République Tunisienne
Ministère de l'Enseignement Supérieur

Université de Sfax

Ecole Nationale d'Electronique et des Télécommunications de Sfax

Licence Fondamentale

Option

Télécommunications

Projet Tuteuré

PROJET TUTEURE

Présenté à

L'Ecole Nationale d'Electronique et des Télécommunications de Sfax

En vue de l'obtention du

Licence Fondamentale en STIC

Option : Télécommunications

Par

Kammoun Bilel Chabchoub Mohamed

Implémentation d'une application de détection des comportements de conduite à l'aide des algorithmes de « Machine Learning »

Soutenu le 12 juillet 2021, devant la commission d'examen :

Mr.	**Ouni Tarak**	*Examinateur*
Mr.	**Louati Mahdi**	*Encadrant*

Dédicaces

A mes chers parents

Aucune dédicace ne pourrait exprimer assez profondément ce que je ressens envers vous. Je vous Dirais tout simplement, un grand merci, je vous aime.

A mes chères sœur et frère et à toute ma famille

Vous occupez une place particulière dans mon cœur. Je vous dédie ce travail en vous souhaitant un avenir radieux plein de bonheur et de succès.

A mes chers amis

En souvenir de nos bons moments, en souvenir de tout ce qu'on a vécu ensemble, j'espère de tout mon cœur que notre amitié durera éternellement.

Kammoun Bilel

A mes chers parents

Pour leurs grands sacrifices et leur affection tout au long de mes études. Aucun mot, aucune dédicace ne peut exprimer mes sentiments, que Dieu leur accorde la santé et le bonheur.

A mes sœurs

Pour leur support et leur soutien que je leur dédie ce travail avec tous mes vœux de bonheur et de santé.

A mes chers amis

Pour leurs efforts fournis afin d'accomplir ce travail et à tous ceux que j'aime.

Chabchoub Mohamed

Remerciements

Que toutes les personnes qui nous ont aidé durant l'élaboration de ce travail trouvent dans ces lignes l'expression de notre profonde gratitude.

Dans un premier temps, nous tenons à remercier tout particulièrement et à témoigner toute notre reconnaissance à notre encadrant M. Louati Mahdi pour l'expérience enrichissante et pleine d'intérêts qu'il nous a fait vivre, pour sa pleine générosité en matière de formation et d'encadrement que ce soit durant la période de l'enseignement présentiel ou à distance, pour les encouragements et conseils qu'il nous a prodigués pour le bien de notre formation et notre parcours professionnel plus tard, et pour le temps précieux pour les corrections et les discussions qu'il nous a réservé malgré ses multiples occupations.

Enfin avec un grand respect, nous tenons à remercier l'examinateur Mr **Ouni Tarak**, pour l'honneur qu'il nous fait en acceptant de juger ce travail, en espérant qu'il trouvera dans ce rapport la clarté et la motivation qu'il attend.

Table des matières

Liste des Figures

Liste des tables

Introduction générale

De nos jours, l'intelligence artificielle est un sujet qui questionne beaucoup et fait partie de notre quotidien. A travers notre regard, nous chercherons surtout à dégager une réflexion autour de l'éthique, de la place de l'Homme face au « robot » dans le futur, du partage des espaces de vie, de la notion d'être vivant et pensant, … Sans oublier de prendre en compte des connaissances scientifiques et les méthodes de création des intelligences artificielles, les aspects techniques de leur fonctionnement, … C'est un sujet qui prend place dans un contexte où le numérique est omniprésent dans notre quotidien et les nouvelles technologies engagent de grands changements dans nos modes de vie.

Notre Projet s'inscrit dans ce cadre. En effet, notre but est de concevoir et développer une application client à partir d'une base de données contenant toutes les informations nécessaires liés à la voitures (exemples : vitesse, zone de conduite, consommation, ...) afin de savoir si le conducteur de la voiture est un chauffeur parfait ou mauvais afin de limiter le nombre des accidents et de bien contrôler les chauffeurs fautifs dans la rue.

Ce rapport se compose de trois chapitres. Le premier chapitre est réservé à la présentation des outils nécessaires au développement de l'application et dans lequel nous allons définir les améliorations et les solutions adoptées. L'objectif du projet et la conception sont présentés dans le deuxième chapitre dans lequel aussi on a réalisé une description complète de toutes les variables (qualitatives et quantitatives) Le troisième chapitre est entièrement réservé à l'implémentation et la réalisation de l'application.
Finalement nous clôturons le rapport par une conclusion générale qui présente le bilan de ce projet en plus des annexes pour une meilleure compréhension du contenu.

Chapitre 1 : Cadre du projet

1.1 Introduction

Dans ce chapitre nous mettons notre travail dans son contexte. On présente les outils nécessaires au développement de l'application et on définit les améliorations et les solutions adoptées.

1.2 Intelligence Artificielle (IA)

L'intelligence artificielle, « le grand mythe de notre temps » comme le déclarait certaines instances, est aujourd'hui au cœur de toutes les attentions. Il n'est pas un jour sans que ne paraisse un article ou une étude sur ses bienfaits. Elle est au cœur de la compétition économique mondiale et au centre des interrogations politiques et géopolitiques. Les chercheurs spécialistes en intelligence artificielle, bien qu'ils soient de plus en plus nombreux, sont devenus une denrée rare et prisée et les entreprises, issues de différents secteurs, investissent de plus en plus dans ce domaine. Ainsi l'intelligence artificielle est considérée comme la principale innovation d'une nouvelle révolution industrielle, celle du travail de l'homme avec des machines dites intelligentes.

L'intelligence artificielle (IA) est « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ». Elle correspond donc à un ensemble de concepts et de technologies plus qu'à une discipline autonome constituée.

Souvent classée dans le groupe des sciences cognitives, elle fait appel aux réseaux de neurones, à la logique mathématique et à l'informatique. Elle recherche des méthodes de résolution de problèmes à forte complexité, pour plus d'informations vous pouvez consulter [1].

1.3 Machine Learning

L'apprentissage automatique ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux machines la capacité d'apprendre à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

L'apprentissage automatique comporte généralement deux phases. La première consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. Cette phase est dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle. La

seconde phase correspond à la mise en production. Le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée. En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits.

L'apprentissage est qualifié de différentes manières. Si les données sont étiquetées (c'est-à-dire que la réponse à la tâche est connue pour ces données), il s'agit d'un apprentissage supervisé. On parle de classification ou de classement si les étiquettes sont discrètes et de régression si elles sont continuées. Dans le cas le plus général, sans étiquette, on cherche à déterminer la structure sous-jacente des données (qui peuvent être une densité de probabilité) et il s'agit alors d' apprentissage non supervisé. L'apprentissage automatique peut être appliqué à différents types de données, les graphes, les arbres, les courbes, ou plus simplement les vecteurs de caractéristiques, qui peuvent être des variables qualitatives ou quantitatives continues ou discrètes. Pour plus d'informations vous pouvez consulter [2].

1.4 Apprentissage supervisé

L'apprentissage supervisé ou (supervised learning) est une forme d'apprentissage machine qui crée des modèles d'intelligence artificielle en se fondant sur des données d'apprentissage « étiquetées ». Dans l'apprentissage supervisé, chaque exemple est un couple constitué d'un objet d'entrée (généralement un vecteur) et d'une valeur de sortie souhaitée. Un algorithme d'apprentissage supervisé analyse les données d'apprentissage et produit une fonction inférée, qui peut être utilisée pour mapper de nouveaux exemples. Pour plus d'informations vous pouvez consulter [3].

1.5 Apprentissage non supervisé

Dans le domaine informatique et de l'IA, l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées. Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées et dans ce cas il est impossible à l'algorithme de calculer de façon certaine un score de réussite.

La Table 1 illustre les différences entre les deux types d'apprentissage supervisé et non supervisé, pour plus d'informations vous pouvez consulter [4].

	Apprentissage supervisé	Apprentissage non supervisé
Données d'entrée	Utilise les données connues et étiquetées comme entrées	Données inconnues en entrée
Complexité informatique	Très complexe	Moins de complexité informatique
Temps réél	Utilise l'analyse hors ligne	Utilise l'analyse en temps réel des données
Sous-domaines	Classification et régression	Exploitation de règles de clustering et d'association
Précision	Produit des résultats précis	Génère des résultats modérés
Nombre de classes	Nombre de classes connues	Le nombre de classes n'est pas connu

Table 1: Comparaison entre l'apprentissage supervisé et non supervisé

1.5.1 Clustering

L'apprentissage non supervisé n'avait pas une variable sortie bien déterminé ce qui rend les composants principales non exploitable.

L'idée générale est donc de classifier les variables à travers une méthode qui s'appelle **'Clustering'**.

Clustering est la technique la plus utilisée pour résoudre les problèmes d'apprentissage non supervisé. La mise en cluster consiste à séparer ou à diviser un ensemble de données en un certain nombre de groupes, de sorte que les ensembles de données appartenant aux mêmes groupes se ressemblent d'avantage que ceux d'autres groupes. En termes simples, l'objectif est de séparer les groupes ayant des traits similaires et de les assigner en grappes. Pour plus d'informations vous pouvez consulter [6].

1.5.2 K-means Clustering

Le partitionnement en k-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

K-means clustering est un algorithme itératif qui sert à chercher la similarité des données à travers les distances séparant un point et un centroïde (centre d'un cluster).

En générale, K-means cherche la position des centres qui minimise la distance entre les points d'un cluster ($\boldsymbol{x}_i$) et le centre ($\boldsymbol{\mu}_i$). Pour plus d'informations vous pouvez consulter [7].

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2\right)$$

Figure 1: Equation de calcul de distance K-means

1.5.3 Méthode Elbow

La méthode Elbow est une méthode empirique permettant de trouver le nombre optimal de clusters pour un ensemble de données. Dans cette méthode, nous choisissons une plage de valeurs candidates de k, puis appliquons le clustering K-Means en utilisant chacune des valeurs de k. Trouvez la distance moyenne de chaque point d'un cluster à son centre de gravité et représentez-la dans un tracé. Choisissez la valeur de k, où la distance moyenne tombe soudainement. Pour trouver le nombre optimal de clusters, nous utilisons une métrique appelée le " Within Cluster Sum Of Squares" (WCSS). Dans 'WCSS', nous prenons la somme des carrés de la distance entre chaque point à l'intérieur du cluster et le centroïde respectif pour tous les clusters.

$$WCSS(k)=\sum_{j=1}^{k} \sum_{x_i \in cluster(j)} \|x_i - \bar{x_j}\|^2 \qquad \bar{x_j} \text{ est la Moyenne de cluster j}$$

Tenant l'exemple de K=3, la somme des carrés des clusters (WCSS) sera :

$$WCSS=\sum_{Pi(\text{cluster1})} \text{distance}(\text{Pi}, \text{c1})^2 + \sum_{Pi(\text{cluster2})} \text{distance}(\text{Pi}, \text{C2})^2 + \sum_{Pi(\text{cluster2})} \text{distance}(\text{Pi}, \text{C3})^2$$

Avec C1, C2 et C3 sont les centroïdes des clusters. Pour plus d'informations vous pouvez consulter [8].

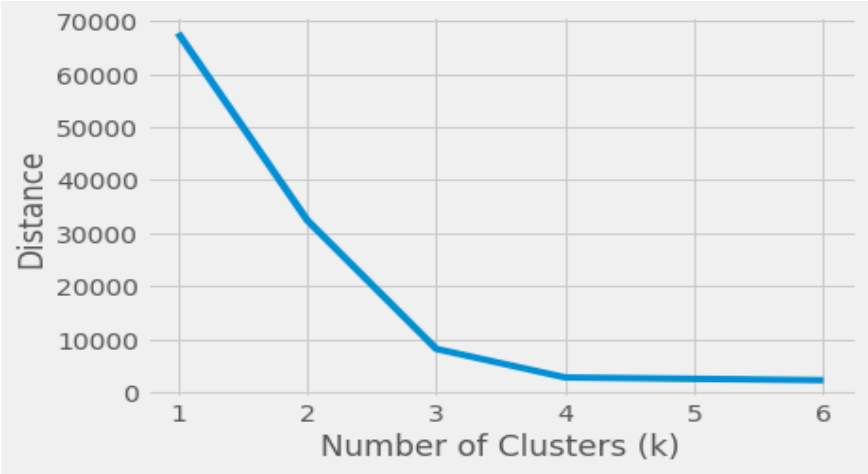


Figure 2: Courbe de la méthode Elbow

1.5.4 Hierarchical Agglomérative Clustering

Il est à signaler que la méthode K-means clustering n'est pas le seul modèle d'apprentissage non supervisé. En effet, il existe d'autres modèles à savoir « Hierarchical Agglomerative Clustering ». Cette méthode est aussi à base de calculs de distance entre les clusters. On commence par initialiser chaque observation dans un cluster (nombre de clusters initiales = nombre d'observations). Dans chaque itération, on cherche une similarité entre deux clusters à travers le calcul des distances entre tous les clusters. On fait donc l'assemblage de ces deux clusters pour obtenir un nombre de cluster égale à nombre de cluster précédant -1. On fixe le nombre de clusters à atteindre en utilisant cette méthode, ce qui signifie que les itérations s'arrêtent lorsqu'on atteint le nombre de clusters souhaités. Pour plus d'informations vous pouvez consulter [9].

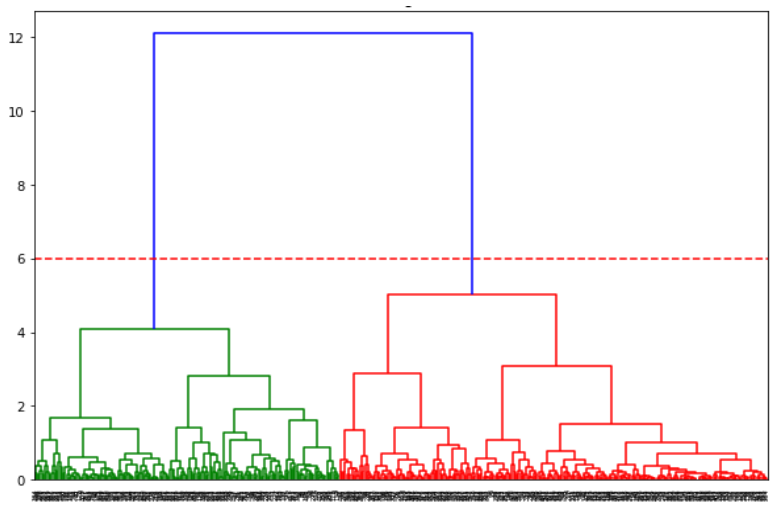


Figure 3: Dendogram

1.6 Analyse en Composantes Principales (ACP)

L'analyse en composantes principales ou ACP est une procédure statistique qui nous permet de résumer ou d'extraire les seules données importantes qui expliquent l'ensemble de données.

Aujourd'hui, l'analyse en composantes principales est l'une des techniques statistiques multivariées les plus populaires. Elle est largement utilisée dans les domaines de la reconnaissance de formes, du traitement du signal et dans l'analyse statistique pour réduire la dimension. Plus précisément, elle permet de comprendre et extraire uniquement les facteurs importants qui expliquent l'ensemble des données. Ainsi, l'ACP, aide à éviter le traitement de données inutiles. En effet, l'analyse en composantes principales permet de transformer des variables corrélées en variables décorrélées. Elle vise réduire le nombre de variables, pour simplifier les observations tout en conservant un maximum d'informations (à partir de n variables indépendantes du dataset, on considère $p < n$ nouvelles variables indépendantes qui expliquent la plus grande variance du jeu de données).

➔ Le fait que le DV Y ne soit pas pris en compte fait du PCA un modèle non supervisé

Cette technique a été bien appliquée dans notre jeu de données qui contient 13 variables et elle nous a permis de conserver uniquement 3 composantes principales.

1.6.1 Les différents étapes de ACP

Pour réaliser l'ACP on suit une démarche en plusieurs étapes (Pour plus d'informations voir [5]).

1) Centrer et réduire la matrice de caractéristiques X

2) Calculer la matrice de variance-covariance de la matrice centrée réduite

3) Déterminer les métriques

4) Calculer les valeurs propres

5) Choisir comme variables explicatives celles qui contiennent les valeurs propres les plus grandes en valeurs absolues

1.7 Data preprocessing

Lorsque nous parlons de données, nous pensons généralement à de grands ensembles de données avec un grand nombre de lignes et de colonnes. Bien que ce soit un scénario probable, ce n'est pas toujours le cas - les données peuvent se présenter sous de nombreuses formes différentes : tableaux structurés, images, fichiers audio, vidéos, etc.

Les machines ne comprennent pas les données de texte libre, d'image ou de vidéo telles qu'elles sont, elles comprennent les 1 et les 0. Donc, ce ne sera probablement pas assez bon si nous mettons en place un diaporama de toutes nos images et attendons que notre modèle d'apprentissage automatique soit formé juste par cela !

Dans tout processus d'apprentissage automatique, le prétraitement des données est l'étape au cours de laquelle les données sont transformées, ou encodées, pour les amener à un état tel que la machine peut désormais les analyser facilement. En d'autres termes, les caractéristiques des données peuvent maintenant être facilement interprétées par l'algorithme. Pour plus d'informations vous pouvez consulter [10].

1.7.1 Encodage

Comme on a mentionné précédemment, tout le but du 'DATA PREPROCESSING' est de coder les données afin de les amener à un état tel que la machine les comprend maintenant.

L'encodage des fonctionnalités effectue essentiellement des transformations sur les données de telle sorte qu'elles puissent être facilement acceptées comme entrée pour les algorithmes d'apprentissage automatique tout en conservant leur signification d'origine. Pour plus d'informations vous pouvez consulter [11].

1.7.2 Missing values (Valeurs manquantes)

Il est très courant d'avoir des valeurs manquantes dans notre ensemble de données. Cela peut s'être produit lors de la collecte de données, ou peut-être en raison d'une règle de validation des données, mais quelles que soient les valeurs manquantes, il faut prendre en considération.

Élimine les lignes avec des données manquantes :

Stratégie simple et parfois efficace. Échoue si de nombreux objets ont des valeurs manquantes. Si une caractéristique a pour la plupart des valeurs manquantes, cette caractéristique elle-même peut également être éliminée.

Estimer les valeurs manquantes :

S'il ne manque qu'un pourcentage raisonnable de valeurs, nous pouvons également exécuter des méthodes d'interpolation simples pour remplir ces valeurs. Cependant, la méthode la plus courante pour traiter les valeurs manquantes consiste à les remplir avec la valeur moyenne, médiane ou de mode de la caractéristique respective. Pour plus d'informations vous pouvez consulter [12].

Conclusion :

Ce chapitre donne une idée générale sur l'apprentissage non-supervisé, en définissant les outils nécessaires.

Chapitre 2 : Etude conceptuelle

2.1 Introduction

Dans ce deuxième chapitre on va monter des modèles afin de classifier les conducteurs en deux catégories mauvais conducteurs ou bon conducteur. Dans notre projet, nous disposons d'une grande base de 95549 observations avec 12 variables.

2.2 Description des données :

L'attitude : La latitude est une coordonnée géographique.

Longitude : La longitude est une coordonnée géographique.

Speed : Vitesse de voiture.

Engine Load : la charge du moteur est tout simplement la quantité de carburant introduite à chaque cycle. (Puissance fournie par le moteur).

AmbientAirtemp : température ambiante.

InsFuel : consommation instantanée de carburants.

X, Y, Z : Paramètre de l'accéléromètre.

2.3 Statistique descriptive

On ne peut jamais commencer l'apprentissage sans connaitre quelques informations et une idée générale sur la base de données pour pouvoir avoir un bon traitement. La figure 4 donne une idée sur les types des variables du dataset. Notre base de données comporte une variable qualitative et 11 variables quantitatives. Pour plus d'information vous pouvez consulter [13].

```
float64    11
object      1
dtype: int64
```

Figure 4: Types des variables explicatives

Table 2 donne des informations concernant le nombre de valeur, la moyenne, l'écart-type, la valeur minimale, la valeur en 25%, la valeur en 50%, la valeur en 75% et la valeur maximale de chaque variable quantitative explicative.

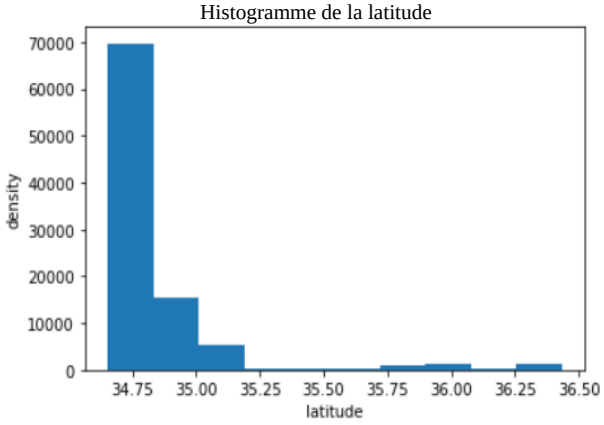
Index	latitude	longitude	Speed	ENGINE_LOAD	AmbientAirTemp	ThrottlePos	insFuel	X	Y	Z
count	95549	95549	86300	86300	86300	85997	53786	92050	92050	92050
mean	34.85	10.65	49.27	35.92	33.87	24.67	18.13	-0.24	4.12	7.21
std	0.28	0.2	31.96	27.65	7.98	16.31	34.77	2.04	3.55	3.36
min	34.65	9.43	0	0	-40	0	0	-16.25	-15.26	-15.15
25%	34.73	10.66	26	14.1	28	15.7	7.03	-0.83	2.45	4.65
50%	34.77	10.71	44	30.2	34	20	10.11	-0.21	3.17	9.03
75%	34.84	10.74	73	56.5	39	25.9	15.88	0.35	6.98	9.41
max	36.43	10.79	145	100	215	88.6	1004.73	18.13	15.22	19.1

Table 2: Description des variables quantitatives du dataset

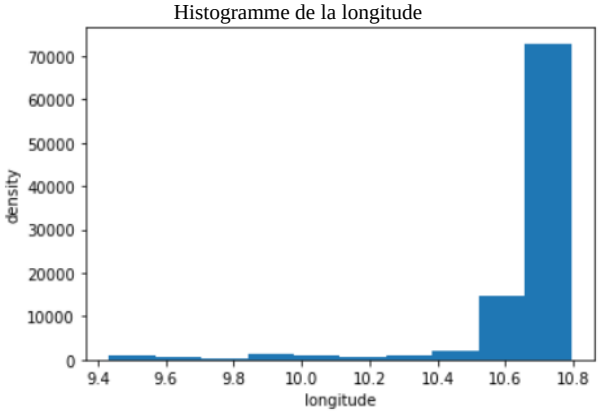
Définition :

On appelle histogramme toute représentation graphique permettant de décrire la répartition d'une variable en la représentant avec des bâtons.

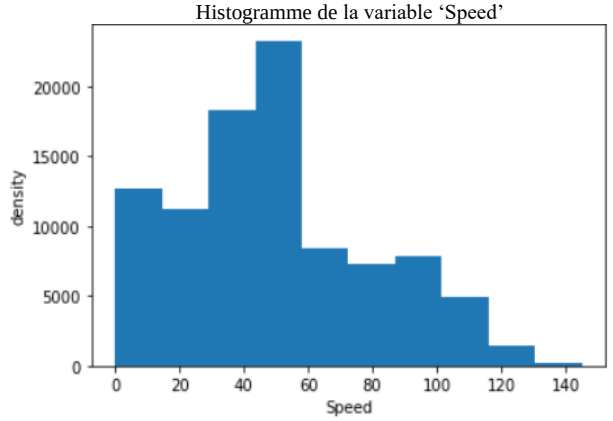
Histogramme de la latitude



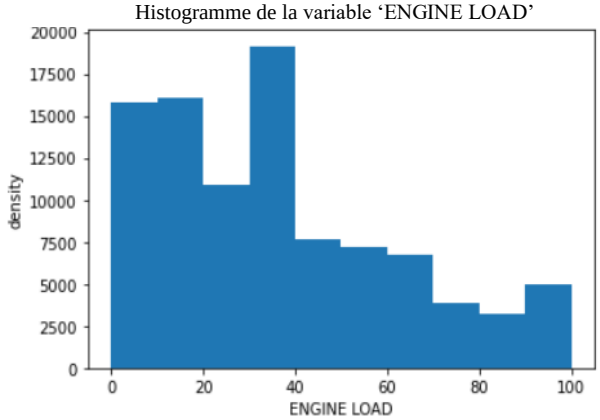
Histogramme de la longitude



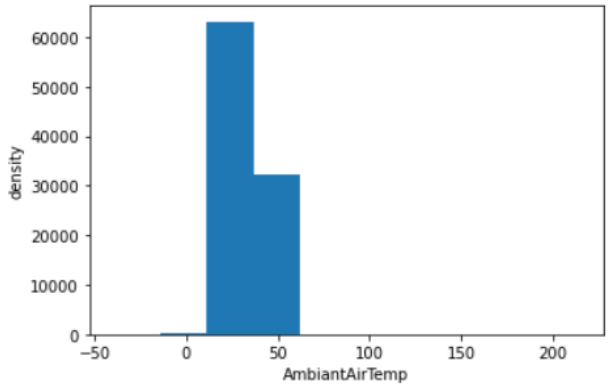
Histogramme de la variable 'Speed'



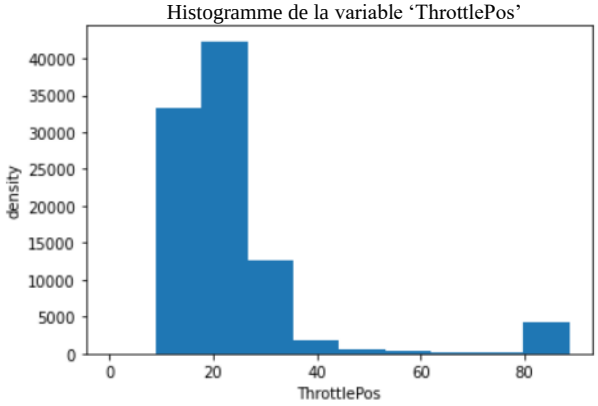
Histogramme de la variable 'ENGINE LOAD'



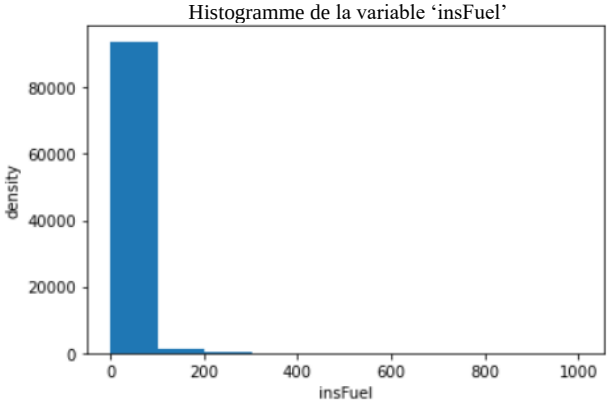
Histogramme de la variable 'AmbientAirTemp'



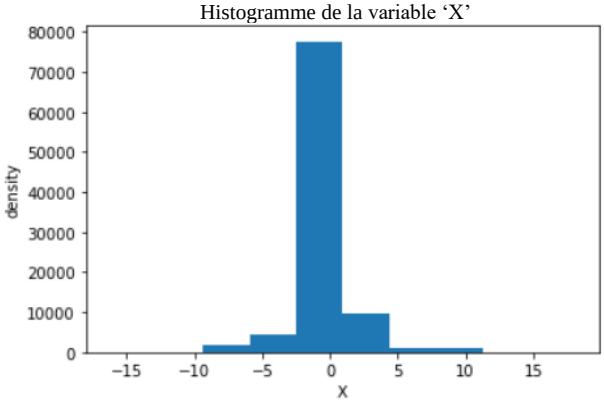
Histogramme de la variable 'ThrottlePos'



Histogramme de la variable 'insFuel'



Histogramme de la variable 'X'



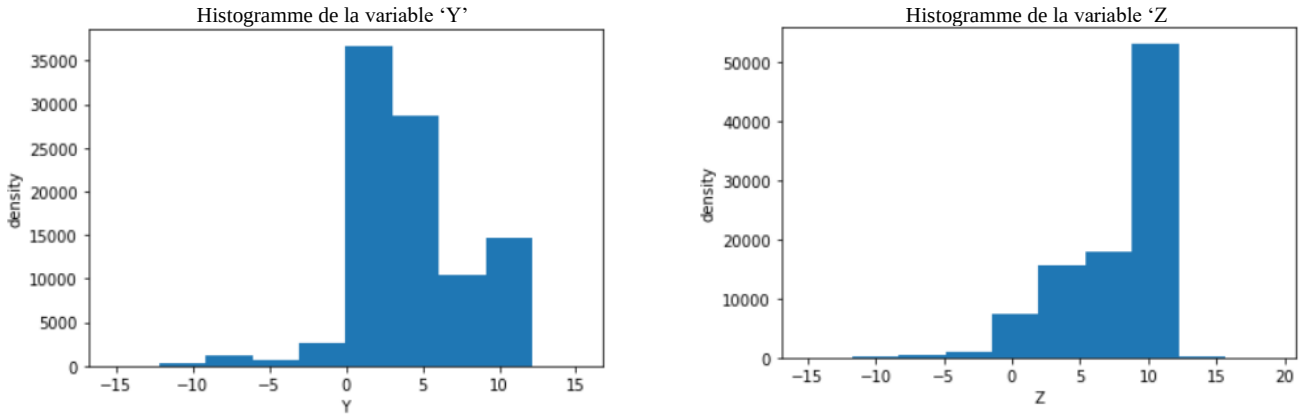


Figure 5: Histogrammes des variables quantitatives

Figure 5 représente les histogrammes des différents variables quantitatives du dataset. On remarque que les variables Speed et X ont des distributions Gaussiennes de moyennes respectives 49.24 et -0.24 et d'écart-types respectives 31.96 et 2.04.

Notre dataset comporte une seule variable qualitative qui est la variable Zone qui prend comme modalité 'urbain' nationale' autoroute. Figure 6 illustre la proportion de chaque modalité

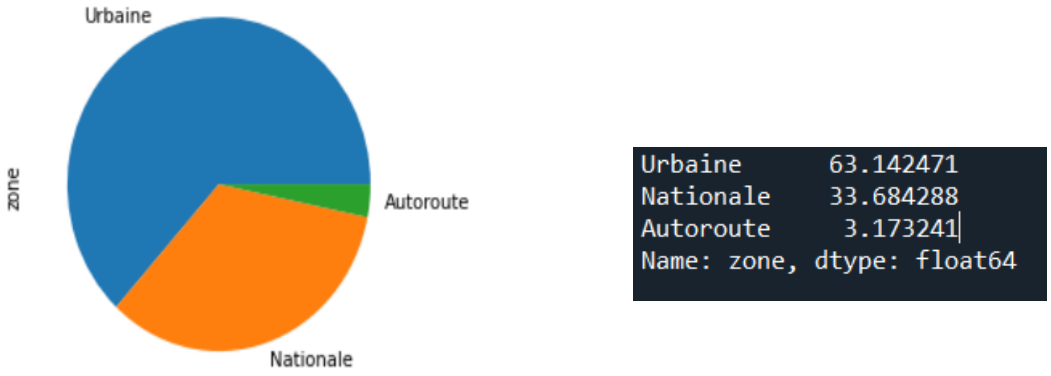


Figure 6: Représentation des variables qualitatives

La modalité urbaine présente 63.14% des observations et la modalité Autoroute présente 3.71% des observations (un faible nombre d'observations).

2.4 Data Pre-processing

Notre base de données comporte des données réellement numériques mais suivis d'une unité « C » ou bien « % » ou bien des valeurs égales à 0 définit par le mot « null », qui n'est pas compréhensible par la machine donc elle les considère comme des chaines de caractères (voir Figure7).

Index	CIN	latitude	longitude	Speed	IGINE_LO/	bientAirTe	hrottlePo	insFuel	X	Y	Z
0	6023226	35.0819	9.86976	8	null	null	nan	nan	nan	nan	nan
1	6023226	35.0819	9.8699	20	40,4%	36C	23,9%	nan	nan	nan	nan
2	6023226	35.0819	9.8702	26	83,5%	36C	33,7%	nan	nan	nan	nan
3	6023226	35.0819	9.87046	37	64,7%	36C	25,9%	nan	nan	nan	nan
4	6023226	35.0819	9.87104	41	64,7%	36C	26,3%	nan	nan	nan	nan

Figure 7: Dataset initiale

Pour pouvoir les intégrer dans des équations mathématiques on doit changer le type vers 'float' (voir Figure 8).

Index	CIN	latitude	longitude	Speed	IGINE_LO/	bientAirTe	hrottlePo	insFuel	X	Y	Z	zone	time
0	6023226	35.0819	9.86976	8	0	0	nan	nan	nan	nan	nan	Nationale	7/5/2020 21:13
1	6023226	35.0819	9.8699	20	40.40	36	23.90	nan	nan	nan	nan	Nationale	7/5/2020 21:13
2	6023226	35.0819	9.8702	26	83.50	36	33.70	nan	nan	nan	nan	Nationale	7/5/2020 21:13
3	6023226	35.0819	9.87046	37	64.70	36	25.90	nan	nan	nan	nan	Nationale	7/5/2020 21:13
4	6023226	35.0819	9.87104	41	64.70	36	26.30	nan	nan	nan	nan	Nationale	7/5/2020 21:13
5	6023226	35.082	9.87146	44	61.60	36	27.50	nan	nan	nan	nan	Nationale	7/5/2020 21:13

Figure 8: Dataset avec features type float

L'étape suivante consiste à éliminer les colonnes inutiles notre dataset contient deux colonnes inutiles 'CIN' 'time' (voir Figure 9) .

Index	latitude	longitude	Speed	IGINE_LOA	bientAirTe	hrottlePo	insFuel	X	Y	Z	zone
0	35.0819	9.86976	8	0	0	nan	nan	nan	nan	nan	Nationale
1	35.0819	9.8699	20	40.4	36	23.9	nan	nan	nan	nan	Nationale
2	35.0819	9.8702	26	83.5	36	33.7	nan	nan	nan	nan	Nationale

Figure 9: Dataset avec les variables finales à traiter

La Figure 10 donne la proportion des valeurs manquantes pour chaque variable

latitude	0.000000
longitude	0.000000
Speed	9.679850
ENGINE_LOAD	9.679850
AmbientAirTemp	9.679850
ThrottlePos	9.996965
insFuel	43.708464
X	3.661995
Y	3.661995
Z	3.661995
zone	0.000000

Figure 10: Pourcentage des valeurs manquantes pour chaque colonne

La colonne 'insFuel' contient plus de valeurs manquantes que les autres, soit 43.7%.
La solution proposée est de remplir les valeurs manquantes de la base de données des variables quantitatives par la stratégie de la médiane (Voir figure 11), car, il s'agit des variables non Gaussienne. (Figure 5)

latitude	longitude	Speed	IGINE_LOA	bientAirTe	hrottlePo	insFuel	X	Y	Z	zone
35.0819	9.86976	8	0	0	24.6659	18.129	-0.241558	4.11843	7.21185	Nationale
35.0819	9.8699	20	40.4	36	23.9	18.129	-0.241558	4.11843	7.21185	Nationale
35.0819	9.8702	26	83.5	36	33.7	18.129	-0.241558	4.11843	7.21185	Nationale
35.0819	9.87046	37	64.7	36	25.9	18.129	-0.241558	4.11843	7.21185	Nationale

Figure 11: Valeurs manquantes bien remplis

Concernant la variable qualitative 'zone', il n'y a pas des valeurs manquantes (voir Figure 7). Afin d'intégrer cette variable (nominale) dans une équation mathématique, on passe à l'encodage en utilisant ''Dummy variables' (voir Figure 12).

Index	1	2	latitude	longitude	Speed	IGINE_LOA	bientAirTe	hrottlePo	insFuel	X	Y	Z
0	1	0	35.0819	9.86976	8	0	0	24.6659	18.129	-0.241558	4.11843	7.2118
1	1	0	35.0819	9.8699	20	40.4	36	23.9	18.129	-0.241558	4.11843	7.2118
2	1	0	35.0819	9.8702	26	83.5	36	33.7	18.129	-0.241558	4.11843	7.2118
3	1	0	35.0819	9.87046	37	64.7	36	25.9	18.129	-0.241558	4.11843	7.2118
4	1	0	35.0819	9.87104	41	64.7	36	26.3	18.129	-0.241558	4.11843	7.2118

Figure 12: Encodage du variable Zone

La variable zone à été remplacer par les deux colonnes '1' et '2' le couple (1,0) indique la modalité 'Nationale'. Le couple (0,1) indique la modalité 'Urbaine'. Le couple (0,0) indique la modalité 'Autoroute' (voir figure 12).

2.5 Analyse en Composantes Principales (ACP)

Notre base de données est prête pour la partie d'entrainement mais ce qu'il faut bien retenir que notre base de données comporte 95549 observations et 12 variables ce qui est très lourd pour la machine.

L'idée est de réduire la base en composantes principales par la méthode ACP. Mais comment déterminé les variables les plus pertinentes (une grande influence) ?

La méthode de « variance ratio » peut donner en pourcentage l'influence de chaque variable sur l'entrainement en ordre décroissant. (voir Figure 13)

```
[5.05944936e-01 2.71855940e-01 1.57227993e-01 3.44077635e-02
 2.03200140e-02 6.74245943e-03 1.87764931e-03 1.47175794e-03
 1.05140693e-04 3.04164279e-05 1.20577181e-05 3.87239872e-06]
```

Figure 13: 'variance Ratio' des variables du dataset

Les trois variables les plus pertinentes ont des pourcentages respectifs 50.59%, 27.18 % et 15.72%.

	0	1	2
0	49.0271	-11.4368	-5.64742
1	19.6217	6.27948	-21.1904
2	-11.3406	27.7262	-42.6616

Figure 14: Les trois premières observations de la nouvelle base de données après ACP

2.6 K-means Clustering

Maintenant on monte le premier modèle d'apprentissage non supervisé à savoir le K-means Clustering dans le but de diviser les conducteurs en des classes bien déterminées. Pour cela on commence par déterminer le nombre de classes adéquat et on utilise la méthode Elbow (voir Figure 15)

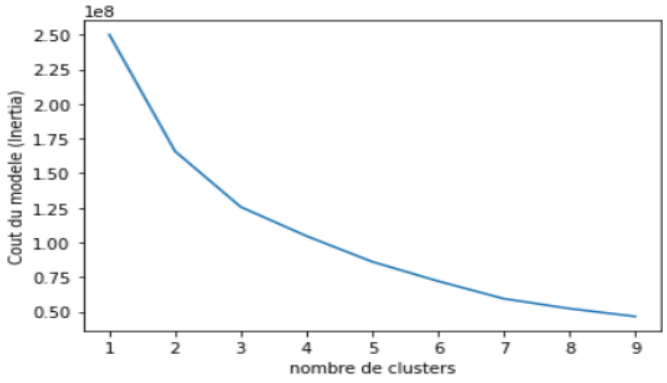


Figure 15: Méthode Elbow

Le meilleur choix est de deux clusters pour cette base d'après la Figure 15.

La prédiction en utilisant la méthode K-means donne la répartition suivante des conducteurs

	0
0	0
1	0
2	1
3	0
4	0

Figure 16: Résultat de prédiction par K-means

Les observations sont bien classifiées en deux classes (classe 0, classe 1). Mais laquelle de ces classes définit un bon conducteur et un mauvais conducteur ?

Dans cette base de données il y a un groupe d'observations qui sont par évidence associé à un mauvais conducteur et un autre associé à un bon conducteur. La solution est qu'on se met à la place d'un expert pour pouvoir estimer ces deux groupes.

On déduit que Cluster '0' convient à un bon conducteur et Cluster'1' convient à un mauvais conducteur.

2.7 Hierarchical Agglomerative Clustering

On applique maintenant le modèle Hierarchical Agglomerative Clustering à notre jeu de données. On obtient le Dendogram suivant.

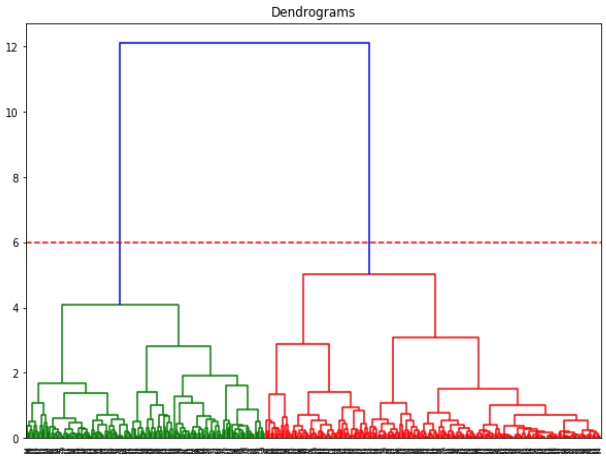


Figure 17: Dendogram

Figure 17 montre que nos observations seront réparties en deux clusters. La prédiction du comportement des conducteurs en utilisant cette méthode donne une grande similarité avec la méthode K-means sauf au niveau de 311 observations qui représentent 311/95548=0.32% qui représente une faible proportion de notre dataset.

Conclusion

Dans ce chapitre, on a essayé de corriger quelques défaillances de notre base de données et de bien manipuler des méthodes d'apprentissages non-supervisé. Ceci nous permet d'avoir un bon entraînement et par conséquent conduit à prédire de bons résultats.

Chapitre 3 : Implémentation de l'application

3.1 Introduction

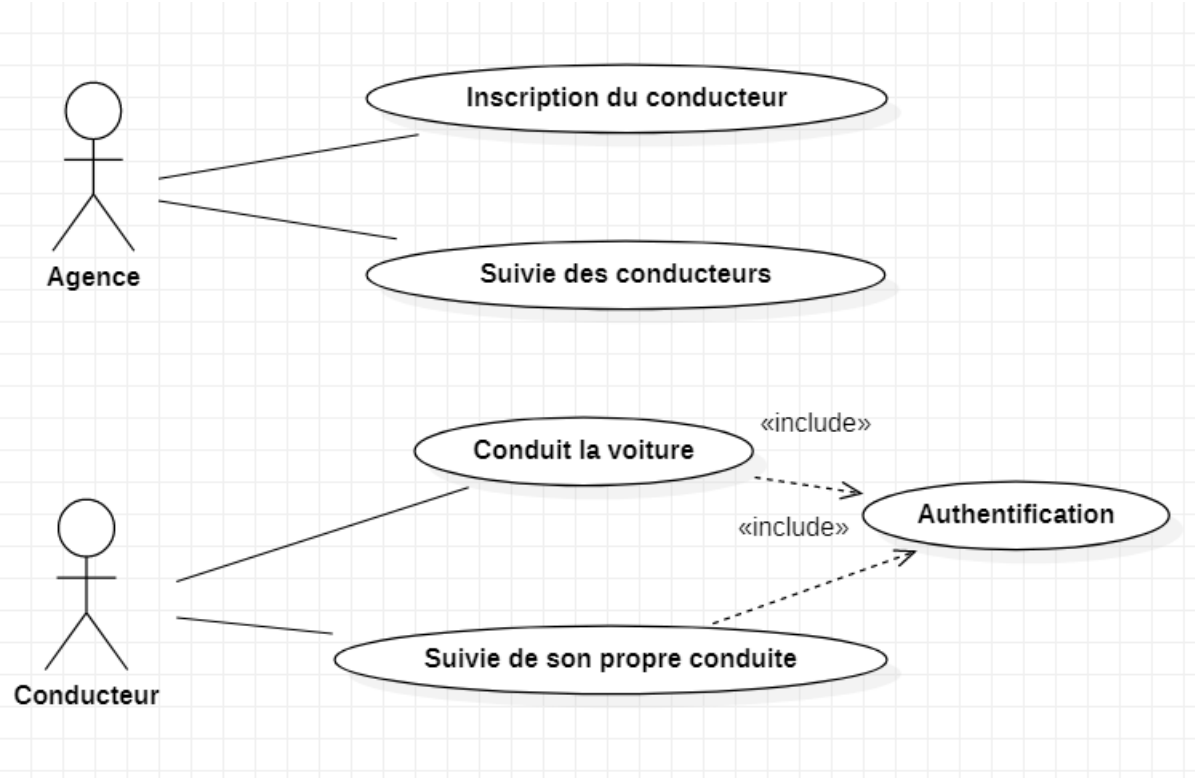
Dans le chapitre précédant, on a réussi de faire un bon apprentissage de la machine sur notre base de données et d'avoir des résultats de prédiction proche de la réalité. Dans ce chapitre on va essayer de profiter ces atouts-là pour avoir une application permettant d'avoir une observation totale sur le chauffeur et de prendre une décision si le chauffeur est un bon ou mauvais conducteur.

3.2 Etude conceptuelle

3.2.1 Diagrammes d'illustration

3.2.1.1 Diagrammes de cas d'utilisation

Le diagramme de cas d'utilisation a pour but de donner une vision globale sur les interfaces de future application. C'est le premier diagramme UML. Constitué d'un ensemble d'acteurs qui agit sur des cas d'utilisation et qui est décrit sous forme d'actions et de réactions, le comportement d'un système du point de vue utilisateur.



3.2.1.2 Diagrammes de séquence

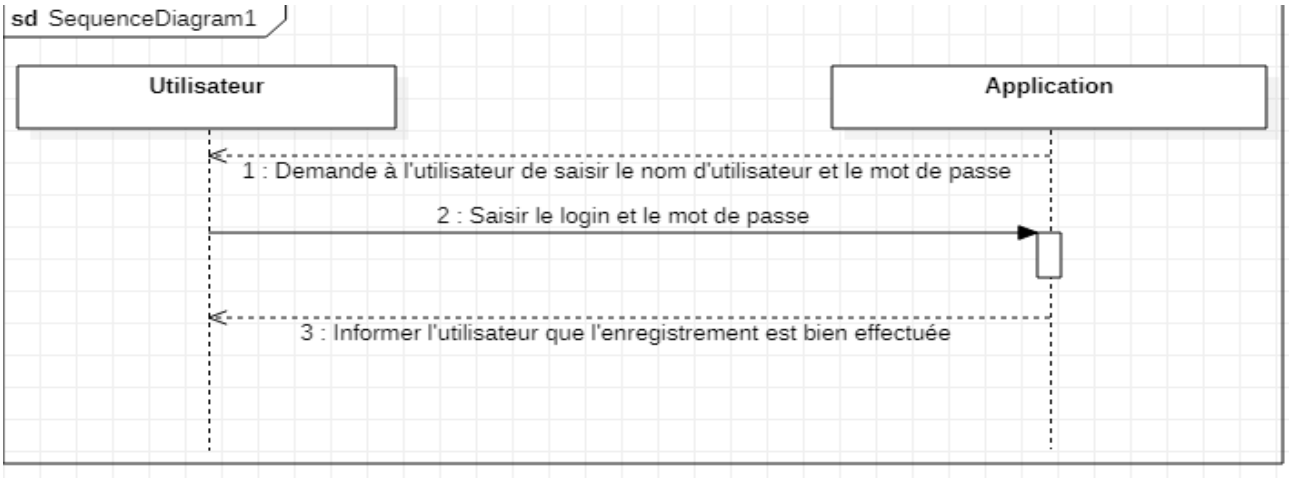


Figure 19: Diagramme de séquence "séquence d'inscription"

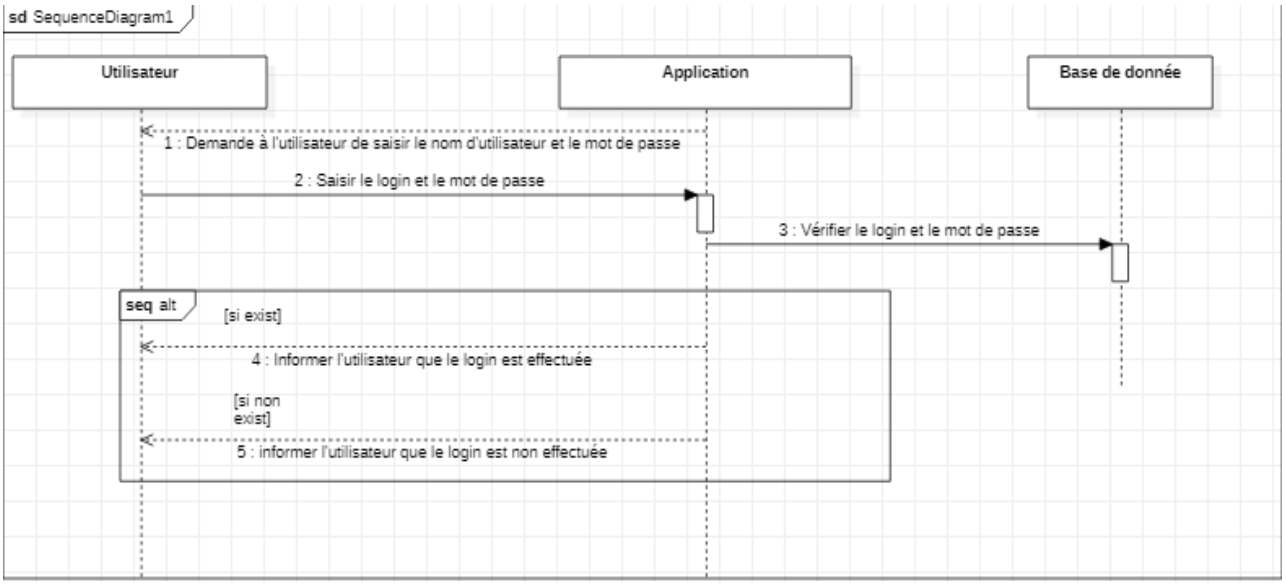


Figure 20: Diagramme de séquence " séquence du login"

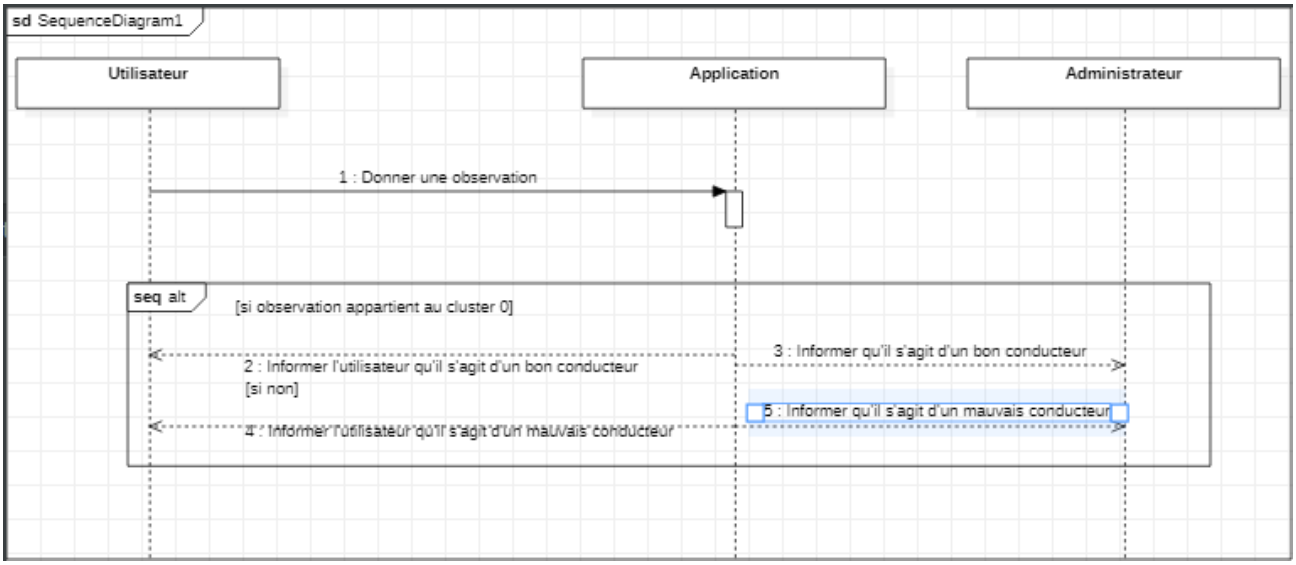


Figure 21: Diagramme de séquence "séquence suivie du conducteur"

3.2.2 Environnement et les outils de développement

Dans cette partie, nous avons présenté l'environnement de développement à travers la spécification des différents outils logiciels que nous avons utilisée pour réaliser notre application.

3.2.2.1 Choix du langage de programmation

- Python

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages.

Il reste aussi accessible pour les débutants, à condition de lui consacrer un peu de temps pour la prise en main. De nombreux tutoriels sont d'ailleurs disponibles pour l'étudier sur des sites Internet spécialisés ou sur des comptes YouTube. Sur les forums d'informatique, il est toujours possible de trouver des réponses à ses questions, puisque beaucoup de professionnels l'utilisent.

Pour plus d'informations vous pouvez consulter [14].

- A quoi sert le langage Python ?

Les principales utilisations de Python par les développeurs sont :

- La programmation d'applications
- La création de services web
- La génération de code
- La métaprogrammation.
- Création des modèles d'intelligence artificielle.

- Le module ' Tkinter' (interface graphique)

Il existe plusieurs modules permettant d'exploiter les interfaces graphiques. Le plus simple est le module tkinter présent lors de l'installation du langage Python. Ce module est simple mais limité. Le module tkinter fait partie de la distribution standard de Python et sera disponible dans toutes les verions de Python. Visuellement, tkinter est moins joli que d'autres extensions mais il vaut mieux vérifier la fréquence des mises à jour de leur code source avant d'en choisir une github/wxPython github/PyQt5. La licence de wxPython est plus souple. D'autres alternatives sont disponibles à Other Graphical User Interface Packages.

Le fonctionnement des interfaces graphiques sous un module ou un autre est presque identique. C'est pourquoi ce chapitre n'en présentera qu'un seul, le module tkinter. Pour d'autres modules, les noms de classes changent mais la logique reste la même : il s'agit d'associer des événements à des parties du programme Python.

3.2.2.2 Les outils logiciels

- Spyder

Spyder est un environnement scientifique gratuit et open source écrit en Python, et conçu par et pour des scientifiques, des ingénieurs et des analystes de données. Il présente une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les belles capacités de visualisation d'un package scientifique. Pour plus d'informations vous pouvez consulter [15].

Figure 22: Logo du logiciel "Spyder".

- StarUML

  StarUML est un logiciel de modélisation UML, qui a été "cédé comme open source" par son éditeur", à la fin de son exploitation commerciale
  Il comprend de nombreux compléments utiles avec diverses fonctionnalités : il génère des codes sources dans des langages de programmation et convertit les codes sources en modèles. Pour plus d'information vous pouvez consulter [16]

Figure 23: Logo Star Uml

- Anaconda :

Anaconda est une distribution libre et open source de langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique),elle vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets conda. La distribution Anaconda est utilisée par plus de 6 millions d'utilisateurs et comprend plus de 250 paquets populaires en science des données adaptés pour Windows, Linux et MacOS.

Figure 24: Logo ANACONDA

3.3 Notion d'administrateur et de Client dans l'application

Cette application est dédié pour n'importe personne veut evaluer la qualité de conduite pour une voiture spécifique. On peut considérer à titre d'exemple un propriétaire d'agence de location des voitures qui a voulu les suivre et d'avertir le conducteur à chaque bêtise commise comme administrateur. il doit tout d'abord faire l'inscription de chaque locataire (qui pourra être plus qu'un personne pour une seule voiture)(voir Figure 26) .Une personne non inscrite ne peut pas être classifier(voir Figure 27). A chaque utilisation de voiture chaque client doit ouvrir sa propre session pour assurer un bon contrôle.

Figure 25: Fenêtre principale de l'application

Une fois la session est bien activée, l'observation est prête pour la classification.

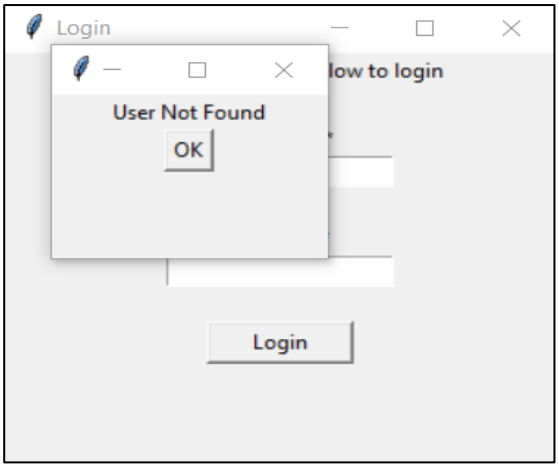


Figure26 : Session introuvable

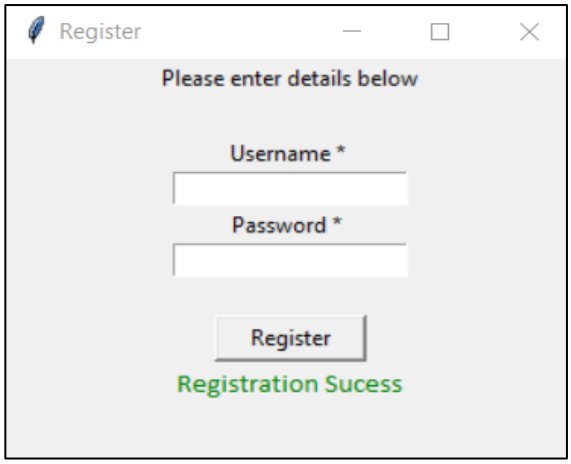


Figure27 : Fenêtre d'inscription

La décision sera prise par la mise des informations liées à la voiture comme des entrées par exemple : 'Speed' , 'zone' , 'Engine Load'... A chaque observation émise pour l'application , elle va prendre la décision si le chauffeur est un mauvais ou bon conducteur (voir Figure 28).

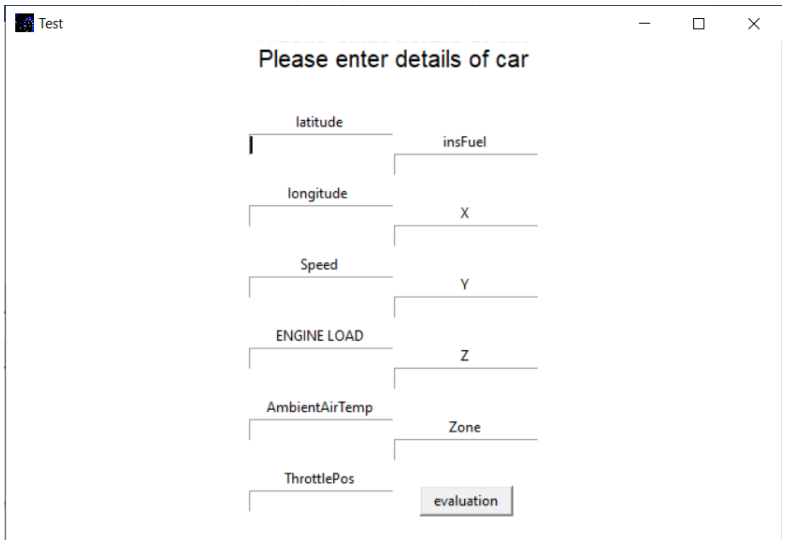


Figure 26 Fenêtre de prise d'informations

3.4 Réalisation du modèle

Notre application va créer des modèles plus précis basés sur les étapes mentionnées au Chapitre 2 , cela signifie qu'elle est prête pour faire une prédiction pour n'importe quelle observation. A la fênetre des entrées , tout en remplissant les champs de la nouvelle observation , l'application va

créer une nouvelle base de données qui admet les mêmes caracteristques que la base de données qui a servi pour l'apprentissage.

Ensuite, elle la gènere tout en encodant la variable qualitative "zone", applique l'ACP et prédit, en utilisant, la méthode K-Means, le comportement du conducteur. Plus précisément, si cette observation appartient au cluster 0, l'application va mentionner que c'est un bon conducteur sinon elle va mentionner qu'il est un mauvais conducteur (voir Figure 29) .

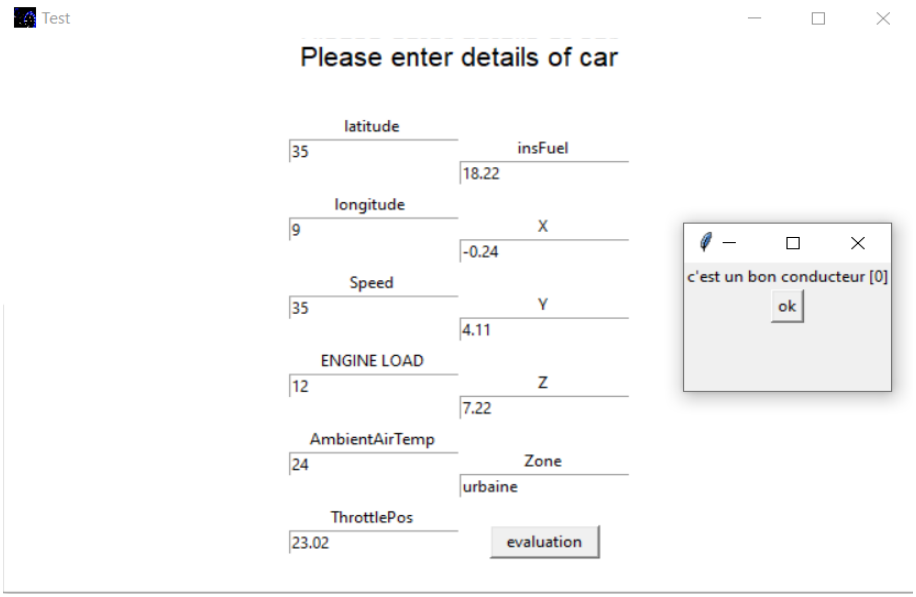


Figure 27 Message d'évaluation

Conclusion

Dans ce chapitre, on a essayé de développer, en utilisant le modèle réalisé au chapitre précédent, une application permettant d'évaluer le comportement d'un conducteur.

Conclusion Générale

Tout au long de la préparation de notre projet tuteuré, nous avons essayé de mettre en pratique les connaissances acquises durant nos études et nos séances d'encadrements dans le but de réaliser une application d'évaluation de la qualité de conduite. Au cours de cette rapport, nous avons commencé par réaliser une étude descriptive des différentes variables intervenant dans le dataset, puis nous avons préparé notre base tout en répondant aux problèmes des valeurs manquantes et la transformation des variables catégoriques en des variables numériques afin de pouvoir les intégrer dans des équations mathématiques. Une fois terminé, notre nouvelle base est prête pour le montage des différents modèles d'apprentissages non-supervisés qui vont servir à la classification du comportement des conducteurs. Plus précisément les modèles "K-means" et "Hierarchical aggromelative clustering" implémenté ont été mis en place afin d'avoir un apprentissage optimal qui permet de prédire à partir d'une nouvelle observation si ce conducteur est un chauffeur parfait ou non. Le but de notre application est de faciliter le contrôle des chauffeurs, pour cela elle peut être utilisée par les agences de location de voiture s'ils veulent surveiller leurs voitures ou bien d'être une solution pour l'état qui minimise les accidents .

A ce stade, nous affirmons que notre projet est une étape qui peut être poussée plus loin pour ouvrir la voie à de futurs travaux visant à rendre l'application plus prometteuses en trouvant un moyen pour récolter des observations successives et périodiques et faire une évaluation journalière du conducteur. De plus, nous pensons qu'il est intéressant de concrétiser ce travail en essayant de construire un prototype qui servira à contrôler le comportement des conducteurs à distance.

Webliographie

- [1] :https://www.netapp.com/fr/artificial-intelligence/what-is-artificial-intelligence/#:~:text=L'intelligence%20artificielle%20(IA)%20est%20un%20processus%20d',agir%20comme%20des%20%C3%AAtres%20humains. (Consulté le 3/07/2021)
- [2] : https://www.ibm.com/fr-fr/analytics/machine-learning (Consulté le 3/07/2021)
- [3] : https://www.24pm.com/117-definitions/512-apprentissage-supervise (Consulté le 3/07/2021)
- [4] : https://dataanalyticspost.com/Lexique/apprentissage-non-supervise/ (Consulté le 3/07/2021)
- [5] : https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501303-analyse-en-composantes-principales-acp-definition-et-cas-d-usage/ (Consulté le 3/07/2021)
- [6] : https://analyticsinsights.io/le-clustering-definition-et-implementations/ (Consulté le 3/07/2021)
- [7] : http://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf (Consulté le 3/07/2021)
- [8] : https://ichi.pro/fr/methode-silhouette-mieux-que-la-methode-elbow-pour-trouver-des-clusters-optimaux-61080390822033#:~:text=La%20m%C3%A9thode%20Elbow%20est%20une,chacune%20des%20valeurs%20de%20k. (Consulté le 3/07/2021)
- [9] : https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379561-partitionnez-vos-donnees-avec-un-algorithme-de-clustering-hierarchique (Consulté le 3/07/2021)
- [10] : https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825 (Consulté le 4/07/2021)
- [11] : https://www.lebigdata.fr/machine-learning-et-big-data (Consulté le 4/07/2021)
- [12] : https://mrmint.fr/donnees-manquantes-data-science (Consulté le 4/07/2021)
- [13] : https://www.math.univ-toulouse.fr/~besse/pub/Appren_stat.pdf (Consulté le 4/07/2021)

- [14] : https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/#:~:text=Python%20est%20le%20langage%20de,domaine%20du%20d%C3%A9veloppement%20de%20logiciels.&text=Ainsi%2C%20d%C3%A9velopper%20du%20code%20avec,'avec%20d'autres%20langages. (Consulté le 4/07/2021)
- [15] : https://koor.fr/Python/Tutorial/python_ide_spyder.wp (Consulté le 4/07/2021)
- [16] : https://www.lucidchart.com/pages/fr/landing/outil-de-diagramme-uml-en-ligne?utm_source=google&utm_medium=cpc&utm_campaign=_fr_allcountries_desktop_search_nb_bmm_&km_CPC_CampaignId=2081333709&km_CPC_AdGroupID=80369426081&km_CPC_Keyword=%2Buml%20%2Blogiciel&km_CPC_MatchType=b&km_CPC_ExtensionID=&km_CPC_Network=g&km_CPC_AdPosition=&km_CPC_Creative=506539085680&km_CPC_TargetID=kwd-314331881428&km_CPC_Country=9073723&km_CPC_Device=c&km_CPC_placement=&km_CPC_target=&mkwid=soyFOi43T_pcrid_506539085680_pkw_%2Buml%20%2Blogiciel_pmt_b_pdv_c_slid__pgrid_80369426081_ptaid_kwd-314331881428_&gclid=Cj0KCQjwxJqHBhC4ARIsAChq4avr-Q89rtqRE2lecoWgLXMq0K83Odxtht_w0zPFNrC1CwN8Mv-1Y80aAkCAEALw_wcB (Consulté le 4/07/2021)

Implémentation d’une application de détection des comportements de conduite à l’aide des algorithmes de « Machine Learning »

Mohamed Chabchoub Bilel Kammoun

Résumé

Ce projet a pour but de développer une application qui permet de classifier le comportement d'un chauffeur. Pour se faire, on a exploité un dataset composé de 95549 observations et 12 variables explicatives. L'étude descriptive et le prétraitement de cette base nous ont permis d'appliquer des algorithmes d'apprentissage non supervisé, à savoir K-Means et Hierarchical Agglomerative Clustering, afin de prédire la classe d'un conducteur. Cette classification permettra de minimiser le nombre d'accidents et de bien contrôler les chauffeurs.

Mots clés

Analyse en composantes principales ; Apprentissage non supervisé ; Regroupement K-moyennes ; regroupement hiérarchique agglomératif.

Abstract

The aim of this project is to develop an application which allows to classify the behavior of a driver. To do this, we used a dataset composed of 95,549 observations and 12 features. The descriptive study and the preprocessing of this database allowed us to apply unsupervised learning algorithms, such as, in order to predict the class of a conductor. This classification will make it possible to minimize the number of accidents and to control the drivers.

Keywords

Principale analysis composant ; unsupervised Learning ; K-means clustering ; Hierarchical agglomerative clustering.

تلخيص

الهدف من هذا المشروع هو تطوير تطبيق يسمح بتصنيف سلوك السائق و للقيام بذلك ، استخدمنا مجموعة بيانات تتكون من 95549 ملاحظة و 12 متغيرًا توضيحيًا. سمحت لنا الدراسة الوصفية والمعالجة المسبقة للقاعدة من تطبيق خوارزميات التعلم غير الخاضعة للإشراف ، وهي « Hierarchical Agglomerative Clustering » و « K-Means » للتنبؤ بفئة السائق. هذا التصنيف سيجعل من الممكن تقليل عدد الحوادث ومراقبة السائقين .