeurs 3 et 0.3 à la constante C et à gamma respectivement , donne des meilleurs résultats.

3.2.3 Classification multi classe à l'aide de SVM

Après les tests que nous avons faits, nous avons trouvé que l'approche OVO, est le bon choix pour notre travail contrairement à l'approche OVA, qui donne des résultats inférieurs aux résultats de l'approche OVO.

Toutes ces étapes nous ont permis de construire un modèle de ML, son fonctionnement est montré par la figure suivant (Figure 32) :

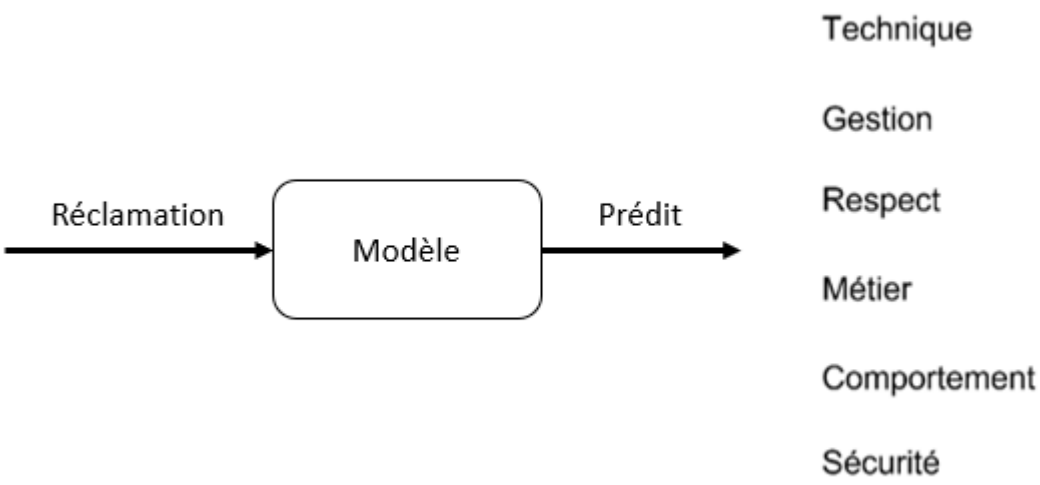


Figure 32. Fonctionnement de modèle de ML.

3.3 Evaluation du Performance

Il est important d'établir la qualité du modèle utilisé, pour cela, diverses mesures d'évaluation peuvent être utilisées et choisies soigneusement, puisque le choix de la mesure peut influencer la manière dont la performance est évaluée et interprétée.

L'une des manières les plus répandues pour mesurer la performance d'un modèle de classification est **la matrice de confusion**, cette dernière correspond à un résumé tabulaire du nombre de prédictions correctes et non correctes, faites par le modèle. **[33]**

3.3.1 Matrice de confusion

La matrice de confusion est un tableau croisé qui enregistre le nombre d'occurrences entre deux évaluateurs, la classification réelle/effective et la classification prédite (Figure 33). **[33]**

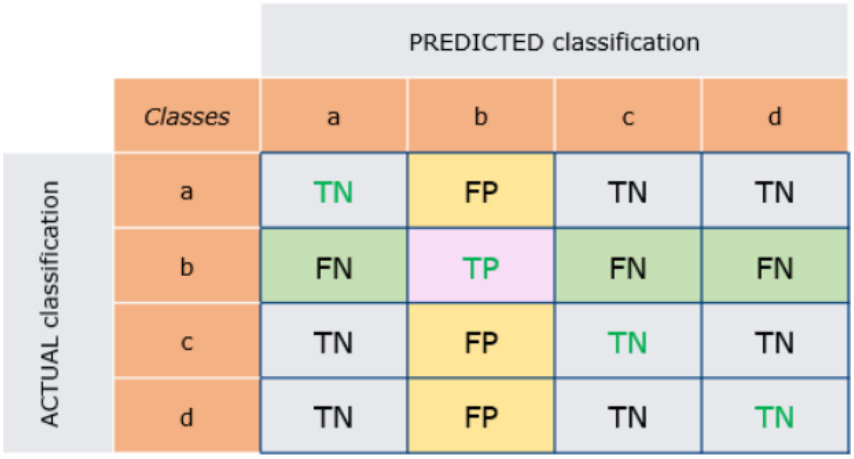


Figure 33. Matrice de confusion pour une classification multi classes **[33]**

Avec :

- **True Negative (TN) :** Vrais négatifs, soit lorsque la classe réelle et la classe prédite sont toutes les deux négatives.
- **True Positive (TP) :** Vrais positif, soit lorsque la classe réelle et la classe estimée prédite sont toutes les deux positives.
- **False Positive (FP) :** Faux positifs, soit lorsque la classe réelle est négative mais que la classe prédite est positive, on appelle ceci une erreur de Type 1.
- **False Negative (FN) :** Faux négatifs, soit lorsque la classe réelle est positive mais que la classe prédite est négative, on appelle ceci une erreur de Type 2.

Lorsque nous passons d'une classe à une autre, nous calculons à nouveau les quantités et les étiquettes des tuiles de la matrice de confusion sont modifiées en conséquence. **[33]**

3.3.2 Mesures de classification

Une fois que la matrice de confusion a été établit, elle peut être utilisée pour des mesures plus approfondies afin d'obtenir une meilleure évaluation de la qualité du modèle. **[34]** Parmi les mesures de classification, on trouve :

3.3.2.1 Accuracy

Accuracy correspond au nombre de prédictions correctes faites par le modèle.

Ceci peut être calculé en utilisant les valeurs de la matrice de confusion et en utilisant la formule suivante :

$$Accuracy_{\mathrm{k}} = \frac{TP_{\mathrm{k}} + TN_{\mathrm{k}}}{TP_{\mathrm{k}} + FP_{\mathrm{k}} + FN_{\mathrm{k}} + TN_{\mathrm{k}}}$$

Avec :

- k : numéro de la classe.

Cette mesure est utilisée lorsque le nombre de TP_{k} et de TN_{k} sont les plus important. **[34]**

3.3.2.2 Précision

La précision correspond au nombre d'éléments corrects rendus par le modèle, elle peut être calculée avec la formule suivante :

$$Précision_{\mathrm{k}} = \frac{TP_{\mathrm{k}}}{TP_{\mathrm{k}}+FP_{\mathrm{k}}}$$ **[34]**

3.3.2.3 Recall (Rappel)

Il détermine la proportion des valeurs positives qui ont été prédites avec précision, On peut utiliser la formule suivante pour le calculé :

$$Recall_{\mathrm{k}} = \frac{TP_{\mathrm{k}}}{TP_{\mathrm{k}}+FN_{\mathrm{k}}} = 1 - Typre\ 2\ error$$ **[34]**

3.3.2.4 Specificiy (Spécificité)

La spécificité correspond au nombre de classes négatives prédites par le modèle **[34]**, elle peut se calculer de la manière suivante :

$$Specificity_{\mathrm{k}} = \frac{TN_{\mathrm{k}}}{TN_{\mathrm{k}}+FP_{\mathrm{k}}} = 1 - Typre\ 1\ error$$ **[34]**

3.3.2.5 F1 Score

Le F1 Score correspond à une combinaison des mesures de recall et de précision, il peut être calculée avec la formule suivante :

$$F1\ Score_{\mathrm{k}} = 2 \times \frac{Recall_{\mathrm{k}} * Précision_{\mathrm{k}}}{Recall_{\mathrm{k}}+Précision_{\mathrm{k}}}$$ **[34]**

4. Conclusion

Au cours de ce chapitre, nous avons présenté l'application RightNow By Brenco et ses objectifs, et les étapes nécessaires afin de réduire notre problème. Dans le chapitre suivant, nous parlerons des outils utilisés, nous discuterons les résultats obtenus et nous évaluerons la performance du modèle, puis nous parlerons sur un travail similaire au notre.

Chapitre 03 : Implémentation

1. Introduction

Dans ce chapitre, on va présenter tout d'abord l'ensemble de outils pour réaliser nos expériences et les résultats expérimentaux, on va parler sur le langage, les bibliothèques et l'éditeur utilisés. Ensuite, on va présenter les résultats obtenus et les discuter. Enfin, nous comparons nos résultats avec un travail similaire.

2. Outils et Langage

Nous avons réalisé nos expériences à l'aide de :

2.1 Python

Python est un langage qui peut s'utiliser dans nombreux contextes et s'adapter aussi à tout type d'utilisation grâce à des bibliothèques spécialisées à chaque traitement. Il est particulièrement utilisé comme un langage de script pour automatiser des tâches simples mais fastidieuses par exemple un script qui récupérerait la météo sur internet ou qui s'intégrerait dans un logiciel de conception assistée par ordinateur afin d'automatiser certains enchaînements d'actions répétitives. On l'utilise comme un langage de développement de prototype lorsqu'on a besoin d'une application fonctionnelle avant de l'optimiser avec un langage de plus bas niveau. **[35]**

2.1.2 L'utilité du Python en Machine Learning

Ce langage de programmation présente des nombreuses caractéristiques intéressantes comme :

- Il est multiplateforme C'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation : Windows, Linux, Android, iOS, mac os x, depuis les mini-Ordinateurs Raspberry Pi jusqu'aux supercalculateurs.
- C'est un langage interprété, Un script Python n'a pas besoin d'être compilé pour être Exécuté, Il convient bien à des scripts d'une dizaine de lignes qu'à des projets complexes de plusieurs dizaines de milliers de lignes.
- Il est gratuit. Vous pouvez l'installer sur autant d'ordinateurs que vous voulez.

C'est un langage de haut niveau, Il demande relativement peu de connaissance sur le

Fonctionnement d'un ordinateur pour être utilisé . **[36]**

2.3 VS Code

Visual Studio Code (VS) est un éditeur de code open-source développé par Microsoft supportant un très grand nombre de langages grâce à des extensions. Il supporte l'autocomplétions, la coloration syntaxique, le débogage, et les commandes git. **[37]**

2.4 Scikit-Learn

Scikit-learn, aussi appelé sklearn, est la bibliothèque la plus robuste et la plus puissante pour ML en Python. Elle fournit une sélection d'outils efficaces pour Machine Learning et la modélisation statistique, surtout pour la classification, la régression et le clustering via une interface cohérente en Python. Cette bibliothèque, qui est en grande partie écrite en Python, s'appuie sur NumPy, SciPy et Matplotlib. forçes de Scikit learn sont :

- Licence BSD : il existe une restriction minimale sur l'utilisation et la distribution du logiciel.
- Facile à utiliser : la popularité de Scikit-learn est due à la facilité d'utilisation qu'il offre.
- Documentation détaillée : Il propose également une documentation détaillée de l'API à laquelle les utilisateurs peuvent accéder à tout moment sur le site Web, ce qui les aide à intégrer l'apprentissage automatique dans leurs propres plateformes.
- Utilisation intensive dans l'industrie : Scikit-learn est largement utilisé par diverses organisations pour prédire le comportement des consommateurs, identifier les activités suspectes, et bien plus encore.
- Organigramme des algorithmes : Scikit-learn dispose d'une antisèche ou d'un organigramme des algorithmes pour aider les utilisateurs.
- Algorithmes d'apprentissage automatique : Scikit-learn couvre la plupart des algorithmes d'apprentissage automatique via un énorme soutien communautaire. **[38]**

2.5 Pandas

La bibliothèque logicielle open-source Pandas est spécifiquement conçue pour la manipulation et l'analyse de données en langage Python. Elle est à la fois flexible, performance et simple d'utilisation . Grâce à Pandas, le langage Python permet enfin

de charger, d'aligner, de manipuler et de fusionner des données. Les performances sont particulièrement impressionnantes quand le code source backend est écrit en C ou en Python**[39]**

2.6 Numpy

Numpy est une bibliothèque mathématique permettant d'implémenter de façon très efficace des calculs standards et de l'algèbre linéaire . Précisons que Pandas est basé sur NumPy. De nombreuses structures de données et fonctionnalités de Pandas proviennent de NumPy. Ces deux bibliothèques sont souvent utilisées conjointement et étroitement liées entre elles . **[40]**

2.7 Matplotlib

Matplotlib est une bibliothèque Python capable de produire des graphes de qualité. Matplotlib peut être utilisé dans des scripts Python, le Shell Python et IPYthon, le notebook Jupyter , des serveurs d'application web et aussi dans quatre outils d'interface graphique.

Matplotlib essai de rendre les tâches simples "simples" et de rendre possible les choses compliqués. Vous pouvez générer des graphes, histogrammes, des graphiques à barres, des graphiques d'erreur, des spectres puissance (lié à la transformée de Fourier), des nuages de dispersion etc… en quelques lignes de code **[41]**

2.8 Django

Django est un framework Web avancé écrit en Python qui utilise le modèle architectural du contrôleur de vue de modèle. Son objectif principal est de faciliter le développement de sites Web complexes et basés sur des bases de données. Django est disponible en tant que framework Web open source et utilise largement Python pour créer des fichiers, des paramètres et des modèles de données. **[42]**

2.9 HTML

HyperText Markup Language (HTML) est un langage de balisage pour le Web qui définit la structure des pages Web, avec :

- **Hypertexte :** texte (souvent avec des intégrations telles que des images également) qui est organisé afin de connecter des éléments connexes

- **Markup :** un guide de style pour la composition de tout ce qui doit être imprimé au format papier ou électronique
- **Language :** un langage qu'un système informatique comprend et utilise pour interpréter les commandes. **[43]**

3. Tests et Résultats

Tout D'abord, nous présentons les résultats des étapes dont nous avons parlé dans le Chapitre 02 :

3.1 Dataset

Nous avons considéré dans notre travail la colonne « Réclamation » et la colonne « Catégories » qui ont une relation avec notre objectif, donc notre dataset est comme montre la Figure 34 :

My New Dataset :

	Reclamation	Categorie
0	Verre de DAT fissuré	sécurité
1	Le kioosque était fermé â 7h 09	gestion
2	DAT fessuré	sécurité
3	L'un des DAT N'etait pas fonctionelle	technique
4	présent autocolant sur l'armoire electrique	gestion
..	...	...
638	Certains agents commettent des actes violents ...	respect
639	l'agent ne respecte pas les emplois du temps.	respect
640	Il y a des policiers qui sont impliqués dans d...	respect
641	Certains officiers sont mêlés à des actes de v...	respect
642	L'agent de la cabine est dur avec le client.	respect

Figure 34. Résultat du Dataset

3.2 Nettoyage de données

Nous avons nettoyé notre dataset des valeurs nulles, les valeurs manquantes et les données dupliquées (Figure 35) :

```
Check if there is null data :
 Reclamation    0
Categorie       0
```

Figure 35. Résultat de nettoyage de données

3.3 Fractionnement de données

Nous avons divisé notre dataset sur Training data et Test data avec un pourcentage de 70 % pour « Training Data » comme suit :

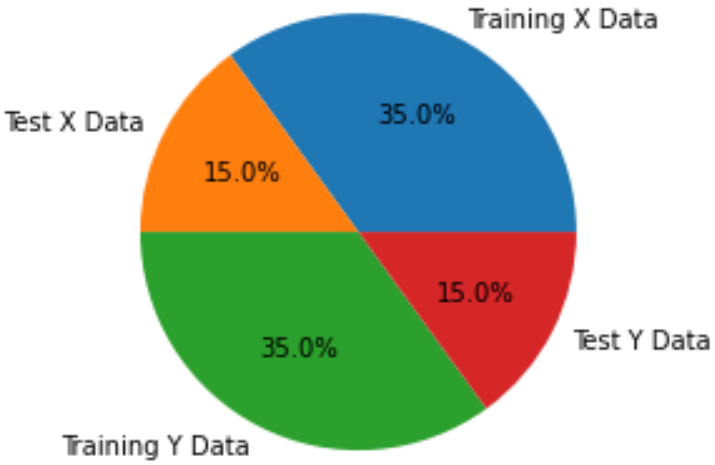


Figure 36. Résultat de fractionnement de données

3.4 Mise en forme et transformation des données

Nous avons utilisé le TF IDF Vectorizer pour la transformation de la colonne « Reclamation » en une représentation significative des nombres :

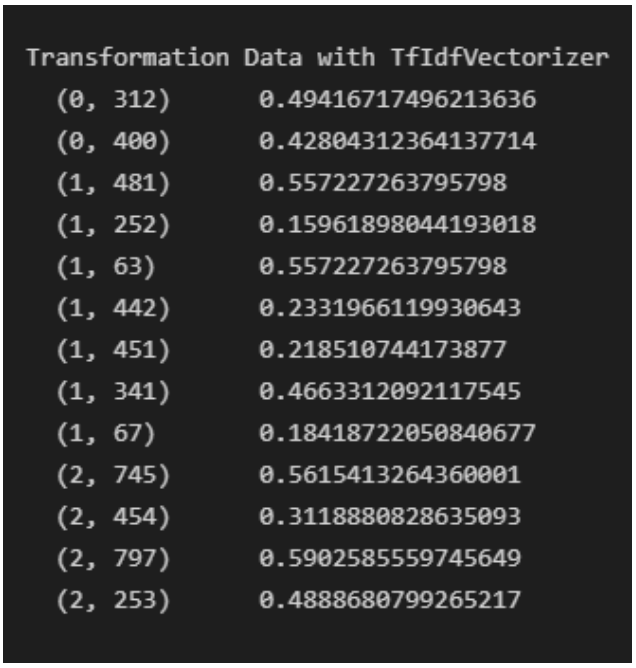


Figure 37. Résultat de transformation de données en utilisant le TF IDF Vectorizer

Après avoir mis en place notre modèle pour le training et les tests, nous évaluons la performance de notre modèle:

3.5 Matrice de confusion

La Figure 38 suivante montre le résultat de la matrice de confusion :

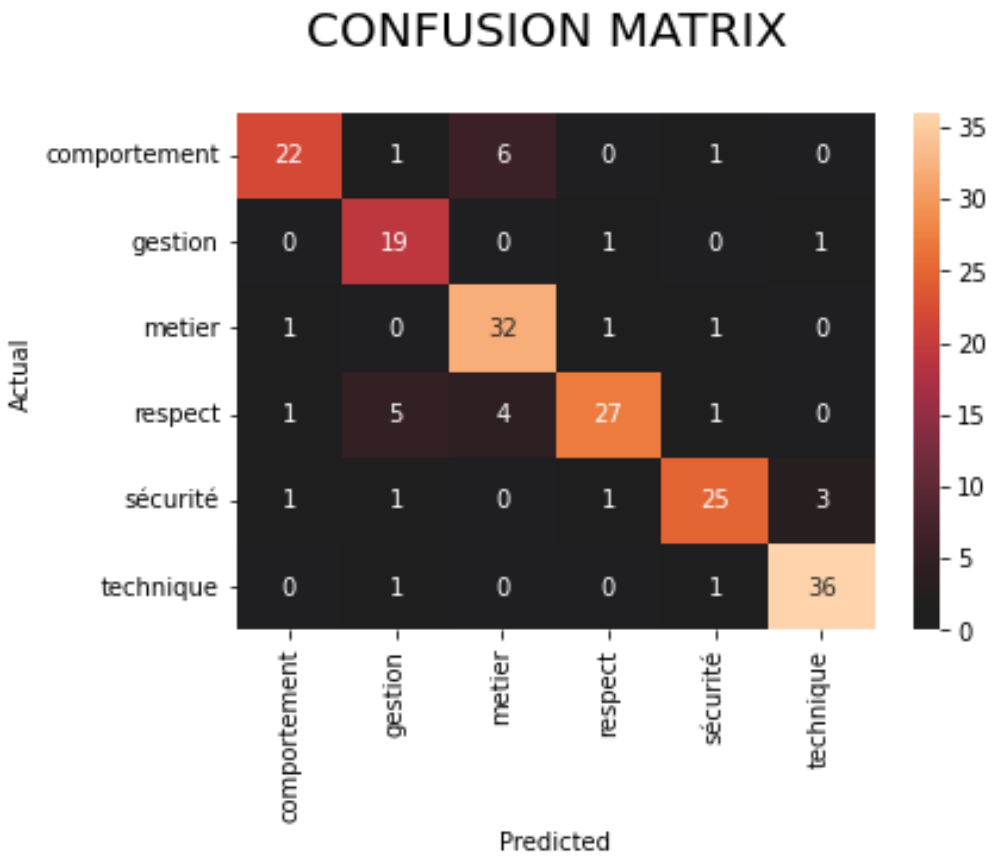


Figure 38. Résultat de la matrice de confusion

La matrice de confusion montre les valeurs de FP, FN qui sont tous les deux inférieurs à la valeur 7, aussi les valeurs de TP qui sont entre 19 et 36. Les valeurs de TN n'apparait pas ici, mais nous les avons obtenues et elles sont entre 148 et 164 (Figure 39) :

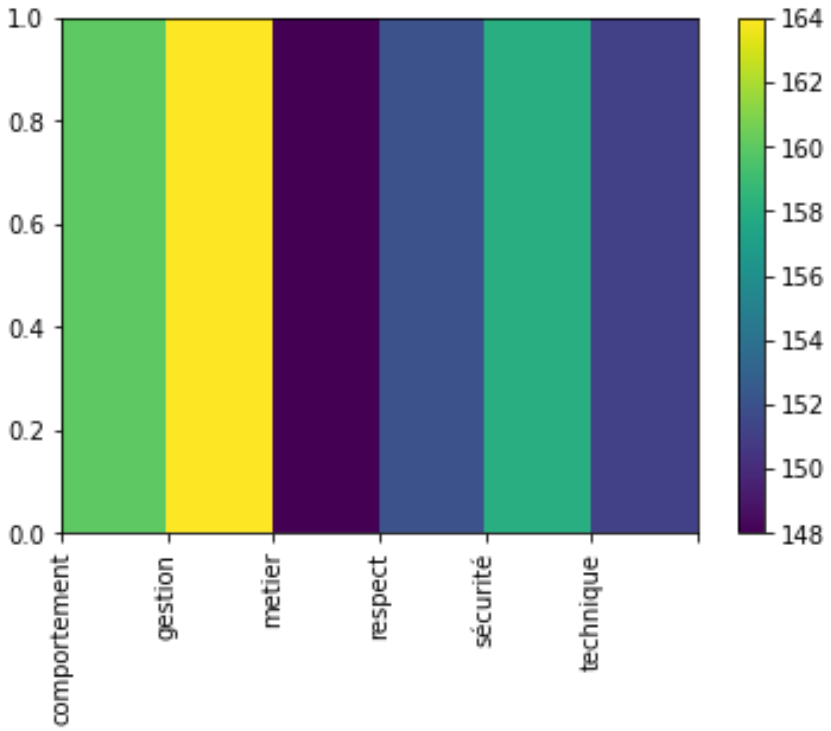


Figure 39. Les valeurs de TN

3.6 Mesure de classification

La table suivante montre les résultats de « Précision », « Recall », « F1 Score », « Sensitivity », « Specificiy » et « Accuracy» de la prédiction du Test Data :

Catégories	Precision	Recall	F1 Score	Specificiy	Accuracy
Comportement	0.88	0.73	0.80	0.98	0.94
Gestion	0.70	0.90	0.79	0.95	0.94
Métier	0.76	0.91	0.83	0.93	0.93
Respect	0.90	0.71	0.79	0.98	0.92
Sécurité	0.86	0.81	0.83	0.97	0.94
Technique	0.90	0.95	0.92	0.97	0.96

Table 2. Résultats de la mesure de classification

Nous avons tracé le graphe correspondant à ces résultats (Figure 40) :

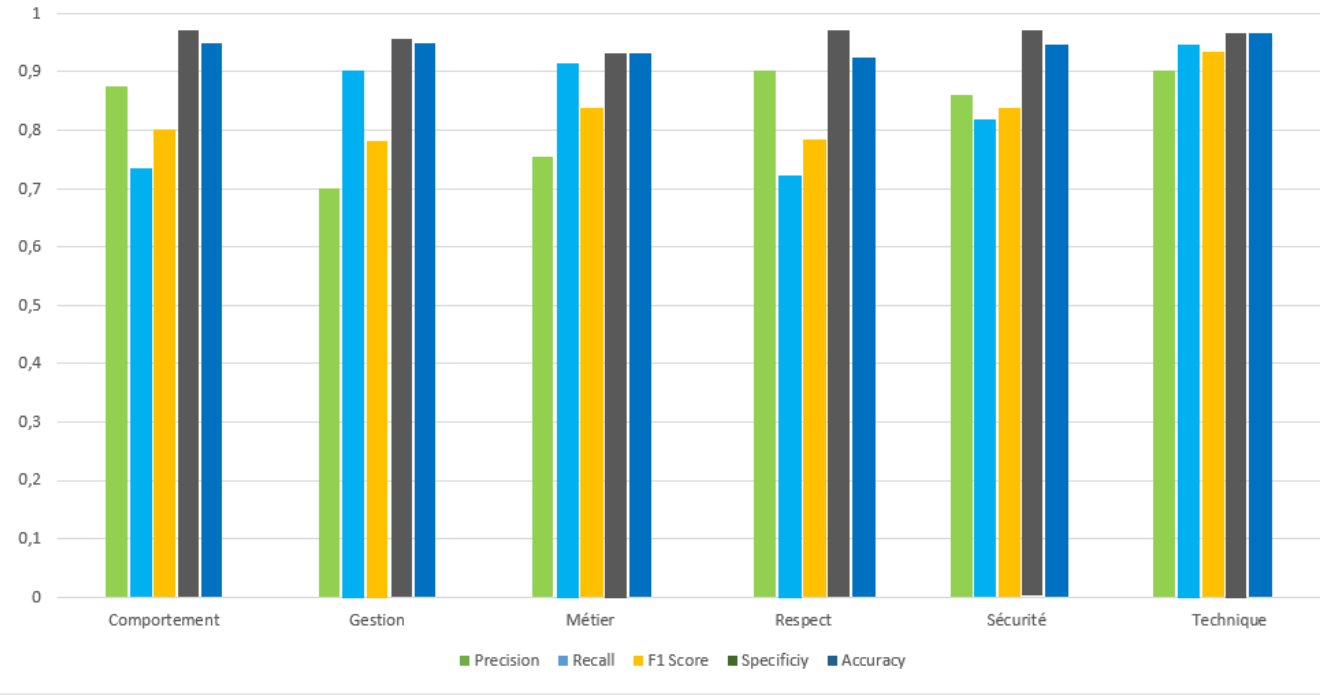


Figure 40. Graphe du résultat de la mesure de classification

Nous remarquons que les résultats sont tout entre 70% et 98% :

Les valeurs de « Recall » sont toujours entre 0.71 % et 0.94 %, ces valeurs présentent la proportion de TP qui sont correctement identifiés dans chaque catégorie, aussi les valeurs de « Specificiy » sont toujours supérieures à 0.90 %, ces valeurs présentent la proportion de TN qui sont correctement identifiés dans chaque catégorie. « Accuracy » aussi donne des bonnes valeurs, elles sont entre 0.92 % et 0.96 %, chaque valeur représente la proportion de vrais résultats positifs (à la fois TP et TN) dans chaque catégorie. La Précision est définie comme l'exactitude du jugement, nous trouvons que sa valeur est supérieure à 0.70 % dans chaque catégorie et c'est un bon signe. Le F1 score est défini comme la moyenne harmonique entre la précision et le rappel, il est utilisé comme mesure statistique pour évaluer les performances, dans chaque catégorie, nous avons trouvé sa valeur est supérieure à 0.79 %.

Selon ces résultats, notre modèle fonctionne bien.

3.6.1 La valeur de « Accuracy Score »

La prédiction du « Test Data » nous a donné une « Accuracy Score» à 0.834 %, c'est une valeur intéressante.

4. Teste du modèle

Afin de tester notre modèle, nous avons développé une application web simple dans laquelle nous pouvons écrire la réclamation et le modèle prédite sa catégorie et l'afficher (Figure 41) :

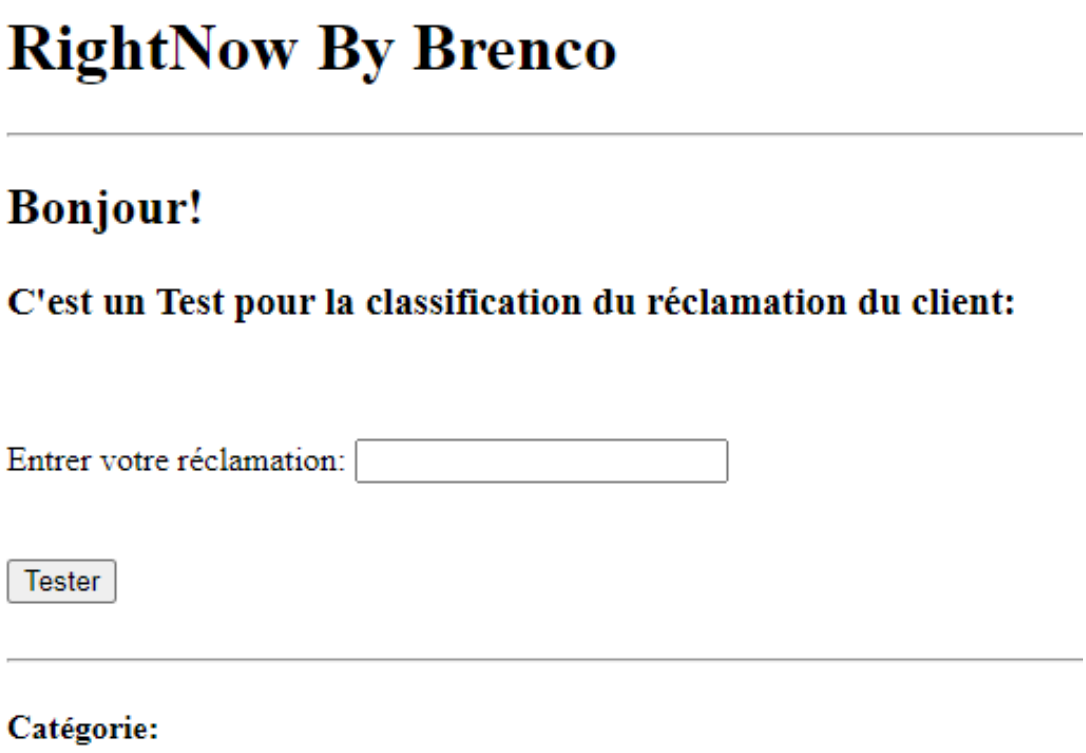


Figure 41. Application web de Test

Voici quelques résultats de tests :

Si la réclamation est « ne porte pas la bavette », donc sa catégorie selon le modèle construit est « comportement » comme montre la figure 42 :

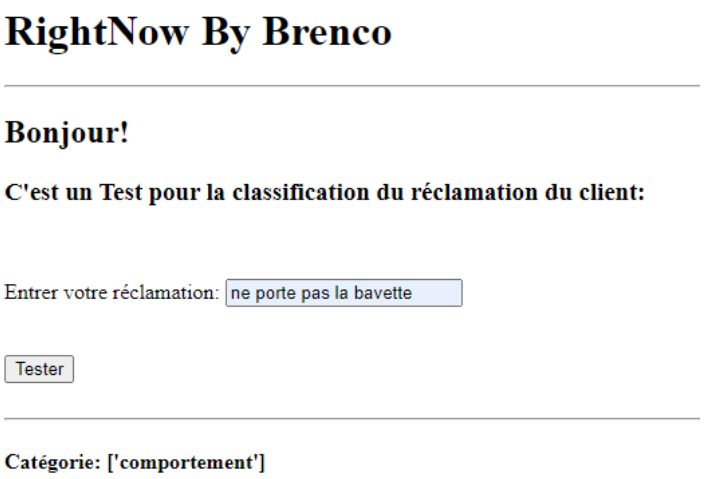


Figure 42. Résultat du Test -Comportement-

Si la réclamation est « Agence est fermé », sa catégorie selon le modèle construit est « gestion » comme montre la figure 43 :

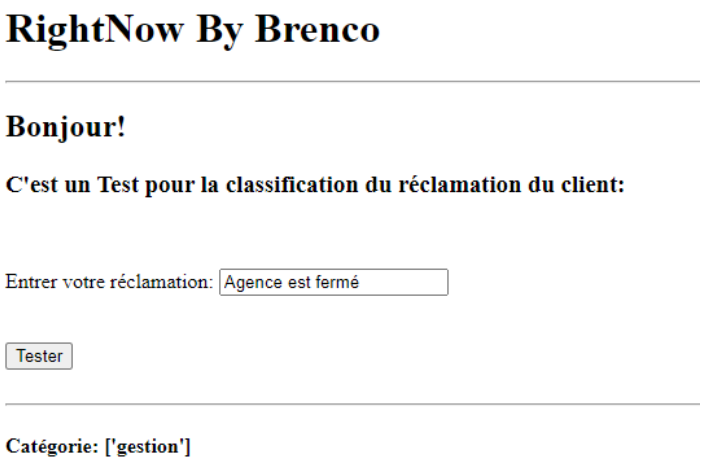


Figure 43. Résultat du Test -Gestion-

D’après la table 2 qui montre les résultats de la mesure de classification, le graphe (Figure 39) et la valeur de « Accuracy Score », nous remarquons que les résultats sont tout entre 70% et 98% :

L'évaluation globale de ce système pour la catégorisation des réclamations du clients a donné un « Accuracy » de 70 à 100% pour les 6 catégories différentes.

En générale, Nous pouvons dire que le modèle qui nous avons construit (Figure 44), prédit les réclamations d'une bonne manière, et comme nous avons dit dans le chapitre 01, les résultats obtenus ne sont pas corrects à 100%, mais avec le temps, la précision des modèles devient plus élevée.

```
32  #get the bath to the new dataset
33  path = 'D:\mydata.txt'
34  #read the data and give the name to each column
35  data = pd.read_csv(path, header=None,names=['Reclamation', 'Categorie'],sep
36  X = data['Reclamation']
37  y = data['Categorie']
38  #separate data to training data and testig data
39  X_train, X_test, y_train,y_test = train_test_split(X, y, test_size=0.30, ra
40  #Convert a collection of raw documents to a matrix of TF-IDF features.
41  vectorizer = TfidfVectorizer()
42  #Choose the best Model, C and gamma Values
43  model= svm.SVC(kernel='rbf', gamma = 0.3, C= 3)
44  #Choose the OVO
45  ovo =  OneVsOneClassifier(model)
46  #Work with Pipeline to enabling data to be transformed and correlated into
47  text_clf = Pipeline([('tfidf', vectorizer),
48                       ('clf', ovo)])
49  #Fit the Model
50  clf =text_clf.fit(X_train, y_train)
51
```

Figure 44. Capture du code - Modèle-

Et pour le code de l'application web :

```python
from django.shortcuts import render
from joblib import load
model = load('./savedModels/model.joblib')
# Create your views here.

def myView (request):
    if request.method == 'POST':
        y_pred = model.predict([request.POST['reclamation']])
        print (y_pred)
        return render(request, 'Test.html', {'result': y_pred})
    else :

        return render(request, 'Test.html')
```

Figure 45. Capture du code -Application Web-

5. Travaux connexes

De nombreux travaux ont porté sur le suivi des catégorisations des réclamations de clients, l'un d'eux est le « Journal of Intelligent Systems: Theory and Applications » qui a publié un article en 2022, intitulé « Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches », écrit par Fatma BOZYİĞİT, Onur DOĞAN et Deniz KILINÇ. Cet article contribue à développer une catégorisation précise des réclamations des clients concernant les produits alimentaires emballés écrites en turc.

Dans l'étude de cet article, les réclamations des clients concernant les produits alimentaires livrés sur les marchés turcs ont été classées en cinq catégories telles que « Hygiene », « Foreign body », « Taste/Smell », « Texture »,et « Package/Label ».

En conséquence, ils ont utilisé des divers algorithmes ML utilisant les stratégies de représentation des caractéristiques Term Frequency-Inverse Document Frequency (TF-IDF) et word2vec ont été exécutés pour déterminer la catégorie de réclamations. Les résultats correspondants des classificateurs de régression linéaire (LR), Naive Bayes (NB), k Nearest Neighbor (kNN), Support Vector Machine (SVM), Random Forest (RF) et Extreme Gradient Boosting (XGBoost) **[44]** comme montre la Figure 46 :

ML algorithm	Feature representations		Precision	Recall	F-measure
LR	TF-IDF		0.80	0.83	0.81
	word2vec	skip-gram	0.66	0.64	0.67
		CBOW	0.54	0.48	0.46
NB	TF-IDF		0.75	0.71	0.73
	word2vec	skip-gram	0.53	0.54	0.53
		CBOW	0.62	0.63	0.62
kNN	TF-IDF		0.80	0.82	0.81
	word2vec	skip-gram	0.56	0.58	0.57
		CBOW	0.57	0.62	0.59
SVM	TF-IDF		0.81	0.81	0.81
	word2vec	skip-gram	0.69	0.74	0.71
		CBOW	0.59	0.66	0.62
RF	TF-IDF		0.82	0.81	0.81
	word2vec	skip-gram	0.53	0.54	0.53
		CBOW	0.62	0.63	0.62
XGBoost	**TF-IDF**		**0.83**	**0.84**	**0.84**
	word2vec	skip-gram	0.62	0.67	0.64
		CBOW	0.76	0.75	0.75

Figure 46. Évaluation de différents algorithmes ML et représentations de caractéristiques **[44]**

Leurs résultats expérimentaux montrent que la méthode la plus performante est XGBoost avec le schéma de pondération TF-IDF et qu'elle atteint un score de mesure de %86. L'autre point important est que les classificateurs ML basés sur word2vec affichent des performances médiocres en termes de mesure F par rapport au schéma de pondération des termes TF-IDF. Ils observent également que chaque algorithme ML basé sur TF-IDF expérimenté donne une meilleure performance de prédiction sur les sous-ensembles optimaux de caractéristiques sélectionnés par la méthode Chi Square (CH2). L'exécution de CH2 sur les fonctionnalités TF-IDF augmente le score de mesure F de 86 % à 88 % dans XGBoost.**[44]**

La table suivante montre une comparaison entre nos résultats et leurs résultats :

Comparaison	Modèle du notre travail	Modèle du travail connexe
Nombre de catégories	Six : Comportement, Gestion, Technique, Sécurité, Métier et Respect.	Cinq : Hygiene, Foreign body, Taste/Smell, Texture, et Package/Label.

Méthode	Algorithme supervisé par classification multi classes	Algorithme supervisé par classification multi classes
Modèle	SVM	XGBoost
Représentation des caractéristiques	TF IDF	TF IDF
Précision	0.83 %	0.83 %
Recall	0.83 %	0.84 %
F1 Score	0.82 %	0.84 %
Accuracy	0.93 %	0.88 %

Table 3. Comparaison entre les résultats du deux travails

Nous utilisons tous les deux l'apprentissage supervisé par classification multi classes, aussi le même caractéristique TF IDF, la différence est dans le modèle utilisé, où nous avons utilisé le SVM et ils ont utilisé le XGBoost. Nous avons calculé la valeur de « Specificity » par rapport à eux ». Les résultats de « Précision », « Recall » et « F1 Sore » sont presque similaires, sauf la valeur de « Accuracy », notre valeur est grande par rapport à leur valeur, sinon les résultats des deux modèles sont acceptables.

6. Conclusion

Nous avons présenté à travers ce dernier chapitre les outils et les langages que nous avons utilisés pour la réalisation du modèle et de l'application web, ainsi que les résultats de toutes les étapes que nous l'avons parlé dans le 2ème chapitre, aussi nous avons présenté les résultats du test, puis nous avons parlé sur un travail similaire au notre dans la dernière section .

Conclusion Générale

La satisfaction du client est spécifiée comme un facteur primordial pour le succès d'une entreprise selon la théorie de base du marketing. Les avantages de la catégorisation automatique des réclamations sont la réduction du coût initial de l'étiquetage des réclamations avec l'étiquette la plus appropriée, la maintenance d'un processus efficace pour diriger les réclamations des clients vers les services concernés, et la suppression du risque de dépendre d'experts dans la gestion des réclamations des clients.

Dans ce sens, les réclamations des clients sont analysées pour déterminer les problèmes éventuels et les stratégies efficaces pour les gérer. Cette analyse peut être une tâche difficile pour un humain, car il peut être nécessaire d'analyser des données à volume élevé pendant de longues périodes. Une alternative consiste à la catégorisation automatiquement des causes d'insatisfaction des clients.

Ce mémoire était dédiée à la construction d'un modèle de Machine Learning qui donne une prédiction correcte que possible de réclamations des clients de la société SETRAM, un des clients de RightNow By Brenco, qui veut trouver une solution pour son problème: mauvais choix des catégories de réclamations envoyées par ses clients à travers des formulaires rend ses statistiques inexacte, donc elle se trouve obligée de recatégoriser manuellement ces réclamations, ce qui prend beaucoup de temps et d’effort, afin d'obtenir des statistiques précises.

Nos contributions sont classées comme suit :

D'abord, nous avons préparé notre dataset par l’ajout des nouvelles catégories, le nettoyage, le fractionnement, la mise en formes et transformation du donnée.

Ensuite, nous avons fait le choix du modèle entre la régression logistique et le SVM et le choix d’approche entre l’OVO et l’OVA, en les testant tous les deux par l’évaluation de la performance du modèle, et nous avons trouvé que le SVM et l’approche OVO fonctionne mieux que la régression logistique et l’approche OVA respectivement.

Enfin, nous avons fait des tests dans une application web que nous l'avons développé afin de tester le fonctionnement du modèle.

En conséquence, les réclamations doivent être classées en catégories «Gestion », « Technique », « Comportement», « Respect » et « Sécurité » et « Métier ».

Comme nous avons mentionné précédemment, les réclamations des clients sont l'une des facteurs les plus importants qui déterminent la dynamique du marché du développement de produits et les contributions de ce mémoire peuvent conduire à de nouvelles voies de réussite

pour les entreprises, l'augmentation d'automatisation des processus mis en place, réduire l’apport humain, gagner en productivité, gagner le temps de traitement des requêtes et augmenter le nombre de requêtes traitées quotidiennement.

La science n'est jamais complète et les idées sont sans fin c'est pour cela nous continuerons à générer de nouvelles idées pour développer ce modèle afin qu'il soit un modèle idéal., nous sommes impatients de catégoriser tous les types de réclamations, qu'il s'agisse de réclamations textuelles, de réclamations audios ou même sous forme de vidéo ou d'image, en plus de répondre automatiquement à ces réclamations, bien sûr, tout cela en utilisant Machine Learning.

Références

[1] Andreas C. Müller, Sarah Guido, Introduction to Machine Learning with Python a Guide for Data Scientists, (2017).

[2] Jean-Louis Laurière, Intelligence artificielle Résolution de problèmes par l'homme et la machine, (1987).

[3] Boris Barraud, L'intélligence artificielle Dans toutes ses dimensions, (2018).

[4] Ludwig Hervé, segmentation visiteurs, (13 décembre 2017 à 10h12). BDM blog

[5] Laurence Moroney, Ai and machine learning for coders a programmers guide to artificial intelligence, (2021).

[6] *Maggioli Developers* [online]. https://www.developersmaggioli.it/blog/machine-learningla-scienza-delle-decisioni-automatiche/, [consulté le 05 Avril 2022].

[7] Andriy Burkov, The Hundred-Page machine learning book, (2019)

[8] Andrew Ng, Machine Learning. *Coursera*. (2022). https://www.coursera.org/learn/machinelearning, [consulté le 02 Mars 2022].

[9] Sonoo Jaiswal, *Javatpoint* [online]. https://www.javatpoint.com/linear-regression, [consulté le 22 Avril 2022].

[10] Shubham-Bansal, Supervised and Unsupervised, https://www.geeksforgeeks.org/supervised-unsupervised-learning/, [consulté le 23 Avril 2022].

[11] Sonoo Jaiswal, *Javatpoint* [online]. https://www.javatpoint.com/unsupervised-machine-learning, [consulté le 23 Avril 2022].

[12] Ta-ying Cheng, Supervised, Semi-Supervised, Unsupervised, and Self-Supervised Learning, (2021)

[13] Dan Lee, Reinforcement Learning, Part 1: A Brief Introduction**,** (2019)

[14] Chris Parsons, What Is a Machine Learning Model?, NVIDIA Blogs, (2021)

[15] Apache Spark™, Delta Lake, MLflow, *Databricks* [online]. https://databricks.com/glossary/machine-learning-models, [consulté le 30 Avril 2022].

[16] *Blogger* [online], milleplateaux, Monday, October 15, 2018, 2:15 AM. http://singaporebusinessintelligence.blogspot.com/2018/10/what-is-automated-machine-learning.html, [consulté le 16 Juin 2022].

[17] Rushikesh Pupale, Support Vector Machines (SVM) — An Overview, (2018)

[18] *ResearchGate* [online]. https://www.researchgate.net/figure/This-figure-shows-the-linear-and-non-linear-SVMfor-2D-dataset-for-text-data-we-have_fig11_332494043, [consulté le 25 Juin 2022].

[19] DataFlair Team, *DataFlair* [online], (2018). https://data-flair.training/blogs/svm-kernel-functions/, [consulté le 26 Juin 2022].

[20] Prateek Bajaj, Creating linear kernel SVM in Python, *GeeksforGeek* [online], (2018)

[21] (2014), https://stats.stackexchange.com/questions/90736/the-difference-of-kernels-insvm, [consulté le 26 Juin 2022].

[22] night_fury1, Major Kernel Functions in Support Vector Machine (SVM), *GeeksforGeek* [online], (2022).

[23] Chirag Goyal, Classification multi classe à l'aide de SVM, (2021)

[24] Tous les modèles de Machine Learning expliqués brièvement, *MONCOACHDATA* [online], (2021)

[25] DataFlair Team, *DataFlair* [online], (2018). https://data-flair.training/blogs/advantages-and-disadvantages-of-machine-learning/, [consulté le 29 Avril 2022].

[26] RUSSELL D. RADTKE, *Sites 4 Students*[online], (2022).

https://sites4students.com/top-10-machine-learning-scientist-courses-to-take-in-2022/, [consulté le 30 Avril 2022].

[27] Rich Montgomery, 6 manières dont les technologies révolutionnent toutes les entreprises, (2017)

[28] Karim Brouri, *Brenco* [online], (2017). https://brenco-algerie.com/, [consulté le 5 Mai 2022].

[29] Karim Brouri, *RightNow By Brenco* [online], (2020). https://rightnow-by-brenco.com/, [consulté le 5 Mai 2022].

[30] *aws* [online]. https://aws.amazon.com/fr/what-is/data-preparation/, , [consulté le 7 Mai 2022]

[31] projeduc, Introduction à l'apprentissage automatique, *GitHub* [online].

[32] Mukesh Chaudhary, *Medium* [online], (2020). https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a, , [consulté le 16 Juin 2022]

[33] Margherita Grandini, Enrico Bagli, Giorgio Visani, METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW, (August 14, 2020). https://arxiv.org/pdf/2008.05756.pdf, , [consulté le 18 Juin 2022]

[34] Amandine Allmang, Apprentissage Supervisé et Classification, Linedata Blog

[35] Noam Bressler, How to Check the Accuracy of Your Machine Learning Model, *deepcheck* [online], (2021).

[36] Programmer en Python - Apprendre la programmation de façon claire, concise et efficace, Auteur : Luciano RAMALHO, (2019)

[37] *Modèle MIT* [online], (2018). https://perso.math.univ-toulouse.fr/fdelebec/files/2018/03/pythonbibliotheques.pdf?fbclid=IwAR0VKbP1tRSXSPtsbU2BAenCjVg_tClBCupB1SgknpVZB4e, [consulté le 18 Juin 2022]

[38] sickit-learn Team, *GitHub* [online]

[39] KAVITA MALI, Pandas: A Hands-On Guide For Beginners, *Analytics Vidhya* [online], (2021)

[40] NumPy community, NumPy User Guide Release 1.22.4, (2022)

[41] Florian Fasmeyer, *HE-Arc* [online], (2016). https://he-arc.github.io/livre-python/matplotlib/index.html, [consulté le 22 Juin 2022]

[42] *techopedia* [online]. https://www.techopedia.com/definition/28227/django, [consulté le 24 Juin 2022]

[43] Kolade Chris, What is HTML – Definition and Meaning of Hypertext Markup Language, *freecodecamp* [online], (2021)

[44] Fatma BOZYİĞİT, Onur DOĞAN, Deniz KILINÇ, Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches, Journal of Intelligent Systems: Theory and Applications, (2022)

[45] *TechTarget*, Ben Lutkevich, [online]. https://www.techtarget.com/searchenterpriseai/definition/natural-languageprocessin%20NLP#:~:text=Natural%20language%20processing%20(NLP)%20is,in%20the%20field%20of%20linguistics, [consulté le 12 Juillet 2022]

[46] *La revur IA*, Walid Chrimni, 12 juin 2022, [online]. https://larevueia.fr/quest-ce-que-le-nlp-natural-language-processing/, [consulté le 12 Juillet 2022]