REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DÉPARTEMENT D'INFORMATIQUE

MEMOIRE DE FIN D'ETUDE

Présenté en vue de l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Informatique Décisionnelle et Optimisation

Présenté par :
Lamara Soumia
Saiahi Khedidja

THEME

Utilisation des Modèles de Machine Learning pour la Prédiction du Trafic Routier

Soutenu publiquement le : / .. /2021

Devant le jury:

Kadri Said	**Président**	**Université de M'sila**
Mehenni Tahar	**Rapporteur**	**Université de M'sila**
Amroune Nasereddine	**Examinateur**	**Université de M'sila**

Promotion : 2020/2021

Dédicaces

A mes très chers parents

Remerciements

Nous remercions Dieu le tout puissant qui nous a donné le courage et la force pour aboutir à L'accomplissement de ce travail.

Nous voudrions exprimer toute mes reconnaissances et nos respects à Monsieur Mehenni Tahar pour son orientation.

Nous remercions les membres du jury pour l'honneur qu'ils nous ont attribué en acceptant d'évaluer et de juger ce modeste travail.

Nous remercions mes amis pour leur aide appréciable, leurs encouragements continus et leur soutien moral ininterrompu.

Ce travail n'aurait jamais été possible sans le soutien et l'appui moral des membres de nos familles. Nous les remercions tous.

TABLE DES MATIERES

CHAPITRE 1 : TRAFIC ROUTIER

CHAPITRE 2 : PREDICTION DE TRAFIC ROUTIER

CHAPITRE 3 : MODELES PROPOSÉS POUR LA PREDICTION DU TRAFIC ROUTIER

Liste Des Figures

Liste Des Tableaux

ملخص:

يشكل ازدحام الطرق مشكلة تعاني منها معظم المدن، خاصة الكبيرة منها. ولأجل ذلك، أجرينا دراسة تتعلق بمشكلة ازدحام الطريق وتوقعها باستخدام نماذج التعلم الآلي للتنبؤ بحركة المرور على الطرق. وقد اعتمدنا على تصنيف حركة المرور باستخدام خوارزميات بايز والانحدار، كما طبقنا ما اقترحناه من تقنيات على بيانات حقيقية.

الكلمات المفتاحية :حركة المرور على الطرق، naivebayes ،الانحدار،التعلم الآلي،التصنيف

Résumé :

Le problème de congestion des routes constitue un défi pour les grandes villes. Dans cette optique, nous proposons d'étudier le problème de prédiction du trafic routier en utilisant les différentes méthodes de machine learning. Nous avons appliqué les méthodes de régression et la classification bayésienne naïve sur des données réelles.

Mots clés : trafic routier, naivebayes, régression, machine learning, classification.

Abstract:

Road congestion is a real problem in the big cities. We aim to study this problem using machine learning techniques in order to predict the road traffic. We used regression models and Naïve Bayes classification algorithms, and apply them on real data.

Keywords: road traffic, naivebayes, regression, machine learning, classification.

INTRODUCTION GENERALE

Le trafic automobile est un problème majeur dans les sociétés modernes. Des millions d'heures et des gallons de carburant sont gaspillés tous les jours par des véhicules bloqués dans le trafic. Cette considération a conduit les ingénieurs et les scientifiques à mettre au travail récemment pour détecter la congestion du trafic et proposer des solutions visant à réduire les effets indésirables. La détection de la congestion n'est qu'une de nombreuses applications de trafic routier et il n'est pas conçu pour être utilisé comme moyen pour conduite automatisée, mais plutôt comme un outil pour fournir des informations au conducteur qui aidez-lui à prendre des décisions pour éviter le trafic lourd. Les congestions de trafic sont formées par de nombreux facteurs. Certains sont (d'une certaine manière) prévisibles comme la construction de routes, les heures de pointe ou les cols de bouteilles et certains sont imprévisibles comme les accidents, la météo et le comportement humain.

Par ailleurs, beaucoup de travaux portant sur la définition puis la détection de la congestion ont été effectués durant ces dernières années, Et nous en décrirons certaines dans notre mémoire.

L’objectif principal de ce mémoire est la prédiction de la congestion dans les réseaux routiers.

Dans le premier chapitre, nous allons aborder le trafic routier et ces problèmes, ainsi que les méthodes de modélisation et la découverte de la congestion.

Le deuxième chapitre, est consacré prévisions de circulation ainsi qu’une taxonomie des approches existantes, ensuite nous présentons plusieurs orientations futures pour les prévisions de trafic.

Dans le dernier chapitre, nous présentons les données et les différents modèles de régression et de classification que nous avons utilisée afin de prédire le trafic routier selon différentes variantes. Enfin, nous clôturons ce mémoire par une conclusion générale.

CHAPITRE 1

TRAFIC ROUTIER

1. Introduction

Les systèmes de transport sont des systèmes complexes. L'optimisation de leur gestion nécessite une bonne compréhension du fonctionnement de ces systèmes, et un développement de stratégies efficaces de gestion et de régulation du trafic la prévision du trafic joue un rôle essentiel dans la Système de transport .Des prévisions précises du trafic peuvent aider à planifier l'itinéraire, guider la répartition des véhicules et atténuer les embouteillages.

Ce problème est difficile en raison de la complexité et dépendances spatio-temporelles dynamiques entre différents régions du réseau routier. Récemment, une quantité importante des efforts de recherche ont été consacrés à ce domaine, faisant progresser considérablement capacités de prévision de la circulation. Le but de ce chapitre est de fournir une enquête complète pour la prévision du trafic.

Plus précisément, nous résumons d'abord les méthodes de prévision du trafic existantes et en donnons une taxonomie. Ensuite, nous énumérons les tâches courantes de prévision du trafic et l'état de l'art dans ces tâches. Enfin, nous collectons et organisons des ensembles de données publiques largement utilisés dans la littérature existante.

De plus, nous donnons une évaluation en effectuant des expériences pour comparer les performances de différentes méthodes liés à la demande de trafic et à la prévision de la vitesse respectivement sur deux ensembles de données.

2. Le trafic routier

Mouvement des véhicules sur la voirie (par extension, s'applique aussi au mouvement des piétons ou à celui des trains sur un réseau ferré) Le terme de trafic (anglais : trafic) n'est pas tout à fait synonyme : il désigne le volume de la circulation. On distingue la circulation interne à une agglomération, d'échange entre agglomérations et de transit à travers une agglomération. La circulation interne à une agglomération qui est spécifiquement urbaine, représente environ 60% du trafic dans une petite ville et jusqu'à 95% dans une très grande agglomération. La circulation est cause de nuisances importantes (bruit, pollution de l'air, accidents, coupure du tissu urbain, dégradation du paysage).[1]

2.1 Route

Itinéraire à suivre pour aller d'un endroit à un autre. C'est une voie carrossable destinée à la liaison entre les localités et à la desserte des zones rurales .Elle permettent à l'homme de se déplacer de chez lui pour pénétrer d'autres régions plus ou moins lointaines. Les routes comportent, outre la chaussée, les fossés, talus, bandes d'arrêts, pistes cyclables et plantations éventuelles. [1]

2.2 Voie

Vient du latin via, qui a donné également le verbe voyagé et le mot voyage, Voie : c'est le chemin, la route, par laquelle on se rend d'un lieu à un autre.

C'est un espace aménagé pour se déplacer en ville (voie urbaine) entre les localités ou en milieu rural (routes).

Les voies de communication sont par conséquent les moyens de liaison à travers l'espace.

La voie permet de circuler dans les deux sens, mais parfois, surtout en ville, pour accroître le débit du réseau de voirie, dans un sens unique.

La voie comporte des trottoirs (1,5 m est un minimum souhaitable) éventuellement plantés d'arbres (5 m de largeur minimale). [2]

3. Les problèmes de trafic routier

La congestion urbaine du trafic existe depuis au moins la première révolution urbaine, celle du XIIe siècle. « Le mal n'est pas nouveau mais il a gagné considérablement, depuis une douzaine d'années, en profondeur et en étendue. Alors qu'il ne sévissait naguère que dans les quartiers d'affaires de quelques grandes métropoles. Malgré les progrès techniques et technologiques accomplis par l'homme dans tous les domaines de la connaissance, le trafic routier reste victime d'une congestion sans cesse croissante. Il sera sans doute bientôt possible de voyager jusqu'à la planète Mars, mais la congestion continue à demeurer un véritable casse-tête pour les gestionnaires routiers de tous les pays. La congestion d'un réseau routier est la condition dans laquelle une augmentation du trafic de véhicules provoque un ralentissement global de celui-ci. Le terme de congestion désigne la dégradation de la qualité de service quand le nombre d'usagers augmente. [3]

Ce phénomène se caractérise par l'apparition de retards, voire de goulets d'étranglement en période de fort trafic, c'est-à-dire quand la capacité de l'infrastructure devient insuffisante pour réguler les flux. Le problème est fréquent localement et périodiquement, notamment dans les grandes villes et lors des grands départs pour les vacances. [3]

4. Modélisation de trafic routier par les graphes

Dans cette section, nous décrivons la conception connexe du modèle des graphes et la définition du problème que nous résoudrions. Respectivement, un graphe G est une paire (V, E), où V est un ensemble fini et E est une relation binaire sur V. L'ensemble V est appelé l'ensemble de sommets de G, noté V (G), et ses éléments sont appelés sommets. L'ensemble E est appelé l'ensemble d'arêtes de G, noté E (G), et ses éléments sont appelés bords. [4]

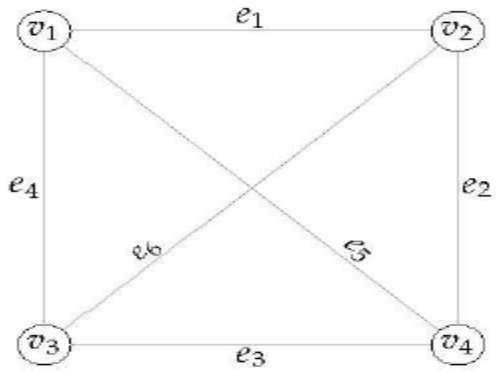


Figure I.1 : Exemple de graphe G. [4]

5. Méthode de modélisation du trafic d'un instant donné

Le trafic en un moment (horaire fixe ou instant donné) se compose de facteurs comprenant les voitures, les piétons et les vélos, peu importe ce qu'ils sont, nous les laissons être les sommets d'un graphe G, quel que soit l'impact d'un sommet à l'autre, laissez les arêtes exister dans les sommets adjacents les uns aux autres, aucune arête ne sort si les sommets ne sont pas directement adjacents. Nous ne prêtons attention qu'aux sommets et aux arêtes entre les sommets de la structure des facteurs de trafic que le modèle graphique décrit, et ignorons l'angle des arêtes, car le principal facteur qui a un impact sur le trafic est la connexion des sommets, pas le l'angle des bords. [4]

6. Structure graphique du trafic

Afin d'être pratique, nous laissons les sommets d'un graphe remplacer le véhicule et le piéton. Ainsi, lorsque nous établissons le modèle de graphe G pour le trafic, nous constatons que si le trafic est encombré, le graphe G doit inclure des sommets dans lesquels on n'a pas moins de six arêtes avec d'autres sommets. [4]

Nous étudions la condition simple du trafic. La figure 2 est un modèle de graphique, et dans le graphique, les cercles représentent les objets en mouvement sur la route à un moment donné. Si le trafic est entravé, il y a beaucoup de sommets du graphe pour le trafic qui sont liés à pas moins que des arêtes de paramètres. Les sommets avec plus de paramètres d'arêtes dans le graphe G sont désignés par le facteur d'encombrement des sommets (VCF). Le graphe composé par VCF est désigné par la congestion de graphe (GC). Le sous-graphe d'un modèle de graphe pour l'isomorphisme du trafic vers le GC est noté STGC, le paramètre est donné par le peuple. [4]

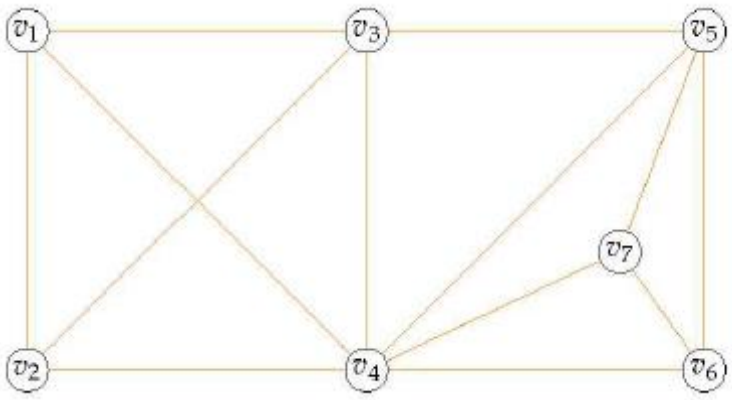


Figure I.2 : Modèle graphique pour le trafic à un instant. [4]

7. Algorithme pour la découverte de la congestion

7.1. Vue d'ensemble de l'algorithme approximatif

Il existe deux approches de base de l'exploitation minière fréquente des sous-structures, les approches de croissance de modèle et les approches basées sur apriori qui doivent utiliser la stratégie de recherche en largeur d'abord (BFS) en raison de sa génération de candidats par niveau. Afin de déterminer si un graphe de taille (k + 1) est fréquent; il doit vérifier tous ses sous-graphes de taille k correspondants pour obtenir une limite supérieure de sa fréquence. [4]

Les algorithmes basés sur apriori sont considérablement complexes, afin d'éviter les complexes des algorithmes de croissance de modèle ont été développés. L'approche peut utiliser la recherche en largeur d'abord ainsi que la recherche en profondeur d'abord (DFS), cette dernière consommant moins de mémoire. Mais les deux approches ne sont pas assez efficaces, car elles doivent rechercher les structures et les sommets du graphe pour le trafic. La recherche de structures prend beaucoup de temps. Ainsi, dans ce chapitre, nous avons proposé un algorithme pour l'exploration fréquente des sous-structures du graphe pour le trafic. L'algorithme recherche d'abord les sommets, pas besoin de considérer la structure du sous-graphe au début. L'algorithme consomme donc moins que les deux approches mentionnées ci-dessus. Selon la fonctionnalité du GC et du STGC, nous trouvons que l'algorithme utilisant DFS est adapté à notre condition. [4]

Le DFS a trouvé tous les VCF adjacents, inspecté si ces nœuds pouvaient former GC. Un graphique composé du VCF de la recherche était autrefois un STGC du graphique pour le trafic, la probabilité de cette condition est grande. [4]

Évidemment, le STGC doit être constitué par VCF. Le GC est composé du VCF. STGC est un sous-graphe d'un isomorphisme de graphe vers le GC. Le STGC est donc constitué par VCF. Et pour trouver STGC, il suffit de trouver le VCF, pas besoin de faire attention à la structure au début. Le STGC est constitué par VCF, donc il n'y a pas de STGC si aucun VCF n'est trouvé.[4]

7.2. Algorithme de minage complet

L'algorithme de minage complet appelé CDA (Algorithme Congestion Discover). Ce algorithme est basé sur DFS. CDA traverse les sommets d'abord, s'il est trouvé le sommet «A» dans VCF, l'algorithme utilisant DFS pour rechercher les nœuds adjacents de «A» jusqu'à ce qu'il ne trouve pas le VCF non traversé en une fois. Stocker le VCF trouvé dans une fois la recherche et dans cette recherche ces VCF sont continus, non interrompus par le non-VCF. Vérifiez ensuite si la structure du VCF pourrait composer un STGC. [4]

En utilisant la même méthode pour trouver s'il y a un STGE près du STGC que nous avons trouvé en dernier recherche de temps. Lorsque la structure du STGC atteint 30% de la structure du graphe recherché, on pense que le trafic est en état d'encombrement. Au fur et à mesure que l'ADC progresse, chaque sommet a une couleur:

- WHITE : signifie non découvert.
- GRAY : signifie découvert, mais pas fini (pas fini d'explorer à partir de celui-ci).
- BLACK : signifie terminé (avoir trouvé tout ce qui était accessible à partir de celui-ci) Soit De [u] une valeur booléenne. Si D [u] vaut 1, le sommet u a plus de six arêtes, sinon moins. [4]

Algorithm CDA

Input: a graph G (V,E), VCF's parameter

Output: STGC

(1) for each u ∈V

(2) do color[u] ← WHITE, D[u] ← 0

(4) for each u ϵV

(5) do if color[u] = WHITE

(6) then DFS-VISIT(u)

(7) scan all the vertices of G, find out all the vertices that D[u] equal 1

(8) use Close Graph algorithm find STGC from graph is composed of VCF

Algorithm DFS-VISIT(u)

Input: a vertex u of a graph

Output: all vertices in one path from u

(1) color[u] ← GRAY

(2) if u has more than six edges

(3) then D[u] ← 1

(4) for each v ϵAdj[u]

(5) do if color[v] = WHITE

(6) then DFS-VISIT(v)

(7) color[u] ← BLACK [4]

8. Conclusion

Ce chapitre a étudié le problème du trafic routier et modélisé le graphique du trafic statique. Un modèle graphique du trafic en un instant a été étudié. Une nouvelle structure du graphique pour le trafic qui a nommé le STGC a été proposé. Nous avons donné des conceptions sur le VCF et le STGC, puis les conceptions ont été expliquées. Selon la caractéristique du STGC, nous formons notre algorithme noté CDA basé sur des algorithmes de recherche DFS. Après analyse, notre approche est plus efficace que les algorithmes d'extraction de sous-graphes fréquents. Ainsi, l'utilisation de notre méthode peut mieux décrire le trafic statique et retrouvez plus facilement la congestion du trafic. Notre approche peut fournir un guide pour soulager la congestion du trafic.

CHAPITRE 2

PREDICTION DU TRAFIC ROUTIER

1. Introduction

La modélisation du trafic urbain dans les villes intelligentes est une activité étudiée dans les systèmes de transport intelligents. Les études antérieures sur ce sujet peuvent être globalement classées en deux types : la prévision du trafic et la découverte de modèles de trafic.

Les études sur la prévision du trafic se sont principalement concentrées sur la prévision de la longueur de la congestion routière sur un segment de route à un moment donné en tenant compte de paramètres tels que le type de route, les données de trafic et météorologiques, le trafic et les accidents, etc. D'autre part, des études sur la découverte de modèles de trafic visant à identifier les événements (par exemple, les accidents de la circulation, les catastrophes naturelles et les horaires de la journée) qui peuvent influencer la circulation et les embouteillages. L'adoption populaire et l'application industrielle réussie de ces modèles de trafic urbain a été entravée par les limites:

1- La plupart des modèles de prévision du trafic peuvent prédire efficacement la congestion sur un segment de route particulier à un moment donné. Cependant, ils sont insuffisants pour fournir des informations globales concernant le comportement de la congestion sur un réseau à une durée donnée.

2- La plupart des modèles de découverte de modèles de trafic précédents ne tenaient compte que de la fréquence et ignoraient les informations d'occurrence temporelle des événements dans la base de données .Par conséquent, ces modèles sont insuffisants pour découvrir des régularités périodiques dans les bases de données temporelles.

2. Enquêtes connexes sur les prévisions de circulation

Il existe quelques enquêtes récentes qui ont passé en revue la littérature sur la prévision de la circulation dans certains contextes à partir de perspectives différentes. Passé en revue les méthodes et les applications de 2004 à 2013, et discuté de dix défis importants à l'époque .Il est davantage axé sur la prise en compte de la prévision du trafic à court terme et les littératures impliquées sont principalement basées sur les méthodes traditionnelles. [5]

Un autre travail c'est également intéressé à la prévision de la circulation à court terme, qui a brièvement présenté les techniques utilisées dans la prévision de la circulation et a donné quelques suggestions de recherche .Fourni des sources d'acquisition de données de trafic, et

principalement axé sur les méthodes traditionnelles d'apprentissage automatique. A souligné l'importance et les directions de recherche de la prévision de la circulation. [5]

A résumé des modèles pertinents basés sur des méthodes classiques et quelques méthodes d'apprentissage en profondeur précoces. "Alexander et coll." a présenté une enquête sur le réseau de neurones profonds pour la prévision du trafic. Il a discuté de trois architectures neuronales profondes communes, y compris le réseau neuronal convolutif, le réseau neuronal récurrent et le réseau neuronal "feed forward". [5]

Cependant, certaines avancées récentes, par exemple l'apprentissage profond basé sur des graphiques, n'ont pas été couvert dans est un aperçu de l'apprentissage profond basé sur des graphes architecture, avec des applications dans le domaine du trafic général, a fourni une enquête axée spécifiquement sur l'utilisation de modèles d'apprentissage pour l'analyse des données de trafic. [5]

Il étudie uniquement la prévision du flux de trafic. En général, différentes tâches de prévision du trafic ont des caractéristiques communes, et il est bénéfique de les considérer conjointement. Par conséquent, il y a encore un manque d'enquête large et systématique sur l'exploration de la prévision de la circulation en général. [5]

3. Machine Learning et data mining

3.1 Data mining

Historiquement, les premières approches statistiques étudient un petit nombre "n" d'individus décrits par un petit nombre "p" de variables. Ces données sont issues de plans d'expériences.

1990s (Mo) : les entreprises commencent à stocker de plus en plus de données concernant leurs clients, sans planification expérimentale. Les méthodes statistiques classiques sont massivement utilisées pour extraire de la connaissance de ces données (CRM, gestion de la relation client). C'est la naissance du data mining.

2000s (Go): première révolution du data mining avec l'avènement de la bioinformatique et des données ohmiques : on observe beaucoup de variables sur peu d'individus ($n << p$). On parle du fléau de la dimension et on doit développer de nouvelles méthodes parcimonieuses.

2010s (To):seconde révolution due au développement d'internet (commerce en ligne, réseau sociaux). On parle de" big data" (volume variété vélocité...) et de science des données. [6]

3.1.1 Les étapes de data mining

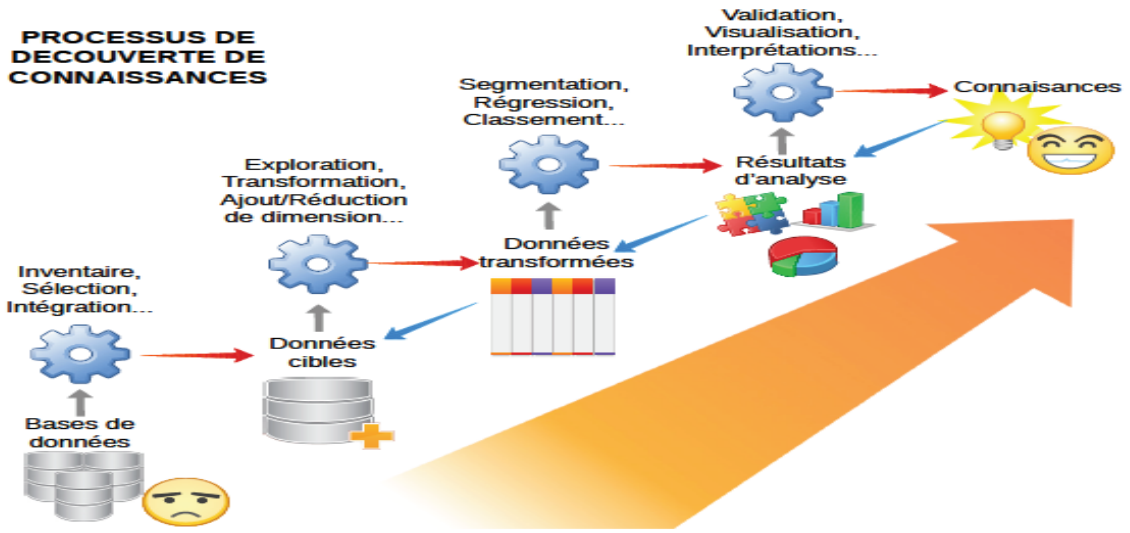


Figure II.1: Les étapes de data mining.[6]

3.1.2 Exemples d'applications réelles

- Vente, marketing
  - ✓ gestion de la relation client (scoring, score d'appétence).
  - ✓ segmentation de la clientèle.
- Banque, finance, assurance
  - ✓ détection de fraude (comportements atypiques).
  - ✓ score de risque (attribution ou non d'un crédit).
- Technologie
  - ✓ reconnaissance faciale dans une image.
  - ✓ reconnaissance de la parole.
- Médecine, industrie pharmaceutique.
  - ✓ réponse d'un patient vis-à-vis d'un traitement.
  - ✓ identification des facteurs de risques.
- Energie, transport...
  - ✓ prévision de consommation d'électricité.
  - ✓ prévision de trafic routier. [6]

3.1.3 Méthodes de data mining

Figure II.2:data mining.[6]

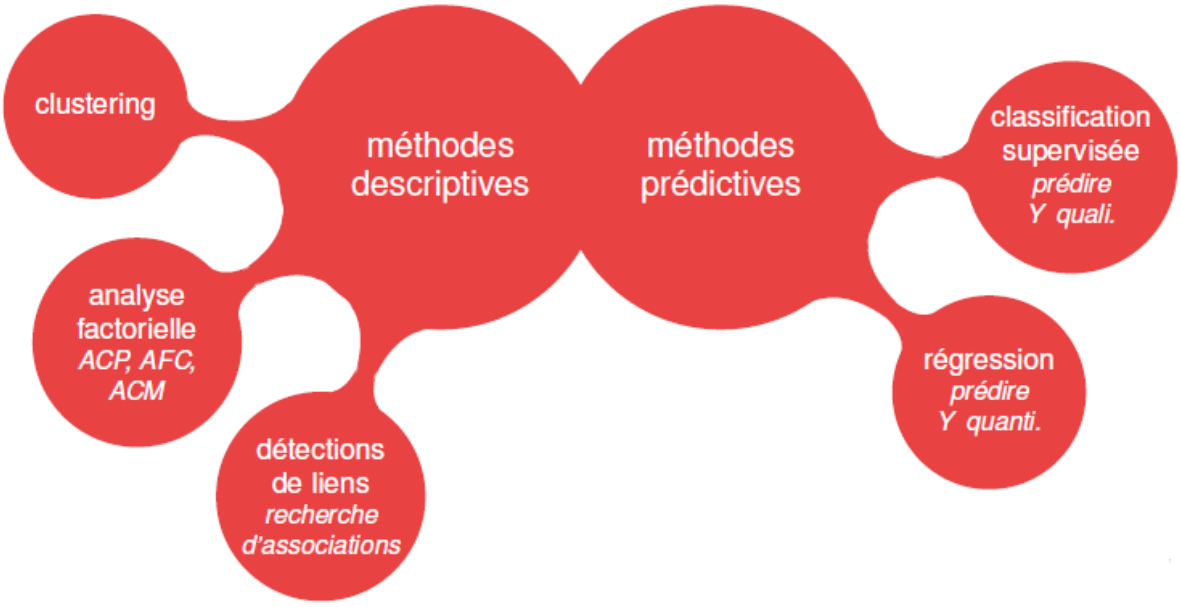


Figure II.3:Méthodes de data mining. [6]

3.2 Machine Learning

Le Machine Learning est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Pour apprendre et se développer, les ordinateurs ont toutefois besoin de données à analyser et sur lesquelles s'entraîner. De fait, le "Big Data" est l'essence du Machine Learning, et c'est la technologie qui permet d'exploiter pleinement le potentiel du "Big Data". Découvrez pourquoi cette technique et le "Big Data" sont interdépendants.

Le Machine Learning ne date pas d'hier, sa définition précise demeure encore confuse pour de nombreuses personnes. Concrètement, il s'agit d'une science moderne permettant de découvrir des patterns et d'effectuer des prédictions à partir de données en se basant sur des statistiques, sur du forage de données, sur les reconnaissances de patterns et sur les analyses prédictives. Les premiers algorithmes sont créés à la fin des années 1950. Le plus connu d'entre eux n'est autre que le perceptron. [15]

3.2.1 Les différents types d'algorithmes de Machine Learning

On distingue différents types d'algorithmes Machine Learning, généralement, ils peuvent être répartis en deux catégories : supervisés et non supervisés. Dans le cas de l'apprentissage supervisé, les données utilisées pour l'entraînement sont déjà étiquetées. Par conséquent, le modèle de Machine Learning sait déjà ce qu'elle doit chercher (motif, élément…) dans ces données.

À la fin de l'apprentissage, le modèle ainsi entraîné sera capable de retrouver les mêmes éléments sur des données non étiquetées. Parmi les algorithmes supervisé, on distingue les algorithmes de classification (prédiction non numériques) et les algorithmes de régression (prédictions numérique).en fonction du problèmes a résoudre, on utilisera l'un de ces deux archétypes**.**

L'apprentissage non supervisé, au contraire, consiste à entraîner le modèle sur des données sans étiquettes. La machine parcourt les données sans aucun indice, et tente d'y découvrir des motifs ou des tendances récurrentes. Cette approche est couramment utilisée dans certains domaines, comme le cyber sécurité. Parmi les modèles non-supervisés, on distingue les algorithmes de "clustering" (pour trouver des groupes d'objets similaires), d'association (pour trouver des liens entre des objets) et de réduction dimensionnelle (pour choisir ou extraire des caractéristiques).[15]

4. Une taxonomie des approches existantes

Après des années d'efforts, la recherche sur la prévision du trafic a réalisé de grands progrès. À la lumière du processus de développement, ces méthodes peuvent être globalement divisées en deux catégories: les méthodes classiques et des méthodes basées sur l'apprentissage en profondeur. Les méthodes classiques incluent les méthodes statistiques et les méthodes traditionnelles d'apprentissage automatique. La méthode statistique consiste à construire un modèle statistique basé sur les données pour la prédiction.

Les algorithmes les plus représentatifs sont la moyenne historique (HA), la moyenne mobile intégrée autorégressive (ARIMA) et la moyenne autorégressive vectorielle(VAR) . Néanmoins, ces méthodes nécessitent des données pour satisfaire certaines hypothèses, et les données de trafic variant dans le temps sont trop complexes pour satisfaire ces hypothèses. De plus, ces méthodes ne sont applicables qu'à des ensembles de données relativement petits. Plus tard, un certain nombre de méthodes traditionnelles d'apprentissage des machines, telles que la régression vectorielle de soutien (SVR) et la régression forestière aléatoire (RFR), ont été proposées pour les problèmes de prévision du trafic. [5]

Ces méthodes ont la capacité de traiter des données de grande dimension et de capturer des relations non linéaires complexes .Il a fallu attendre l'avènement des méthodes basées sur l'apprentissage profond que le plein potentiel de l'intelligence artificielle dans la prévision du trafic a été développé. Cette technologie étudie comment apprendre un modèle hiérarchique pour mapper directement l'entrée d'origine sur la sortie attendue. En général, les modèles d'apprentissage en profondeur empilent des blocs ou des couches d'apprentissage de base pour former une architecture profonde, et l'ensemble du réseau est formé de bout en bout.

Plusieurs architectures ont été développées pour traiter des données spatio-temporelles complexes et à grande échelle. Généralement, le réseau neuronal convolutif (CNN) est utilisé pour extraire la corrélation spatiale des données structurées en grille décrites par des images ou des vidéos, et le réseau convolutionnel graphique (GCN) étend l'opération de convolution à des données structurées graphiquement plus générales, qui sont plus approprié pour représenter la structure du réseau de trafic. [5]

De plus, le réseau neuronal récurrent (RNN).Nous résumons ici les techniques clés couramment utilisées dans les méthodes de prévision du trafic existantes. [5]

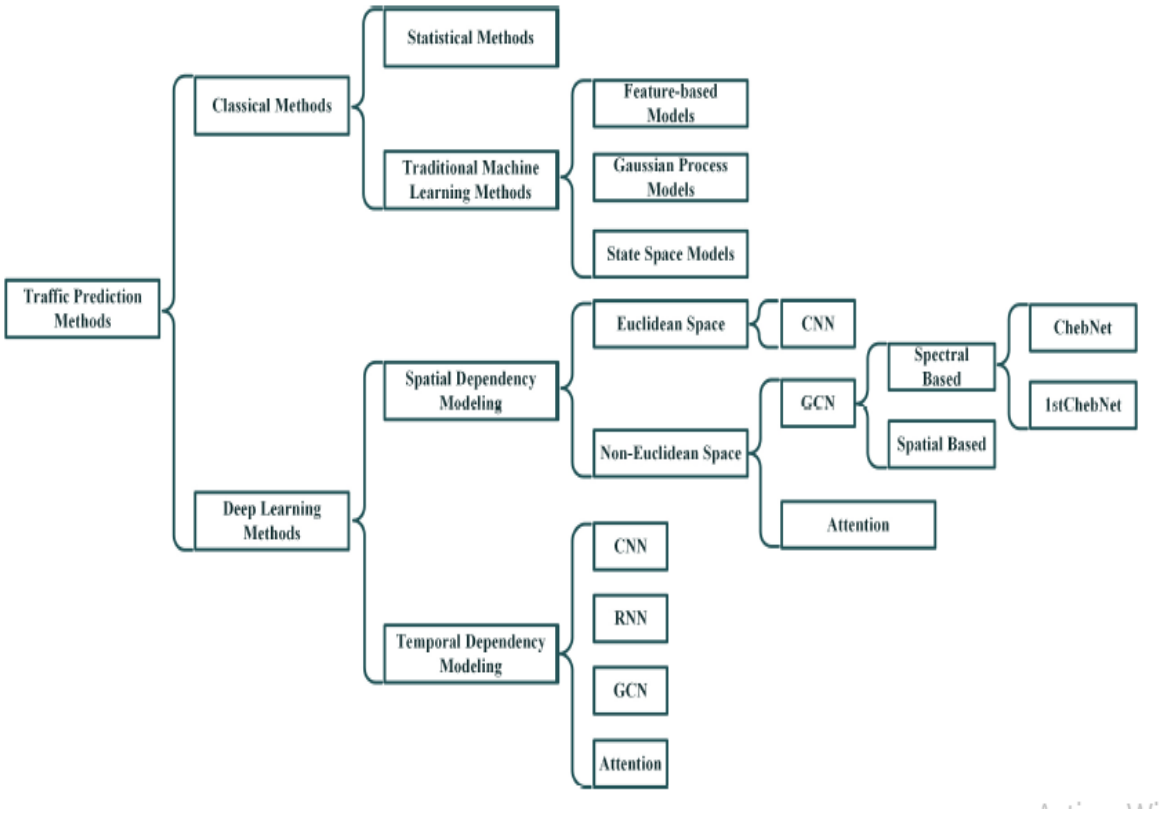


Figure II.4: Techniques clés des méthodes de prévision du trafic. [5]

4.1. Méthodes classiques

Les modèles d'apprentissage automatique statistiques et traditionnels sont deux principales méthodes représentatives basées sur les données pour la prévision du trafic .Dans l'analyse des séries chronologiques, la moyenne mobile intégrée autorégressive (ARIMA) et ses variantes sont l'une des approches les plus consolidées basées sur des statistiques classiques et ont été largement appliquées pour les problèmes de prévision du trafic.

Cependant, ces méthodes sont généralement conçues pour de petits ensembles de données, et ne conviennent pas pour traiter des données chronologiques complexes et dynamiques. De plus, étant donné que seules les informations temporelles sont généralement prises en compte, la dépendance spatiale des données de trafic est ignorée ou à peine prise en compte.

Les méthodes traditionnelles d'apprentissage automatique, qui peuvent modéliser des données plus complexes, sont globalement divisées en trois catégories : [5]

Modèles basés sur des fonctionnalités, modèles de processus gaussiens et modèles d'espace d'états. Les méthodes basées sur les caractéristiques résolvent le problème de la prévision du trafic en entraînant un modèle de régression basé sur des caractéristiques de trafic conçues par l'homme. Ces méthodes sont simples à mettre en œuvre et peuvent fournir des prédictions dans certaines situations pratiques. Le processus gaussien modélise les caractéristiques internes des données de trafic à travers différentes fonctions du noyau, qui doivent contenir simultanément des corrélations spatiales et temporelles.

Bien que ce type de méthodes se soit avéré efficace et réalisable dans la prévision du trafic, comparé aux modèles basés sur les caractéristiques, ils ont généralement une charge de calcul et une pression de stockage plus élevées, ce qui n'est pas approprié lorsqu'une masse d'échantillons d'apprentissage sont disponibles. Les modèles d'espace d'état supposent que les observations sont générées par des états cachés markoviens.

L'avantage de ce modèle est qu'il permet de modéliser naturellement l'incertitude du système et de mieux capter la structure latente des données spatio-temporelles. Cependant, la non-linéarité globale de ces modèles est limitée et la plupart du temps ils ne sont pas optimaux pour modéliser des données de trafic complexes et dynamiques. [5]

En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives. On parle aussi de modèle linéaire ou de modèle de régression linéaire. Parmi les modèles de régression linéaire, le plus simple est l'ajustement affin. Celui-ci consiste à rechercher la droite permettant d'expliquer le comportement d'une variable statistique y comme étant une fonction affine d'une autre variable statistique x.

En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de y connaissant x est une fonction affine des paramètres. Cependant, on peut aussi considérer des modèles dans lesquels c'est la médiane conditionnelle de y connaissant x ou n'importe quel quantile de la distribution de y connaissant x qui est une fonction affine des paramètres. [7]

4.1.1 Régression linéaire

La régression est une technique utilisée pour modéliser et analyser les relations entre les variables et souvent la façon dont elles contribuent et sont liées à la production d'un résultat particulier ensemble. Une régression linéaire fait référence à un modèle de régression entièrement composé de variables linéaires.

La forme générale de l'équation de régression linéaire simple est Y = a + bX + r, où 'X' est une variable indépendante, 'Y' est une variable dépendante, a 'est une intersection,' b 'est la pente de la droite et' r ' est un terme d'erreur. Cette équation peut être utilisée pour estimer la valeur de la variable de réponse (Y) sur la base des valeurs données de la variable prédictive (X) dans son domaine. [7]

4.1.2 Régression polynomiale

Lorsque nous voulons créer un modèle adapté à la gestion de données séparables non linéairement, nous devrons utiliser une régression polynomiale. Dans cette technique de régression, la ligne de meilleur ajustement n'est pas une ligne droite. C'est plutôt une courbe qui s'insère dans les points de données. Pour une régression polynomiale, la puissance de certaines variables indépendantes est supérieure à 1. [7]

La forme générale de l'équation de régression polynomiale est :

$$\boldsymbol{Y = a_1 X + a_2 X^2 + \cdots + a_n X^n + b} \qquad (1)$$

4.1.3 Régression Logistique

La régression logistique est utilisée pour trouver la probabilité d'un événement. On veut déterminer le succès ou l'échec d'un événement. On utilise la régression logistique lorsque la variable dépendante est de en semble de valeurs fini (0/1, Vrai / Faux, Oui / Non). Ici, la valeur de Y va de 0 à 1. [8]

4.2. Méthodes d'apprentissage profond

Les modèles d'apprentissage profond exploitent beaucoup plus de fonctionnalités et d'architectures complexes que les méthodes classiques et peuvent obtenir de meilleures performances. [5]

4.2.1 Modélisation de la dépendance spatiale

4.2.1.1. CNN

Une série d'études ont appliqué CNN (convolutional neural network) pour capturer les corrélations spatiales dans les réseaux de trafic à partir de données de trafic spatio-temporelles bidimensionnelles. Le réseau de trafic étant difficile à décrire par des matrices 2D, plusieurs recherches tentent de convertir la structure du réseau de trafic à différents moments en images et de diviser ces images en grilles standard, chaque grille représentant une région. De cette manière, les CNN peuvent être utilisés pour apprendre les caractéristiques spatiales entre différentes régions.[5]

4.2.1.2 GCN

Le CNN traditionnel est limité à la modélisation des données euclidiennes, et GCN (Graph convolutional network) est donc utilisé pour modéliser les données de structure spatiale non euclidienne, ce qui est plus conforme à la structure du réseau routier de trafic. GCN se compose généralement de deux types de méthodes, les méthodes spectrales et spatiales. Les approches basées sur le spectre définissent les convolutions de graphe en introduisant des filtres du point de vue du traitement du signal de graphe où l'opération de convolution de graphe est interprétée comme la suppression du bruit des signaux de graphe. Les approches spatiales formulent des convolutions de graphe sous forme d'agrégation d'informations sur les caractéristiques des voisins. Dans ce qui suit, nous présenterons respectivement les GCN spectraux et les GCN spatiaux. [5]

4.2.1.2.1 Méthodes spectrales

Bruna et coll. premier réseau spectral développé, qui a effectué l'opération de convolution pour les données de graphe du domaine spectral en calculant la composition propre de la matrice la placienne du graphe L. Plus précisément, l'opération de convolution de graphe G d'un signal x avec un filtre.[5]

4.2.1.2.2 Méthodes spatiales

Les méthodes spatiales définissent les convolutions directement sur le graphe à travers le processus d'agrégation qui opère sur le nœud central et ses voisins pour obtenir une nouvelle représentation du nœud central, comme le montre la figure II.5. Dans le réseau de trafic a d'abord été modélisé comme un graphe orienté, la dynamique du flux de trafic a été

capturée sur la base du processus de diffusion. Ensuite, une opération de convolution de diffusion est appliquée pour modéliser la corrélation spatiale, qui est plus interprétation intuitive et s'avère efficace dans la modélisation spatio-temporelle. Plus précisément, la convolution de diffusion modélise le processus de diffusion bidirectionnelle, permettant au modèle de capturer l'influence du trafic en amont et en aval. [5]

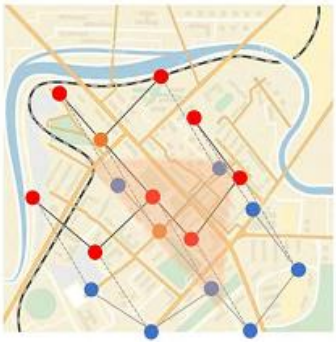


Figure II.5:Réseau de convolution de graphes spatiaux. Chaque nœud du graphique peut représenter une région du réseau de trafic. Pour obtenir une représentation cachée d'un certain nœud (par exemple, le nœud orange), GCN agrège les informations sur les caractéristiques de ses voisins (zone grisée). Contrairement aux données de grille dans les images 2D, les voisins d'une région ne sont pas ordonnés et varient en taille.[5]

4.2.1.3 Mécanisme attention

Le mécanisme d'attention est d'abord proposé pour les traitements du langage, et a été largement utilisé dans divers domaines. L'état de la circulation d'une route est affecté par d'autres routes avec des impacts différents. Un tel impact est très dynamique et évolue avec le temps. Pour modéliser ces propriétés, le mécanisme d'attention spatiale est souvent utilisé pour capturer de manière adaptative les corrélations entre les régions du réseau routier.

L'idée clé est d'attribuer dynamiquement différents poids à différentes régions à différents pas de temps. Par souci de simplicité, nous ignorons les coordonnées temporelles pour le moment. [5]

4.2.2 Modélisation de la dépendance temporelle

4.2.2.1 CNN

Ont d'abord introduit le modèle entièrement convolutif pour l'apprentissage de séquence en séquence. Un travail représentatif dans la recherche sur le trafic, a appliqué des

structures purement convolution elles pour extraire simultanément des caractéristiques spatio-temporelles à partir de données de séries chronologiques structurées graphiquement.

De plus, la convolution causale dilatée est un type spécial de convolution unidimensionnelle standard .Il ajuste la taille du champ réceptif en modifiant la valeur du taux de dilatation, ce qui est propice à capturer la dépendance périodique à long terme. Donc adopté la convolution causale dilatée comme couche de convolution temporelle de leurs modèles pour capturer les tendances temporelles d'un nœud. Par rapport aux modèles récurrents, les convolutions créent des représentations pour des contextes de taille fixe, cependant, la taille de contexte effective du réseau peut facilement être agrandie en empilant plusieurs couches les unes sur les autres.

Cela permet de contrôler précisément la longueur maximale des dépendances à modéliser. Le réseau convolutif ne repose pas sur le calcul du pas de temps précédent, il permet donc la parallélisassions de chaque élément de la séquence, ce qui peut faire un meilleur usage du matériel GPU et plus facile à optimiser. Ceci est supérieur aux RNN, qui conservent tout l'état caché du passé, empêchant les calculs parallèles dans une séquence. [5]

4.2.2.2 RNN

Dans l'apprentissage de séquence basé sur RNN (Recurrent Neural Network) une structure de réseau spéciale connue sous le nom de codeur-décodeur a été appliquée pour la prévision du trafic. L'idée clé est d'encoder la séquence source comme un vecteur de longueur fixe et utilisez le décodeur pour générer la prédiction. [5]

4.2.2.3 Mécanisme attention

A conçu un mécanisme d'attention temporelle pour modéliser de manière adaptative les corrélations non linéaires entre les différents pas de temps. Ont incorporé un mécanisme standard de convolution et d'attention pour mettre à jour les informations d'un nœud en fusionnant les informations aux pas de temps voisins, et exprimer sémantiquement l'intensité de dépendance entre les différents pas de temps. Considérant que les données de trafic sont très périodiques, mais pas strictement périodiques, a conçu un mécanisme d'attention périodiquement décalé pour faire face à la dépendance périodique à long terme et au décalage temporel périodique.[5]

4.2.2.4 GCN

Song et coll. a d'abord construit un graphe spatio-temporel localisé qui comprend à la fois des attributs temporels et spatiaux, puis a utilisé la méthode GCN spatiale proposée pour modéliser simultanément les corrélations spatio-temporelles. [5]

5. Modèle de Machine Learning utilisé pour la classification

La classification aussi appelé apprentissage supervisé est l'activité consistant à examiner les caractéristiques d'un objet présenté et l'attribuer à l'un d'un ensemble prédéfini de Des classes. [9]

5.1 Apprentissage non supervisé (Clustering)

L'apprentissage non supervisé consiste à apprendre à classer sans supervision. Au début de processus nous ne disposons ni de la définition des classes, ni de leurs nombres. C'est l'algorithme de classification qui va déterminer ces informations. Nous ne disposons pas non plus de données en entrée qui sont déjà classées, c'est aussi à l'algorithme de découvrir par lui-même la structure plus ou moins cachée des données et de former des groupes d'individus dont les caractéristiques sont communes. [10]

5.2 Apprentissage supervisé (Catégorisation)

Contrairement à l'apprentissage non supervisé, nous commençons ici par un ensemble de classes connues et définies à l'avance. Nous disposons aussi d'une sélection initiale de données dont la classification est connue. Ces données sont supposées indépendantes et identiquement distribuées. Elles nous servent pour l'apprentissage de l'algorithme. La classification se fait par l'algorithme selon le modèle qu'il a appris. [10]

5.3 Algorithmes d'apprentissage

5.3.1 Naïve Bayes

Naïve Bayes est un algorithme d'apprentissage automatique supervisé basé sur le théorème de Bayes qui est utilisé pour résoudre des problèmes de classification en suivant une approche probabiliste. Il est basé sur l'idée que les variables prédictives d'un modèle d'apprentissage automatique sont indépendantes les unes des autres. Cela signifie que le résultat d'un modèle dépend d'un ensemble de variables indépendantes qui n'ont rien à voir les unes avec les autres. [11]

Dans les problèmes du monde réel, les variables prédictives ne sont pas toujours indépendantes les unes des autres, il existe toujours des corrélations entre elles. Étant donné que Naïve Bayes considère que chaque variable prédictive est indépendante de toute autre variable du modèle, elle est appelée «Naïve».Voyons maintenant la logique de l'algorithme Naïve Bayes.

Le principe derrière Naïve Bayes est le théorème de Bayes également connu sous le nom de règle de Bayes. Le théorème de Bayes est utilisé pour calculer la probabilité conditionnelle, qui n'est rien d'autre que la probabilité qu'un événement se produise sur la base d'informations sur les événements du passé. Mathématiquement, le théorème de Bayes est représenté par: [11]

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \quad (2)$$

Dans l'équation ci-dessus:

P (A | B): Probabilité conditionnelle que l'événement A se produise, étant donné l'événement B.

P (A): Probabilité que l'événement A se produise.

P (B): Probabilité que l'événement B se produise.

P (B | A): Probabilité conditionnelle que l'événement B se produise, étant donné l'événement A. [11]

5.3.2Les arbres de décision

Les arbres de décision (AD) sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille. [12]

5.3.3 L'algorithmes de K-NN

L'algorithme K-NN (K-nearest neighbors) est une méthode d'apprentissage supervisé. Il peut être utilisé aussi bien pour la régression que pour la classification. Son fonctionnement peut être assimilé à l'analogie suivante "dis moi qui sont tes voisins, je te dirais qui tu es…". [13]

Pour effectuer une prédiction, l'algorithme K-NN ne va pas calculer un modèle prédictif à partir d'un Training Set comme c'est le cas pour la régression logistique ou la régression linéaire. En effet, K-NN n'a pas besoin de construire un modèle prédictif. Ainsi, pour K-NN il n'existe pas de phase d'apprentissage proprement dite.

C'est pour cela qu'on le catégorise parfois dans le" Lazy Learning" Pour pouvoir effectuer une prédiction, K-NN se base sur le jeu de données pour produire un résultat, pour effectuer une prédiction, l'algorithme K-NN va se baser sur le jeu de données en entier**.**

En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation. Ensuite pour ces K voisins, l'algorithme se basera sur leurs variables de sortie (output variable) Y pour calculer la valeur de la variable Y de l'observation qu'on souhaite prédire. Par ailleurs : [13]

- Si K-NN est utilisé pour la régression, c'est la moyenne (ou la médiane) des variables Y des K plus proches observations qui servira pour la prédiction.
- Si K-NN est utilisé pour la classification, c'est le mode des variables Y des Kplus proches observations qui servira pour la prédiction.

6. Directions futures

Nous présentons plusieurs orientations futures pour les prévisions de trafic.

6.1. Prévision du trafic dans les cas extrêmes

Bien que les tendances du trafic dans des conditions normales soient faciles à prévoir, une question plus intéressante dans les prévisions de trafic consiste à prévoir le trafic dans des conditions extrêmes, qui comprennent à la fois les heures de pointe et les prévisions de trafic post-accidentelles. Dans, les auteurs proposent d'apprendre une représentation des

caractéristiques de l'accident avec l'auto-encodeur, puis de le combiner avec un réseau neuronal récurrent pour la prévision du trafic post-accidentel. Bien que des performances améliorées soient observées, le modèle proposé ne tient pas compte de la corrélation entre les différents capteurs et les résultats peuvent encore être améliorés. [14]

6.2. Prévision du trafic de fusible

Avec d'autres applications De nombreuses applications importantes dans le domaine des transports sont étroitement liées à la prévision du trafic. Un exemple est l'estimation du temps de trajet (ETA). Actuellement, la prévision du trafic et l'estimation du temps de trajet sont généralement effectuées indépendamment. Il est souhaitable de disposer d'un modèle qui modélise conjointement ces deux problèmes et aboutisse à de meilleurs résultats pour l'une ou l'autre tâche. [14]

6.3. Modélisation de la dépendance temporelle à long terme

Une dépendance temporelle à très long terme existe généralement dans les données de trafic, par exemple, la situation actuelle du trafic peut être fortement corrélée avec un jour, une semaine ou même plusieurs mois. Actuellement, les approches les plus populaires pour modéliser la dépendance temporelle non linéaire sont les réseaux de neurones récurrents (RNN). Cependant, en raison de la nature séquentielle du (RNN), il est difficile de modéliser les dépendances à très long terme. De plus, (RNN) n'est pas efficace pour s'entraîner car il est difficile de le paralléliser. Ainsi, des approches efficaces capables de capturer les dépendances temporelles non linéaires à long terme sont indispensables. [14]

6.4. Conception de la métrique d'évaluation

Les mesures populaires pour évaluer les prévisions de trafic comprennent l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE) et l'erreur en pourcentage absolu moyen (MAPE) qui sont calculées en faisant la moyenne de tous les capteurs.

Ces métriques accordent une importance égale à tous les capteurs et créneaux horaires, cependant, nous soutenons que tous les capteurs et intervalles de temps ne sont pas également informatifs. Évaluer la performance.

La figure II.6 montre l'écart type de chaque emplacement par rapport à sa moyenne historique. En général, plus l'écart type est grand, plus il est difficile de prévoir le trafic à cet endroit. On peut soutenir que les emplacements et les heures avec une erreur plus élevée, par

exemple, les intersections achalandées pendant les heures de pointe, sont plus importants à prédire. Alternativement, prédire la vitesse moyenne sur toutes les autoroutes de minuit à 5 heures du matin n'est pas très difficile. Ainsi, il pourrait être avantageux d'avoir une métrique qui donne plus de récompenses à la prévision du trafic dans des endroits et des moments plus difficiles. [14]

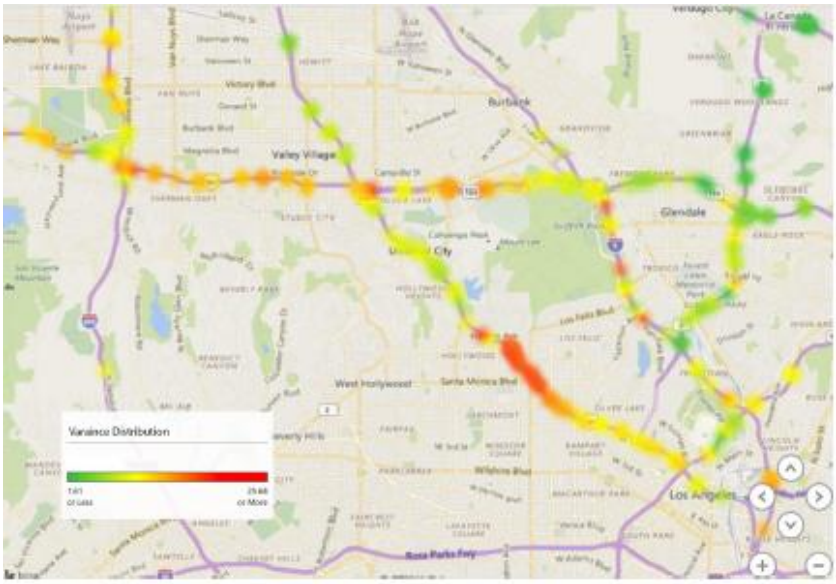
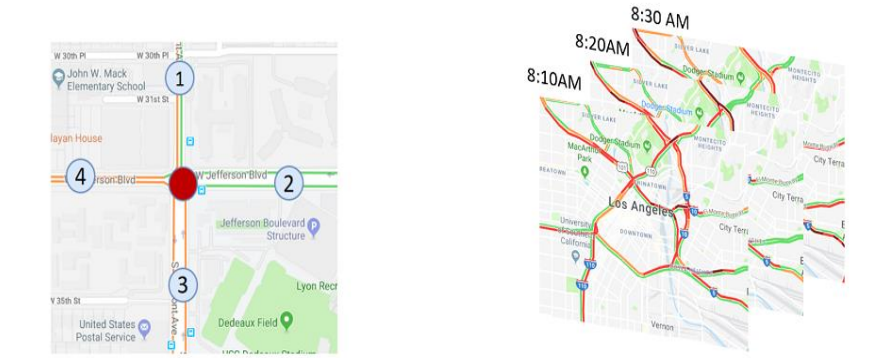


Figure II.6: Distribution de l'écart type par rapport a la moyenne historique.la dureté de la prévision de la circulation varie d'un endroit à l'autre. [14]

6.5. Prédiction de trafic interprétable:

De nombreux modèles d'apprentissage automatique sont utilisés pour la prévision du trafic. Bien que de bonnes performances soient obtenues, les prévisions faites par le modèle ne sont généralement pas interprétables. Comme le montre la figure II.7, il est souhaitable d'identifier les composantes spatiales et temporelles qui affectent la prédiction du modèle. En outre, plutôt qu'une seule prédiction, il est plus informatif de prédire une distribution, par exemple, la moyenne et la variance de la distribution gaussienne, ce qui aiderait la prise de décision ainsi que d'autres applications connexes, par exemple, l'estimation du temps de trajet.[14]



(a) Which neighbor road has the largest effects? (b) Which historical observation is more relevant?

Figure II.7: Prévisions de trafic interprétables**.**

7. Conclusion

Dans ce chapitre, nous avons mené une étude exhaustive des différentes approches et techniques utilisées dans la prévision du trafic routier. Plus précisément, nous résumons d'abord les méthodes de prévision du trafic existantes et en donnons une taxonomie. Enfin, certains défis majeurs et orientations futures de la recherche sont discutés. Il convient aux lecteurs intéressés pour comprendre rapidement les prévisions de trafic, afin de trouver les branches qui les intéressent. Ce chapitre fournit également une bonne référence et une enquête pour les chercheurs dans ce domaine, ce qui peut faciliter la recherche pertinente.

CHAPITRE 3

MODELES PROPOSES POUR LA PREDICTION DU TRAFIC ROUTIER

1. Introduction

Au cours de ce chapitre, nous allons présenter notre travail qui consiste à réaliser la prévision de trafic routier dans New York Metropolitan Transportation. Celui-ci est basé sur la prédiction par régression et la prédiction par la classification bayésienne naïve.

Nous commençons par présenter les différents outils utilisés : logiciel de Excel et logiciel de R et Rstudio.

Excel est un logiciel de la suite bureautique Office de Microsoft et permet la création de tableaux, de calculs automatisés, de plannings, de graphiques et de bases de données. On appelle ce genre de logiciel un tableur. Permet également de générer de jolis graphiques pour mieux visualiser les valeurs et les interpréter. C'est un puissant outil de visualisation mathématique.[16]

Le principe d'analyse de donnée sera implémenté et intégrés dans la plateforme de R.

R est un logiciel permettant de faire des analyses statistiques et de produire des graphiques. Mais R est également un langage de programmation complet, c'est cet aspect qui fait que R est différent des autres logiciels statistiques et un clône gratuit du logiciel S-Plus commercialisé par MathSoft et développé par Statistical Sciences autour du langage S. [17]

R studio est un outil apparu récemment et qui vient combler un manque dans la collection des outils associés à R : il s'agit d'un environnement de développement intégré (IDE en anglais) fonctionnel, libre, gratuit et multiplateforme. a été écrit en langage C++, et son interface graphique utilise l'interface de programmation Qt. est développé par RSTUDIO.Inc, une entreprise commerciale fondée par JOSEPH J.ALLAIRE.[1]

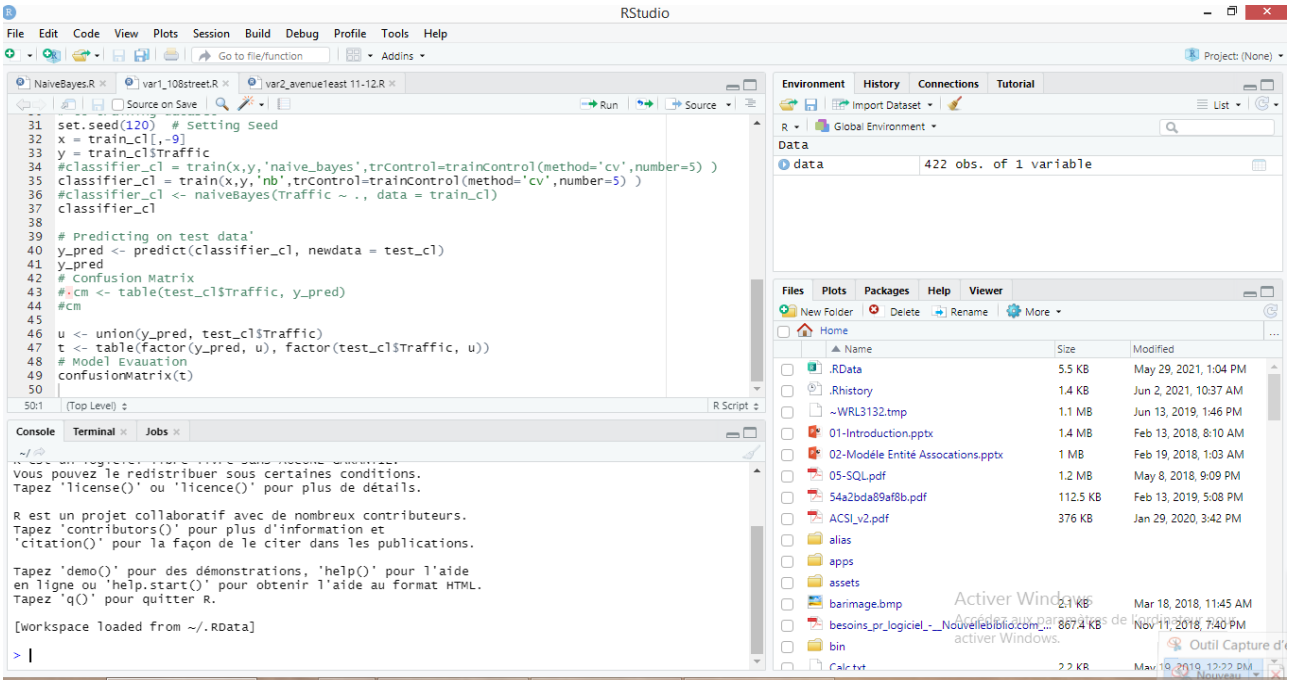


Figure III.1: Interface de RStudio**.**

2. Les données

Les données sont collectées parDOT pour New York Metropolitan Transportation Council (NYMTC) et validées par the New York Best Practice Model (NYBPM).

Source: https://data.cityofnewyork.us/d/ertz-hr4r.

La taille :4 ,17 MB

Nombre de lignes :27290 .

Colonnes de table :id,segment_id,roadway_name,from,to,direction,date et les heures:

12:00_1:00_am, 1:00_2:00am, 2:00_3:00am, 3:00_4:00am, 4:00_5:00am,5:00_6:00am
6:00_7:00am, 7:00_8:00am, 8:00_9:00am, 9:00_10:00am, 10:00_11:00am, 11:00_12:00pm
12:00_1:00pm, 1:00_2:00pm, 2:00_3:00pm, 3:00_4:00pm, 4:00_5:00pm,5:00_6:00pm
6:00_7:00pm, 7:00_8:00pm, 8:00_9:00pm,9:00_10:00pm,10:00_11:00pm, 11:00_12:00am

ID	Segment I	Roadway	From	To	Direction	Date	12:00-1:00 A	1:00-2:00A	2:00-3:00A	3:00-4:00A	4:00-5:00A	5:00-6:00A	6:00-7:00A	7:00-8:00A	8:00-9:00A	9:00-10:00	10:00-11:0	11:00-12:0
2	70376	3 Avenue	East 154 S	East 155 S	NB	09/13/2014	204	177	133	126	141	134	121	180	223	272	386	339
2	70376	3 Avenue	East 155 S	East 154 S	SB	09/13/2014	140	51	128	116	144	146	153	219	226	273	317	325
56	176365	Bedford P	Grand Cor	Valentine	EB	09/13/2014	94	73	65	61	64	73	65	113	169	210	182	245
56	176365	Bedford P	Grand Cor	Valentine	WB	09/13/2014	88	82	75	60	65	67	71	142	198	212	205	237
62	147673	Broadway	West 242	240 Street	SB	09/13/2014	255	209	149	148	128	136	199	354	473	567	634	781
62	158447	Broadway	West 242	240 Street	NB	09/13/2014	255	209	149	148	128	136	199	354	473	567	634	781
62	255653	Broadway	West 242	240 Street	SB	09/13/2014	87	86	78	56	47	80	98	133	171	177	215	235
71	139620	Bronx Rive	Bronx Rive	East Gun H	NB	09/13/2014	802	445	388	318	400	527	969	1443	1712	1923	2248	2481
71	139618	Bronx Rive	East Gun H	Bronx Rive	SB	09/13/2014	618	345	309	277	301	365	699	1194	1662	1830	2058	2493
76	70364	Brook Ave	East 152 S	East 153 S	SB	09/13/2014	79	58	41	30	42	34	55	92	114	150	194	200
89	69780	East 138 S	Canal Plac	Rider Ave	EB	09/13/2014	339	233	239	206	257	226	231	319	360	450	386	389
89	69780	East 138 S	Rider Ave	Canal Plac	WB	09/13/2014	282	229	182	187	200	164	188	254	304	331	360	395
102	69979	Concourse	East 156 S	East 156 S	NB	09/13/2014	101	67	50	52	34	30	48	85	98	159	224	203
104	9001519	Courtladt	East 153 S	East 154 S	NB	09/13/2014	97	72	36	46	52	29	35	58	100	122	155	152
108	276970	East 144 S	A J Griffin	Grand Cor	WB	09/13/2014	77	42	29	28	29	32	50	98	142	101	125	123
108	276970	East 144 S	Grand Cor	A J Griffin	EB	09/13/2014	81	49	55	40	46	46	93	172	168	150	166	160
109	173496	East 149 S	A J Griffin	Grand Cor	WB	09/13/2014	348	303	280	217	214	170	236	308	397	448	531	519
109	173496	East 149 S	Grand Cor	A J Griffin	EB	09/13/2014	454	365	344	277	252	203	300	342	476	540	608	616
111	69836	East 149 S	Morris Av	Park Aven	WB	09/13/2014	338	304	271	188	185	176	261	342	425	507	611	646
111	69836	East 149 S	Park Aven	Morris Av	EB	09/13/2014	333	238	209	158	172	150	227	251	376	407	490	498
120	111919	Eastchest	East Gun H	Adee Ave	NB	09/13/2014	109	71	49	28	30	34	65	178	200	247	273	357
120	111919	Eastchest	East Gun H	Adee Ave	SB	09/13/2014	57	35	32	37	36	34	113	155	167	195	197	272
121	88518	Eastchest	Givan Ave	Tillotson A	NB	09/13/2014	169	133	81	48	35	53	99	211	235	273	321	391
121	88518	Eastchest	Givan Ave	Tillotson A	SB	09/13/2014	116	83	60	58	47	49	110	135	187	224	226	277

Table III.1: Extrait de l'ensemble de données du trafic routier (2014-2019) sous forme Excel.

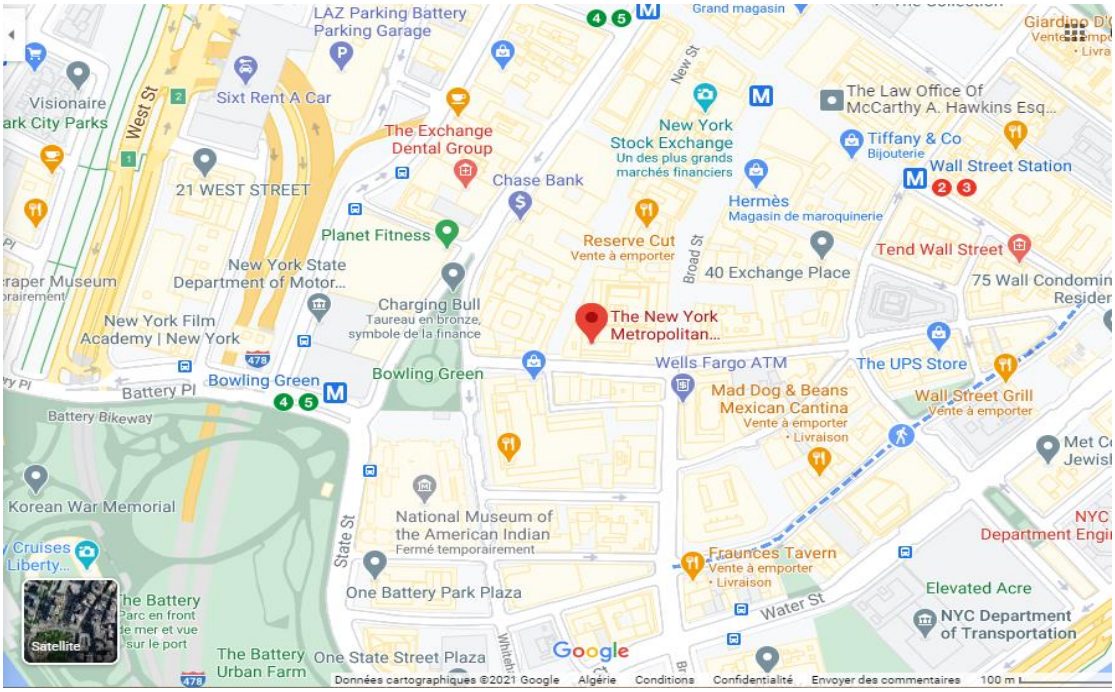


Figure III.2: New York Metropolitan Transportation Council.

Nous avons mené notre étude en utilisant quatre (04) variantes différentes :

Variante1 : Prédiction du trafic routier sur un segment donné (avenue) à un instant donné (heure fixe)

Variante 2 : Prédiction de la moyenne du trafic routier sur un segment donné durant la journée (les 24 heures)

Variante 3 : Prédiction du trafic routier sur tous les segments à une heure fixe

Variante 4 : Prédiction de la moyenne du trafic routier sur tous les segments durant la journée (les 24 heures).

3. Les données utilisées

Nous avons sélectionné 4 avenues parmi les données en possession :

3.1 Avenue 1

Nous avons extrait l'avenue 1de Google map comme indiqué dans la figure 3

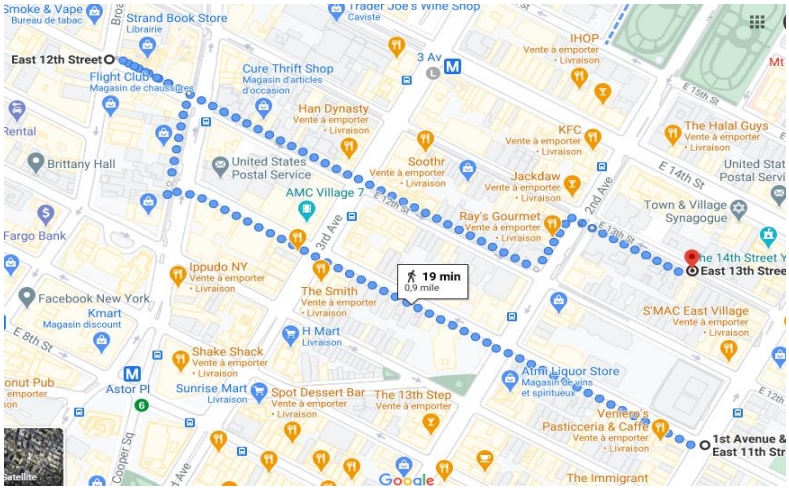


Figure III.3:avenue 1(east11street to east12 street to east13 street).

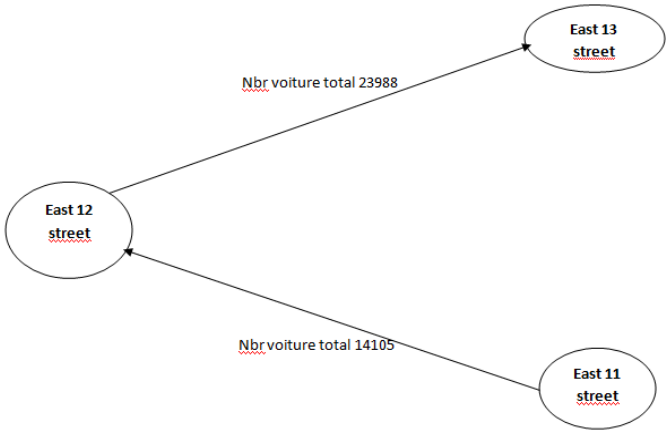


Figure III.4: Graphed'avenue 1**.**

3.2 (65 street 10 avenue-11 avenue)

Nous avons extrait l'avenue 65 street de Google map comme indiqué dans la figure 5

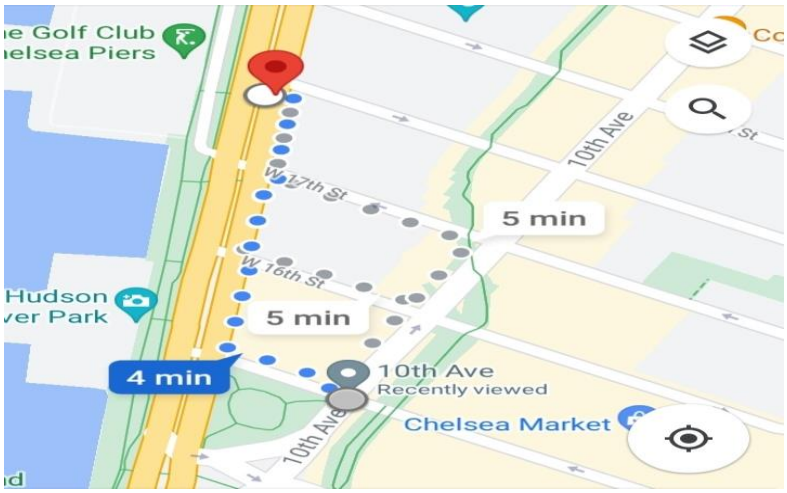


Figure III.5: 65 street (10 avenue-11 avenue)**.**

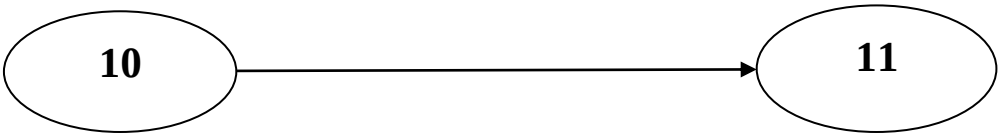


Figure III.6: Graphe de 65 street**.**

3.3 Centrale Park West (100street-97street)

Nous avons extrait l'avenue centrale park west de Google map comme indiqué dans la figure 7

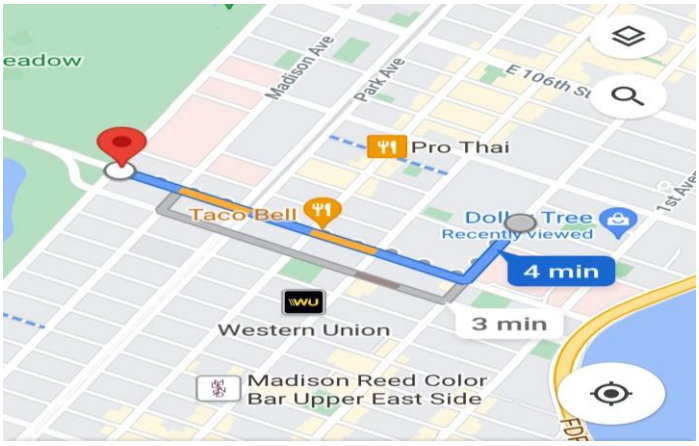


Figure III.7: Centrale Park West (100street-97street).

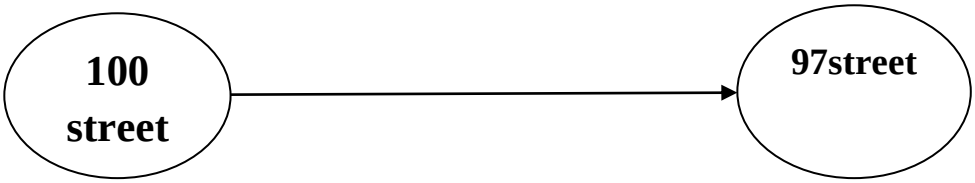


Figure III.8: Graphe de Centrale Park West.

3.4. (108street 62 avenue-apex place):

Nous avons extrait l'avenue 108 street de Google map comme indiqué dans la figure 9

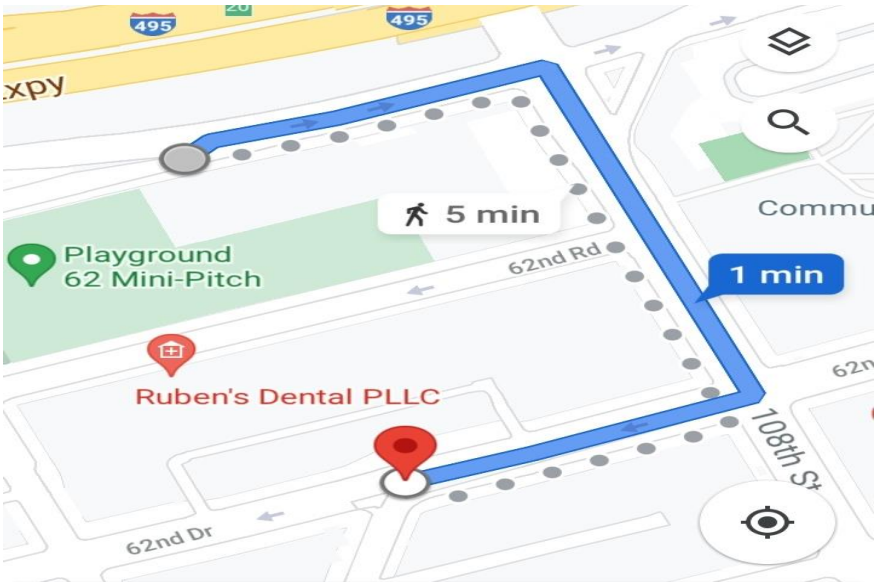


Figure III.9: 108street(62 avenue-apex place)**.**

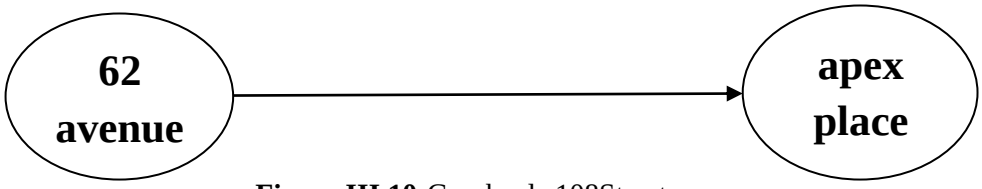


Figure III.10: Graphe de 108Streets**.**

4. Prédiction du trafic routier par régression

Nous avons mené dans cette partie une étude du trafic routier en utilisant les modèles de régression. La stratégie que nous avons suivie est la suivante : pour chaque variante, nous utilisons Microsoft Excel pour l'élaboration du modèle de régression, nous commençons par l'utilisation de la régression linéaire, si le taux d'erreur résiduelle (R^2) n'est pas bon (généralement $R^2<50\%$) alors nous utilisons le modèle de régression polynomiale qui généralement donne un meilleur taux d'erreur.

4.1 Variante 01 : segment fixe, heure fixe, date variable

On s'intéresse dans cette variante à la prédiction du trafic routier sur un segment donné à un instant (ou une heure) fixe.

On cherche à modéliser la relation entre la date et le nombre des voitures .On pose :

y = nombre des voitures.

x = la date.

Afin de pouvoir faire des prédictions, nous avons changé la date en valeurs, à l'aide de la fonction dateval dans Excel

On suppose que cette relation est linéaire de la forme :

$$y = ax+b \qquad (3)$$

4.1.1 avenue 1 (east11 street-east12street)

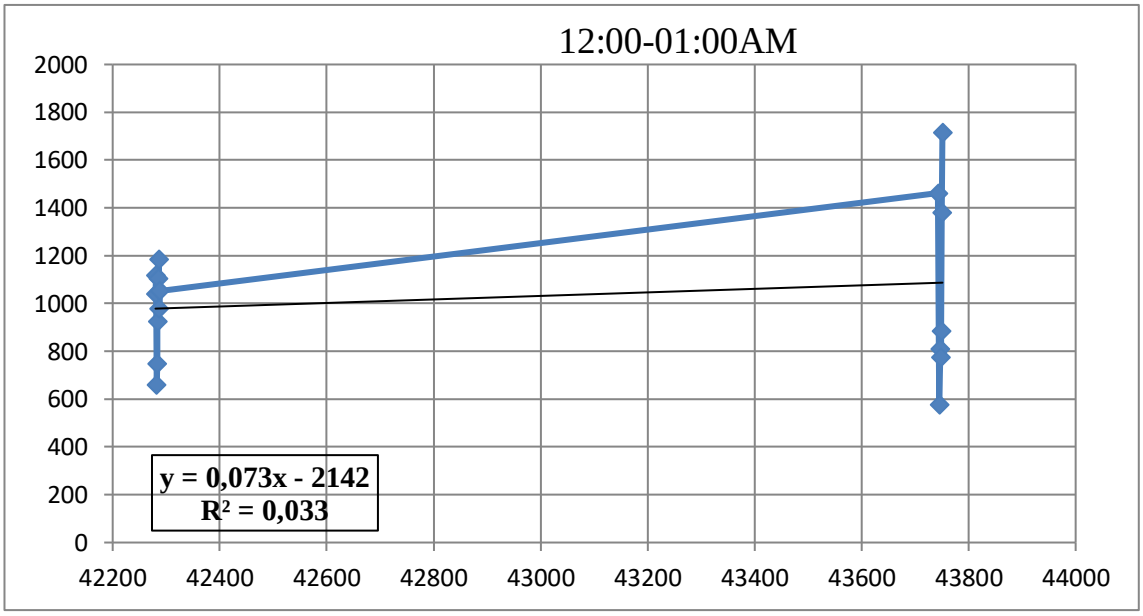


Figure III.11: La régression linéaire simple pour east11 street-east12street.

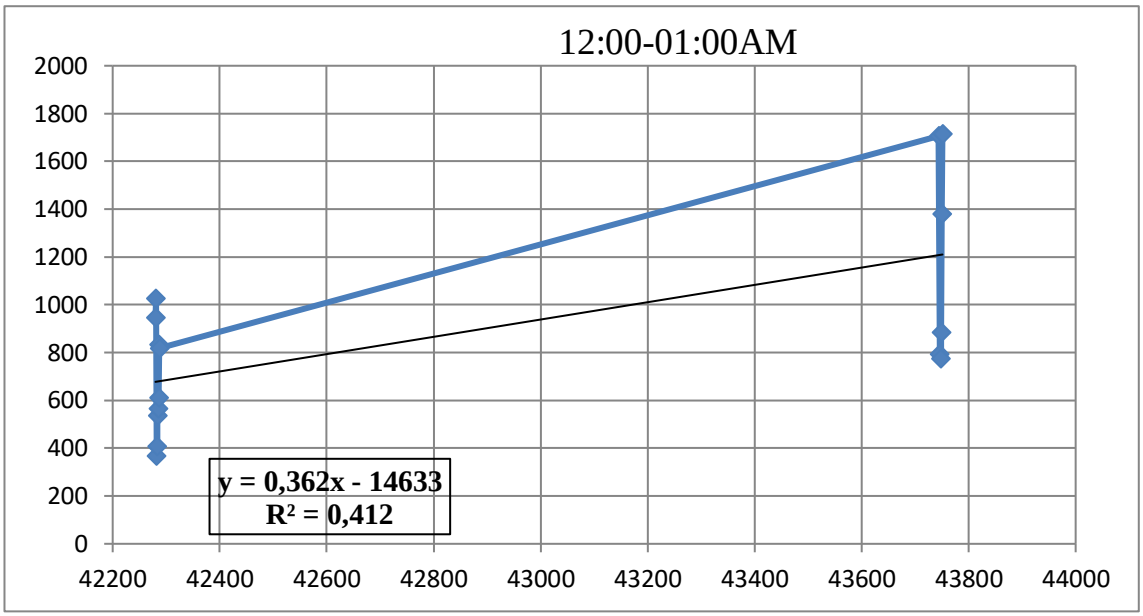


Figure III.12: La régression linéaire simple pour east12 street-east13 street.

On a :

$$\mathbf{y= 0,362x - 14633 + } \boldsymbol{R^2} \qquad (4)$$

$\hat{y} = \hat{a}x + \hat{b}$ est appelé la valeur prédite

Où

- ✓ X est la variable explicative
- ✓ Y est la variable à expliquer
- ✓ a=0,362 est le paramètre de pente de la régression
- ✓ b=14633est le paramètre d'ordonnée à l'origine de la régression
- ✓ R^2=0,412 est l'erreur résiduelle

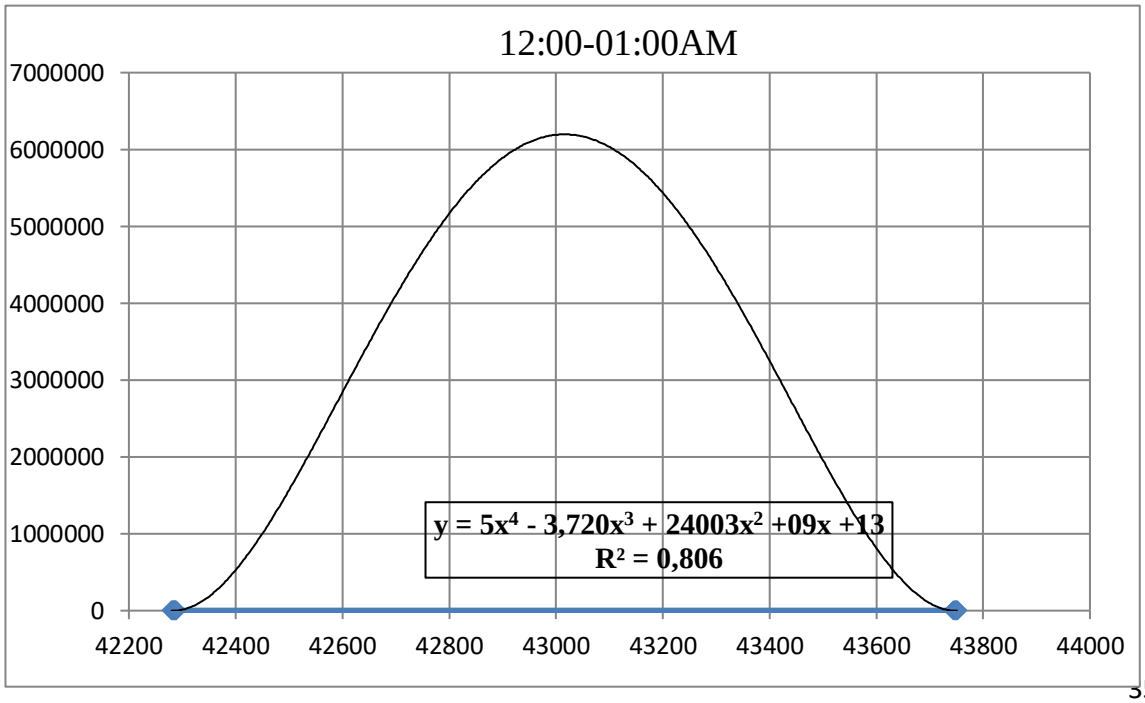


Figure III.13: La régression polynomiale pour east12 street-east13 street.

On a :

$$\mathbf{y=5x^4 - 3{,}720x^3 + 24003x^2 + 9x + 13 + R} \qquad (5)$$

$\hat{y} = \hat{a}x + \hat{b}$ est appelé la valeur prédite

Où:

- ✓ *X* est la variable explicative
- ✓ *Y* est la variable à expliquer
- ✓ a4=5, a3= 3,720, a2=24003 , a1=9 est les paramètres de pente de la régression
- ✓ b=13est le paramètre d'ordonnée à l'origine de la régression
- ✓ R^2 =0,821est l'erreur résiduelle$=y-\hat{y}$
- ✓ Pour la régression polynomiale R^2 = 0,806

Pour la régression linéaire le pourcentage est faible dans le segment east12 street-east13 street , dans east11 street-east12 street pour la régression linéaire est 41 % par contre dans la régression polynomiale est 82%très bon taux pour la prédiction.

4.1.2 (65 street 10 avenue-11 avenue)

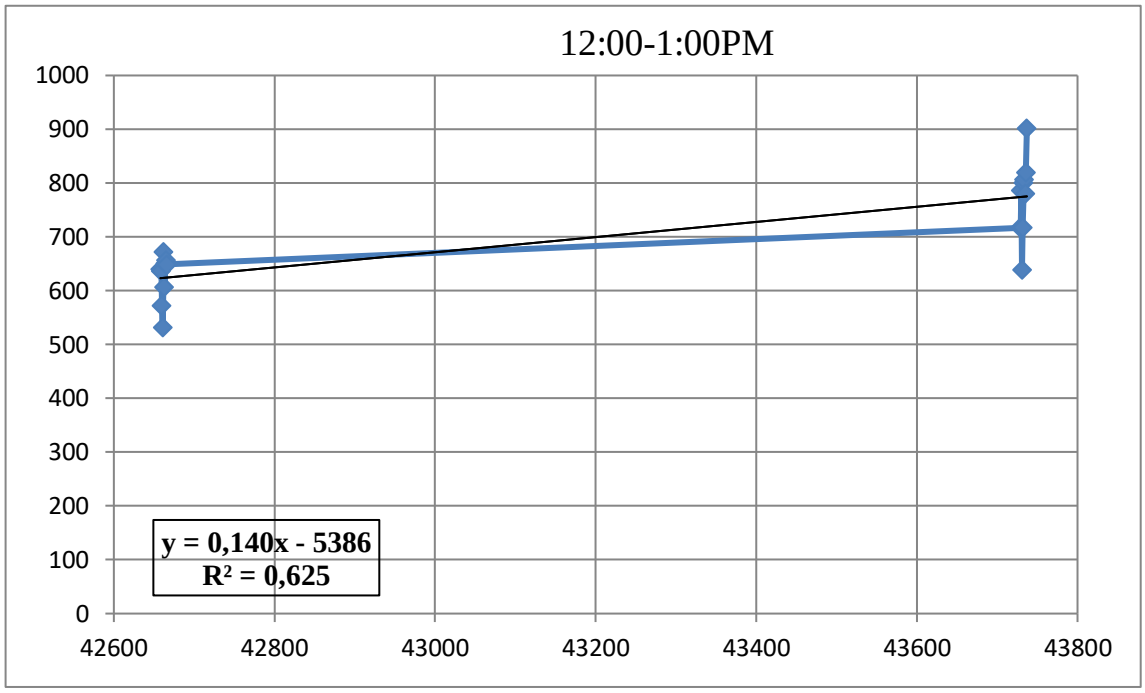


Figure III.14: La régression linéaire simple pour 65 street (10 avenue-11 avenue).

On a :

$$\mathbf{y= 0,140x - 5386+ } \boldsymbol{R^2} \qquad (6)$$

$\hat{y} = \hat{a}x + \hat{b}$ est appelé la valeur prédite

Où:

- ✓ X est la variable explicative
- ✓ Y est la variable à expliquer
- ✓ a=0,140est le paramètre de pente de la régression
- ✓ b=5386est le paramètre d'ordonnée à l'origine de la régression
- ✓ R^2= 0,625 est l'erreur résiduelle

A noter que le ratio R est proche de 62%, ce qui est un bon ratio par rapport au premier.

4.1.3 Central Park West (100street-97street)

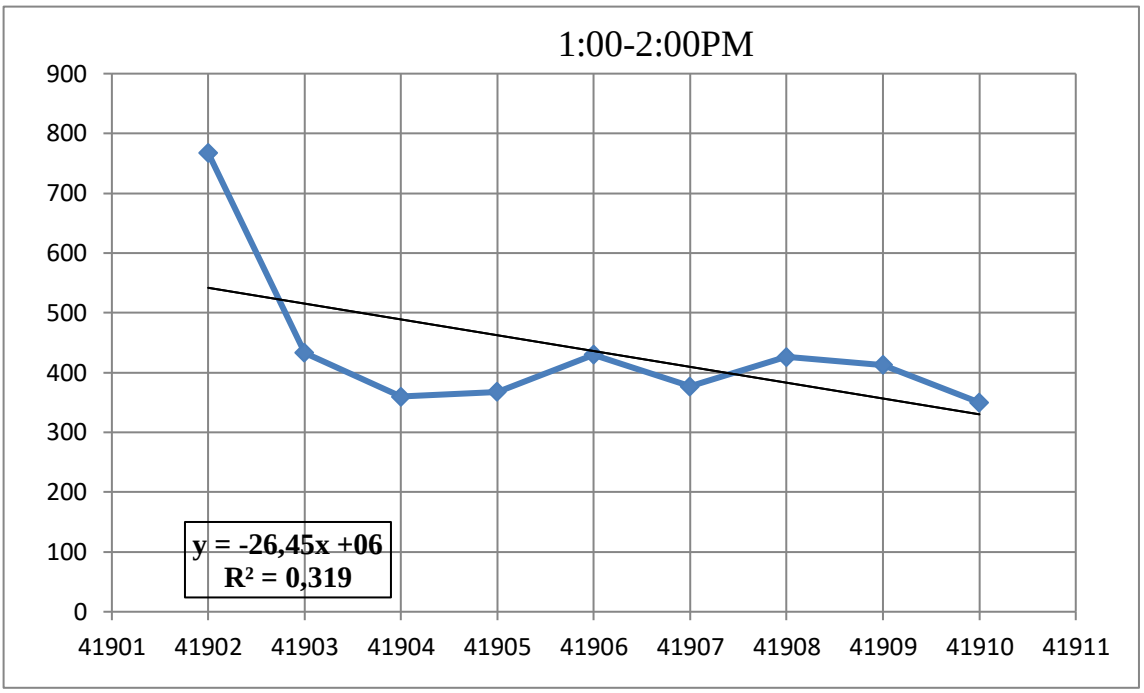


Figure III.15: La régression linéaire simple pour Central Park West (100street-97street)**.**

R^2=0,319 ; C'est à peu près 32% aussi reste Plutôt faible.

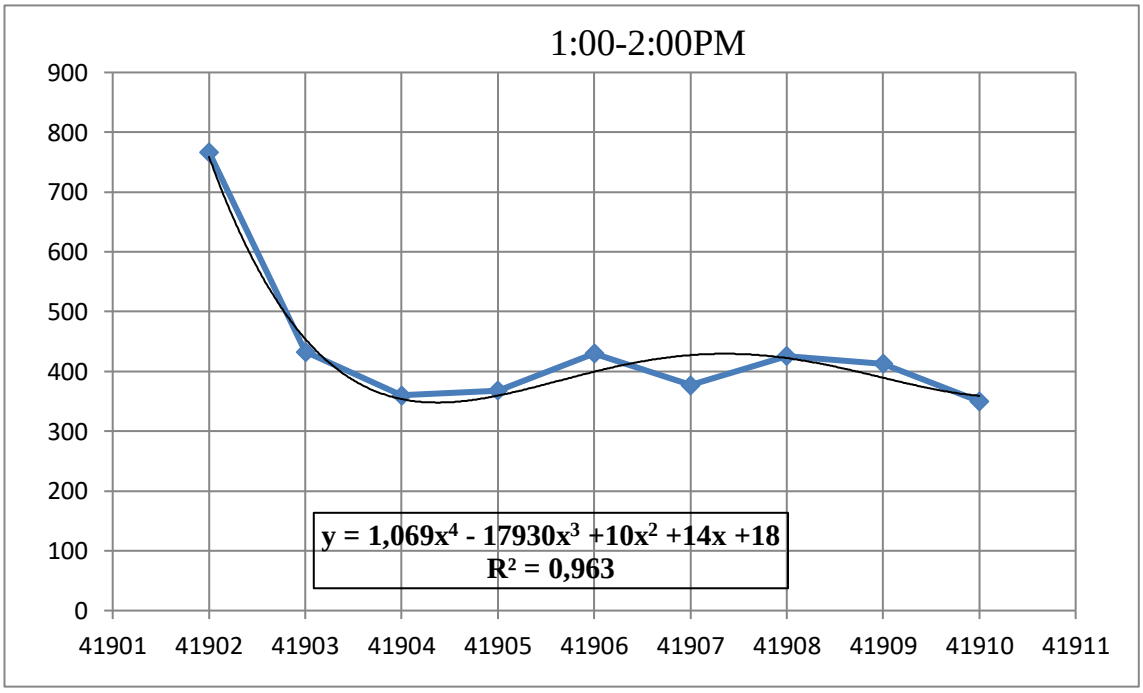


Figure III.16: La régression polynomiale pour Central Park West (100street-97street)**.**

Par contre Central Park West dans la valeur de R = 31% C'est pourcentage faible pour faire une prédiction, de plus, lors du changement de régression polynomiale, le rapport est passé à 96%.

4.1.4 108street (62 avenue-apex place)

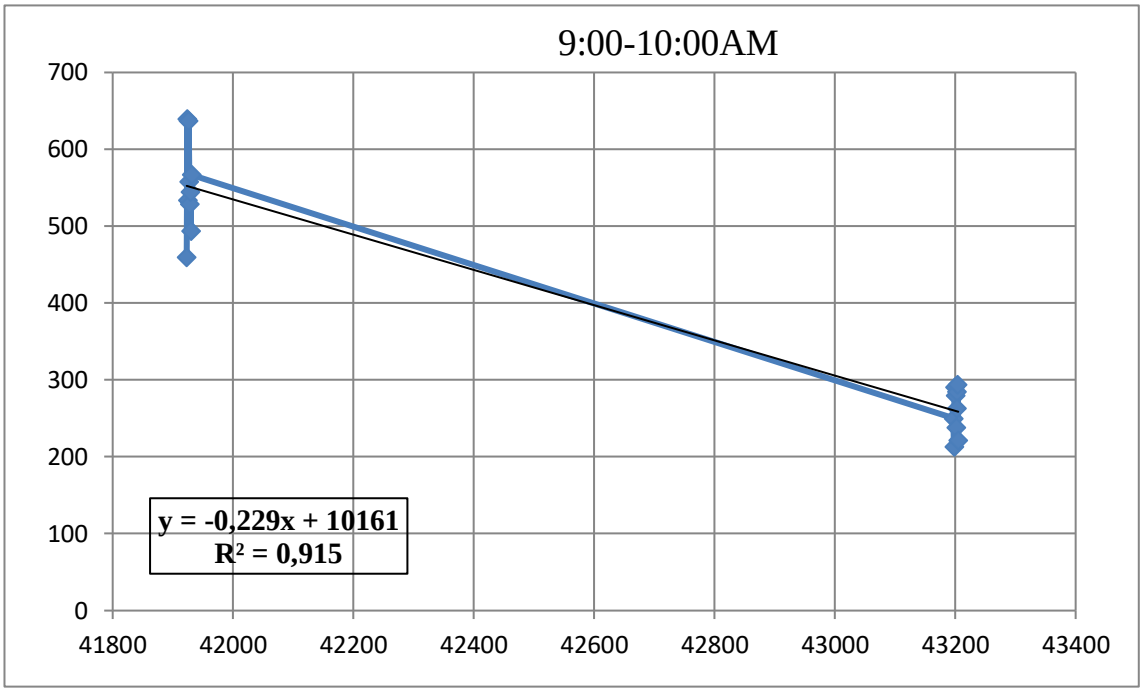


Figure III.17: La régression linéaire simple pour108street (62 avenue-apex place)**.**

4.2 Variante 02 : segment fixe

La prédiction de la moyenne du trafic pour toutes les heures. On pose :

y = la moyenne des voitures.

x = la date.

On suppose que cette relation est linéaire de la forme :

$$y = ax+b \qquad (7)$$

4.2.1 avenue 1(east 11 street vers east 12 street)

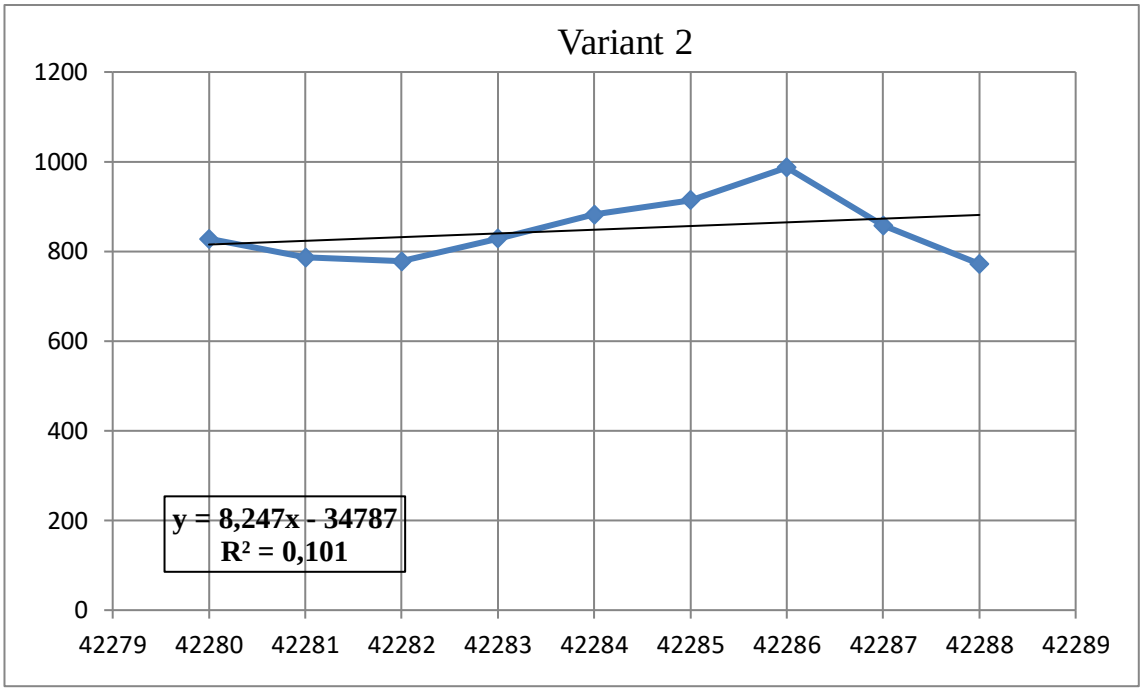


Figure III.18: La régression linéaire simple pour avenue 1(east 11 street vers east 12 street).

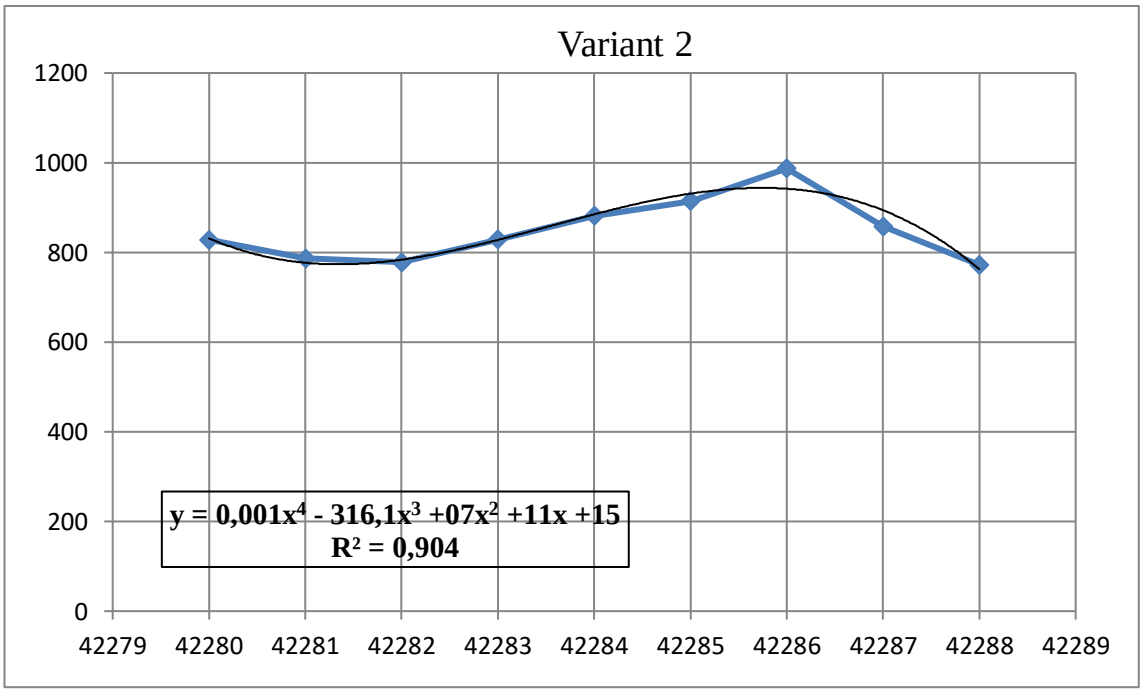


Figure III.19: La régression polynomiale pour avenue 1(east 11 street vers east 12 street).

R^2 =90% Dans le segment (east 11 street vers east 12 street) avec la régression polynomiale c'est un très bon pourcentage.

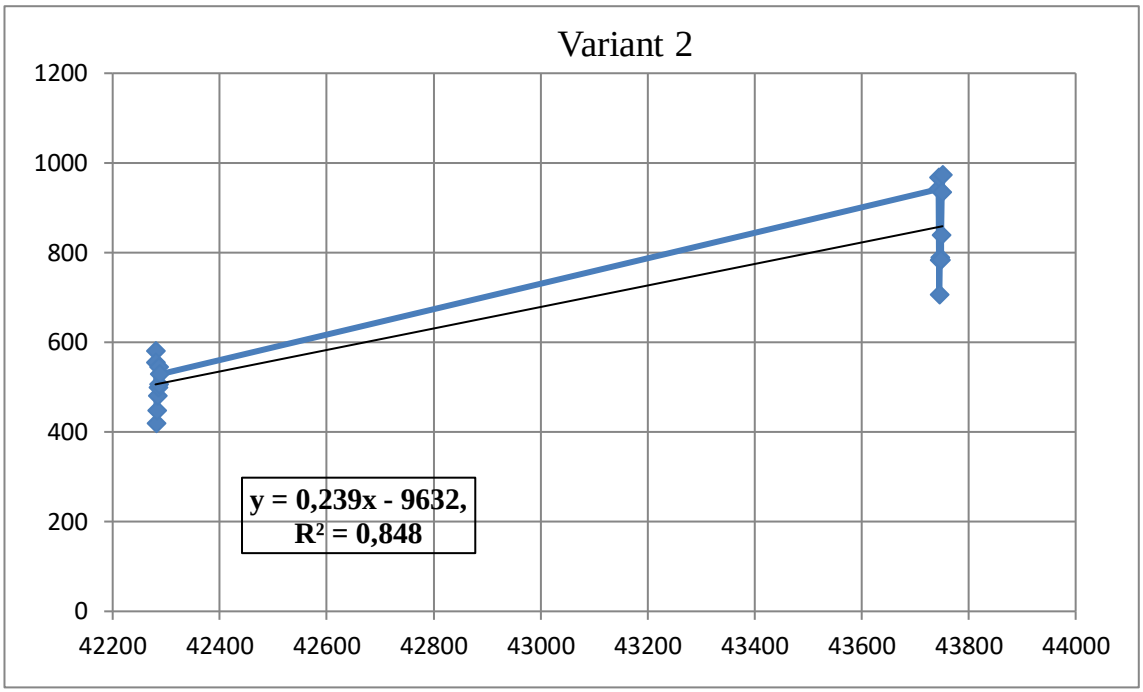


Figure III.20: La régression linéaire simple pour l'avenue 1(east 12 street vers east 13 street).

R^2= 85% Dans le segment (east 12 street vers east 13 street) avec la régression linéaire simple, c'est aussi un bon pourcentage.

4.2.2.(65 street 10 avenue-11 avenue)

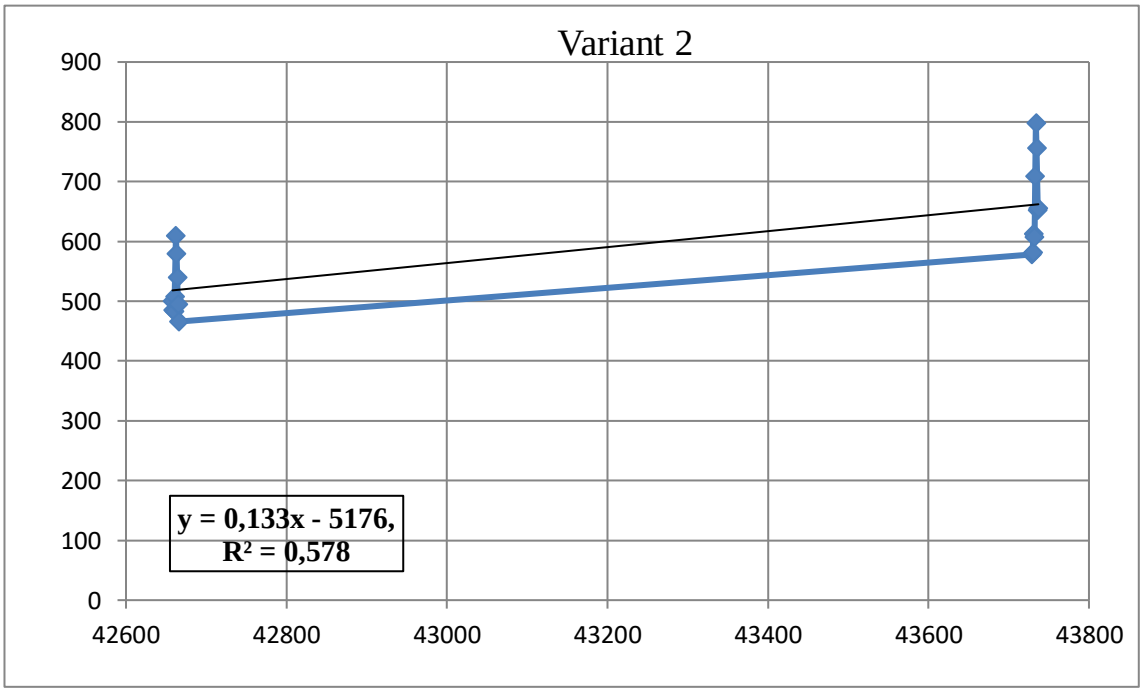


Figure III.21: La régression linéaire simple pour65 street (10 avenue-11 avenue).

Pour le segment 65 street l'erreur résiduelle pour la régression linéaire simple est 58%, c'est acceptable.

4.2.3 Central Park West (100street-97street)

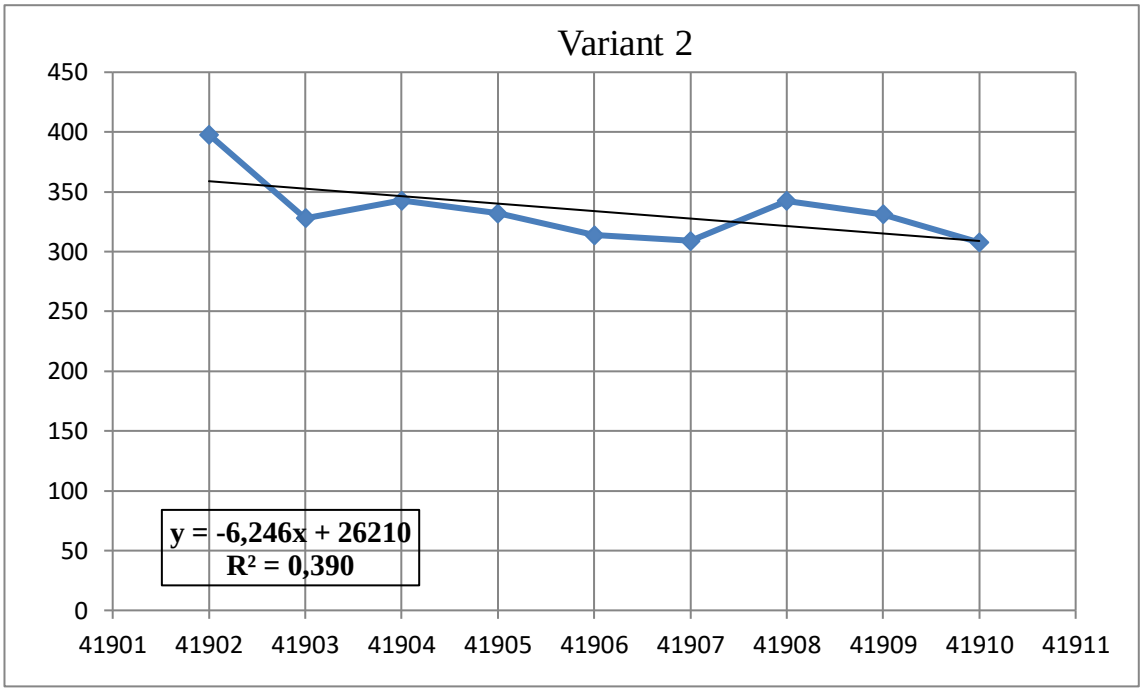


Figure III.22: La régression linéaire simple pour Centrale Park West (100street-97street).

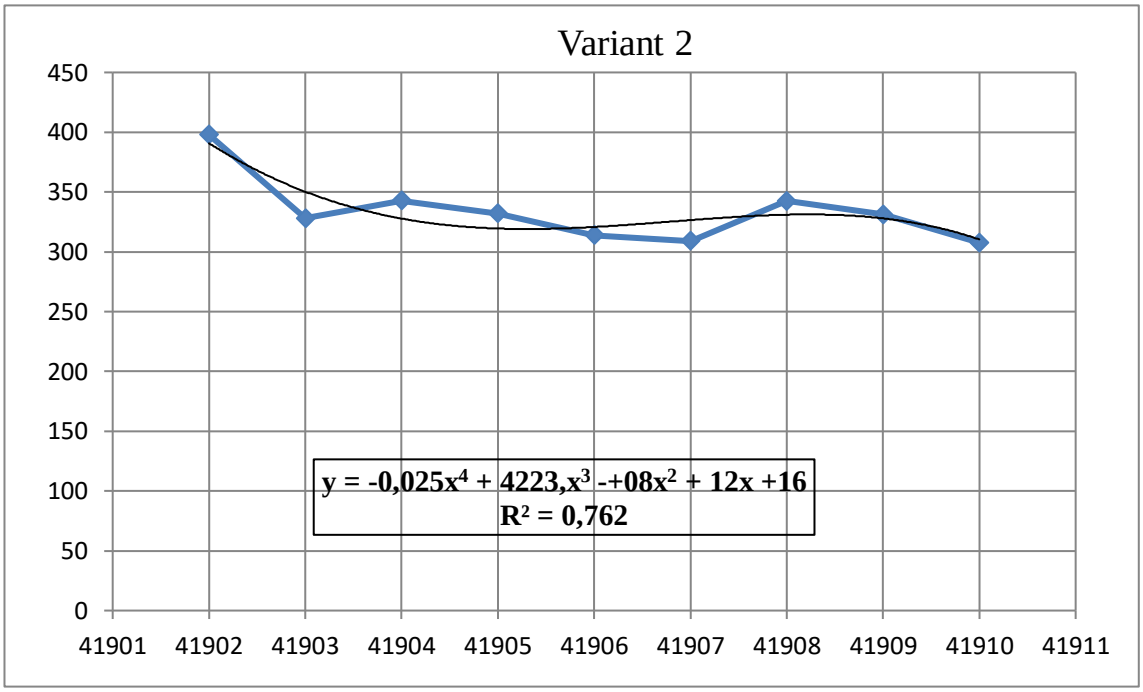


Figure III.23: La régression polynomiale pour Central Park West (100street-97street).

Le pourcentage de R^2 dans la régression linéaire un peu faible ,par contre dans la régression polynomiale est bon.

4.2.4 (108street 62 avenue-apex place)

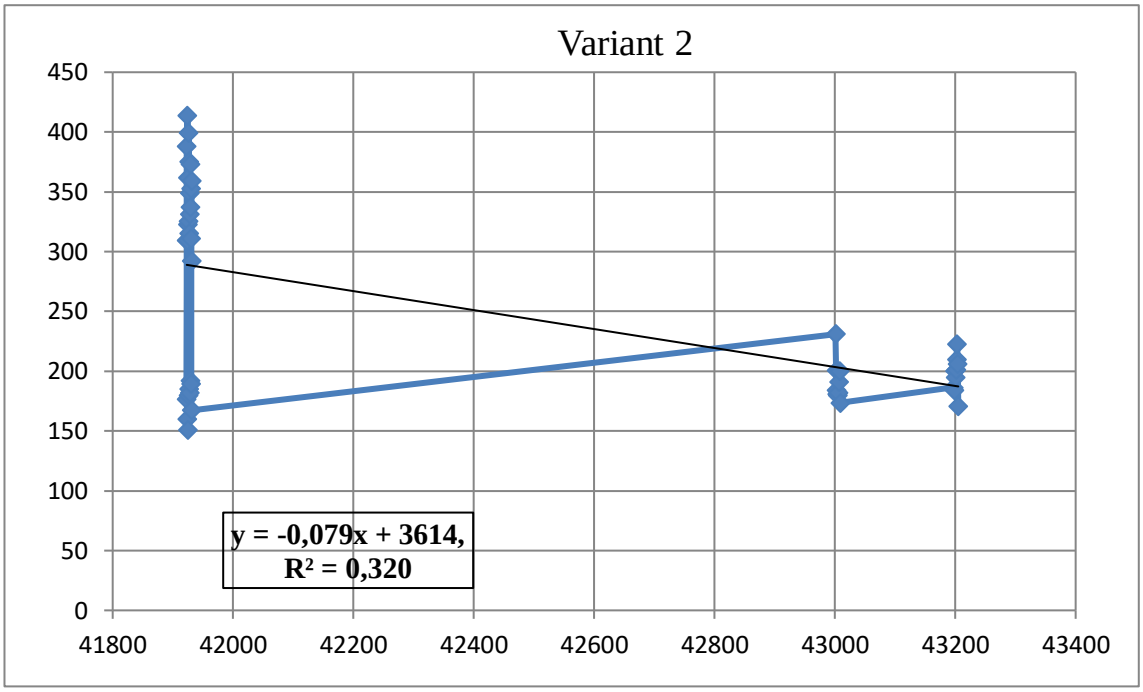


Figure III.24: La régression linéaire simple pour 108street (62 avenue-apex place)**.**

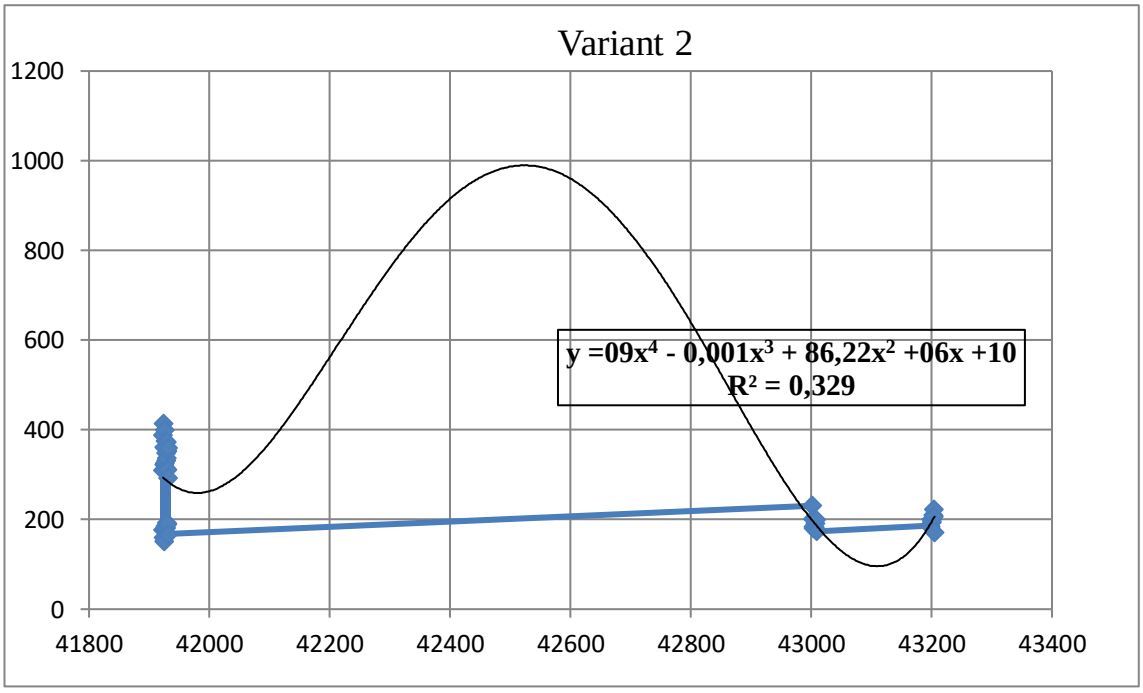


Figure III.25: La régression polynomiale pour 108street (62 avenue-apex place)**.**

Dans les deux cas (linéaire et polynomiale) le pourcentage reste faible.

4.3. Variante 03 : tous les segments, heure fixe, date variable

Prédiction de total count de tous les segments pour une date. On pose :

y = nombre totale des voitures.

x = la date.

On suppose que cette relation est linéaire de la forme :

$$y = ax+b \qquad (8)$$

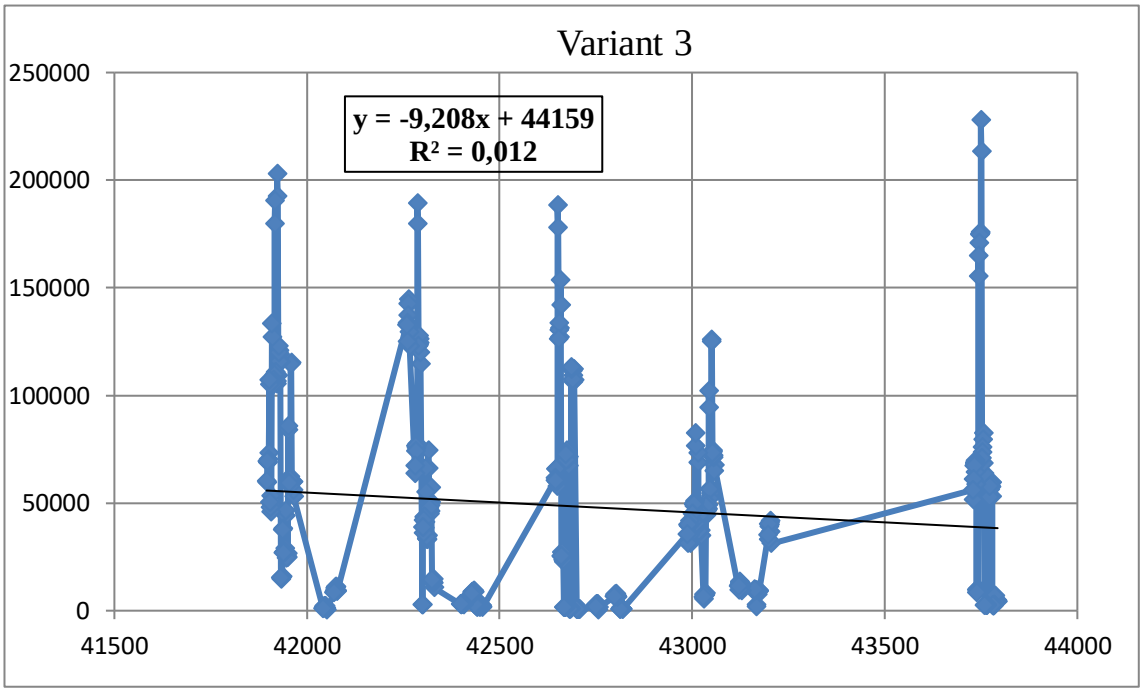


Figure III.26: La régression linéaire simple pour tous les segments avec le total de count.

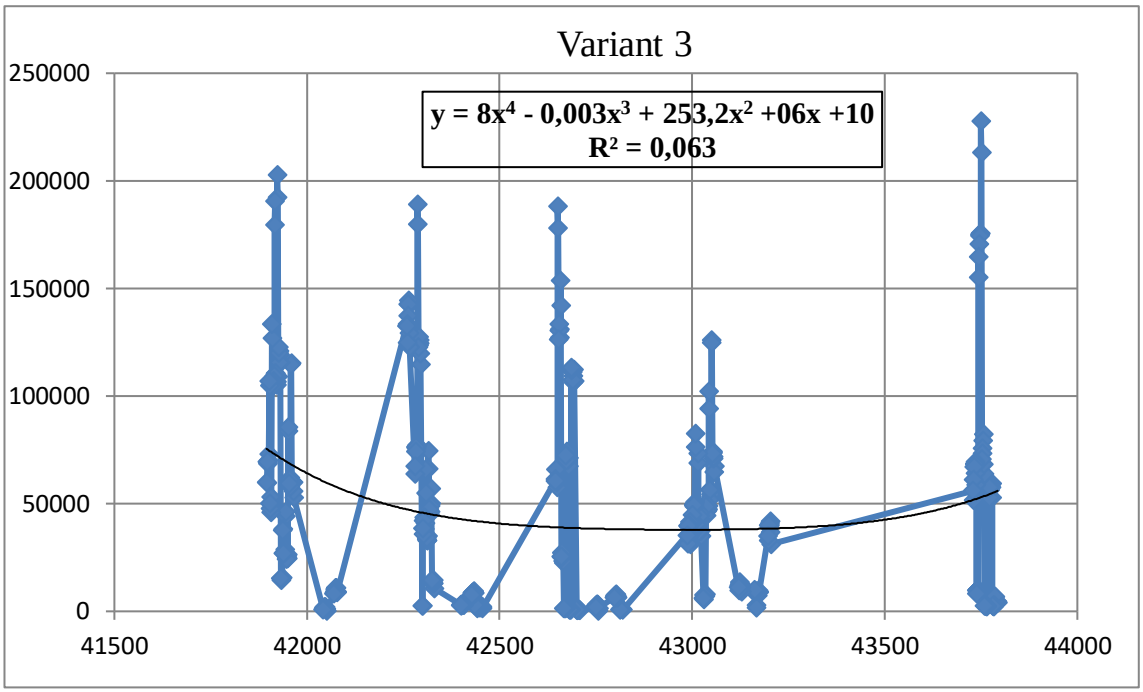


Figure III.27: La régression polynomiale pour tous les segments.

R^2= 0 ,012 dans la régression linéaire et 0,063 dans polynomiale, reste toujours faible.

4.4. Variante04 : tous les segments

Prédiction de moyenne count de tous les segments pour une date

On pose :

y = nombre moyenne des voitures.

x = la date.

On suppose que cette relation est linéaire de la forme :

$$y = ax+b \qquad (9)$$

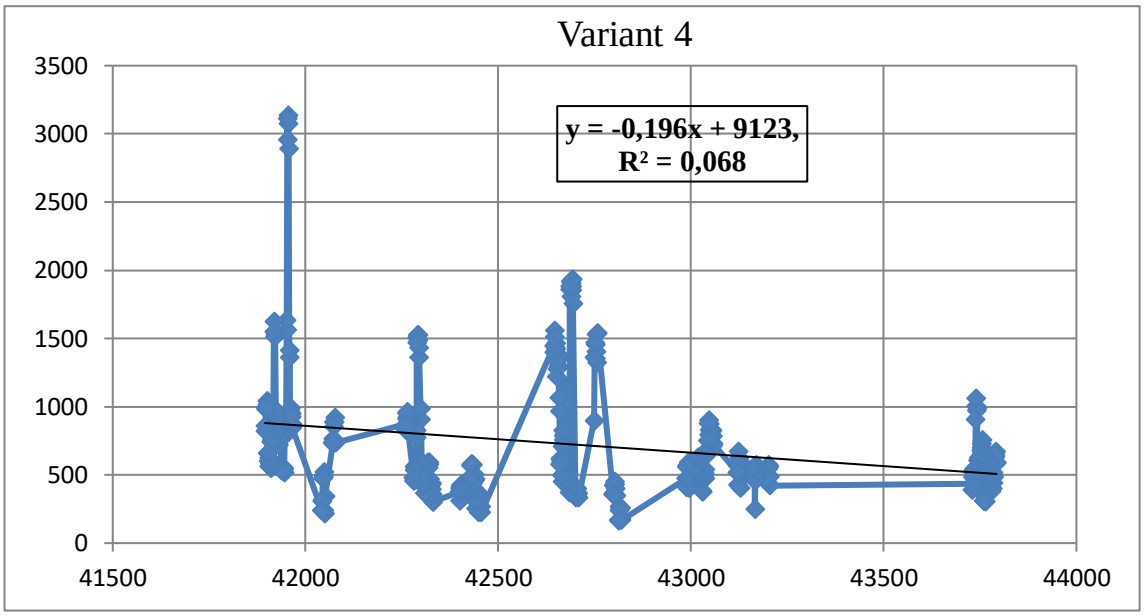


Figure III.28: La régression linéaire simple pour tous les segments avec le moyen de count**.**

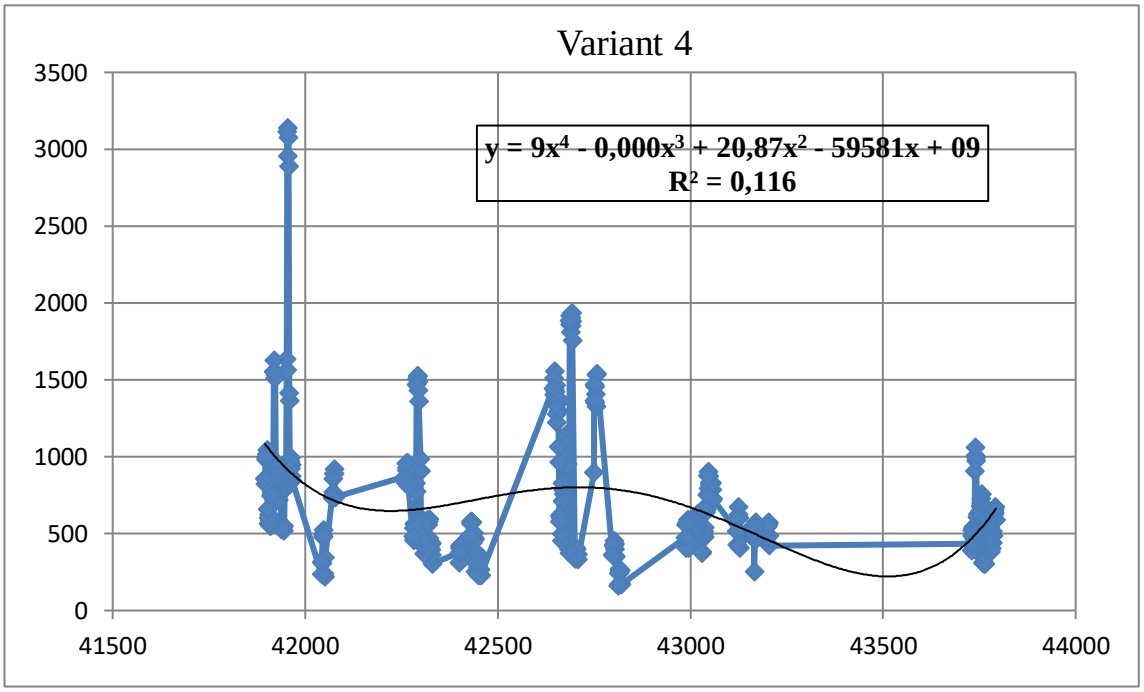


Figure III.29: La régression polynomiale pour tous les segments avec le moyen de count.

R^2= 0 ,068 dans la régression linéaire et 0,12 dans polynomiale, Cependant, le pourcentage est encore faible.

5. Prédiction du trafic routier par classification

Dans cette deuxième partie de notre étude, nous avons utilisé un algorithme de classification (apprentissage supervisé) pour la prédiction du trafic routier. Nous avons choisi d'utiliser la classification bayésienne naïve, vu sa simplicité d'implémentation et sa robustesse.

Pour mieux classifier le trafic, nous avons défini quatre classes essentielles, à savoir : Trafic Léger(TL), Trafic Moyen(TM), Trafic Dense(TD) et Trafic Très Dense(TTD). Ces classes ont été définies comme suit :

$$\begin{cases} \text{TL} \leq 450 \text{ véhicules} \\ 450 < TM \leq 550 \text{ véhicules} \\ 550 < TD \leq 850 \text{ véhicules} \\ \text{TTD} > 850 \text{ véhicules} \end{cases} \quad (10)$$

Toutefois, cette classification a été légèrement modifiée pour les variantes 3 et 4, où le trafic routier prend en considération tous les segments en question.

5.1 Implémentation pratique de Naïve Bayes dans R

Énoncé du problème: pour étudier un ensemble de données sur le trafic routier et créer un modèle d'apprentissage automatique qui prédit quelle classe de congestion peut-on avoir.

Description de l'ensemble de données: l'ensemble de données donné contient :

- ✓ La date
- ✓ Nombre des voitures dans une heure fixe

La variable de réponse ou la variable de sortie est la classe du trafic routier parmi les classes prédéfinies (TL, TM, TD, TTD)

Pour mieux avoir un ensemble de données adapté pour la classification, nous avons transformé le champ date en trois autres champs, à savoir : jour de la semaine, numéro de la semaine et nom du mois.

5.2 Algorithme de "naïve bayes"

Les packages installés pour mieux implémenter la méthode bayésienne naïve sont les suivants :

e1071:contient Fonctions d'analyse de classe latente, transformée de Fourier à court terme, clustering flou, machines à vecteurs de support, calcul du plus court chemin, clustering en sac, classificateur naïf de Bayes, k-plus proche voisin généralisé... [19]

caTools: L'un des principaux objectifs du package cartools est de comprendre le fonctionnement d'un système routier

caret: (classification and regression training) est une librairie pour R. Il couvre une large fraction de la pratique de l'analyse prédictive (classement et régression).[20]
klaR: pour la Classification et visualisation

Nous donnons dans ce qui suit, un extrait du programme implémenté.

```
install.packages("e1071")
install.packages("caTools")
install.packages("caret")
install.packages("klaR")
# Chargement du packages
library(e1071)
library(caTools)
library(caret)
library(klaR)
setwd("C:\traffic routtier")
traffic<- read.csv("nom_de_fichier.csv", sep = ';', header = TRUE)
# Diviser les données en train
# et tester les données
intrain<- createDataPartition(y = traffic$Traffic, p= 0.7, list = FALSE)
train_cl<- traffic[intrain,]
test_cl<- traffic[-intrain,]
# Mise à l'échelle des fonctionnalités
# Ajustement du modèle naïf de Bayes à l'ensemble de données d'entraînement
set.seed(120)
 # Réglage de la graine
x = train_cl[,-9]
y = train_cl$Traffic
classifier_cl = train(x,y,'nb',trControl=trainControl(method='cv',number=5) )
```

```
classifier_cl
# Prédire sur les données de test
y_pred<- predict(classifier_cl, newdata = test_cl)
y_pred
# Matrice de confusion
u <- union(y_pred, test_cl$Traffic)
t <- table(factor(y_pred, u), factor(test_cl$Traffic, u))
# Évaluation du modèle
confusionMatrix(t)
```

5.3 Variante 01

5.3.1. avenue 1

Jour	Semaine	Mois	Traffic
Samedi	1	Octobre	TTD
Dimenche	2	Octobre	TTD
Lundi	2	Octobre	TL
Mardi	2	Octobre	TM
Mercredi	2	Octobre	TD
Jeudi	2	Octobre	TTD
Vendredi	2	Octobre	TTD
Samedi	2	Octobre	TD
Dimenche	3	Octobre	TTD
Samedi	1	Novembre	TTD
Lundi	2	Novembre	TL
Mercredi	2	Novembre	TD
Jeudi	2	Novembre	TM
Vendredi	2	Novembre	TD

Table III.2: La classification pour l'avenue 1, variante 1.

```
Confusion Matrix and Statistics


      TL TTD TD
  TL   0   1  0
  TTD  0   1  0
  TD   0   0  1

Overall Statistics

               Accuracy : 0.6667
                 95% CI : (0.0943, 0.9916)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.7407

                  Kappa : 0.5

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TL Class: TTD Class: TD
Sensitivity                 NA     0.5000    1.0000
Specificity             0.6667     1.0000    1.0000
Pos Pred Value              NA     1.0000    1.0000
Neg Pred Value              NA     0.5000    1.0000
Prevalence              0.0000     0.6667    0.3333
Detection Rate          0.0000     0.3333    0.3333
Detection Prevalence    0.3333     0.3333    0.3333
Balanced Accuracy           NA     0.7500    1.0000
```

Figure III.30:Résultats d'Implémentation pratique de Naïve Bayes de l'avenue 1, variante 1.

Une matrice de confusion, également appelée matrice d'erreur est une disposition de tableau spécifique qui permet de visualiser les performances d'un algorithme, chaque ligne de la matrice représente les instances d'une classe réelle tandis que chaque colonne représente les instances d'une classe prédite, Le nom vient du fait qu'il permet de voir facilement si le système confond deux classes (c'est-à-dire qu'il étiquette généralement à tort l'une comme l'autre).[21]

Le taux de précision est de 66%, ce qui est un bon pourcentage, L'intervalle de confiance à 95% est fournie.

La sensibilité (sensitivity) est définie comme la proportion de résultats positifs sur le nombre d'échantillons qui étaient réellement positifs. Lorsqu'il n'y a pas de résultats positifs, la sensibilité n'est pas définie et une valeur NA est renvoyée. De même, lorsqu'il n'y a pas de résultats négatifs, la spécificité n'est pas définie et une valeur NA est renvoyée. Des déclarations similaires sont vraies pour les valeurs prédictives.

5.3.2 65 street (10 avenue-11 avenue)

Jour	Semaine	Mois	Traffic
Samedi	3	Septembre	TM
Dimenche	4	Septembre	TM
Lundi	4	Septembre	TL
Mardi	4	Septembre	TL
Mercredi	4	Septembre	TM
Jeudi	4	Septembre	TM
Vendredi	4	Septembre	TM
Samedi	4	Septembre	TM
Dimenche	5	Septembre	TM
Samedi	3	Octobre	TD
Dimenche	4	Octobre	TD
Lundi	4	Octobre	TM
Mardi	4	Octobre	TD
Mercredi	4	Octobre	TTD
Jeudi	4	Octobre	TTD

Table III.3: La classification pour 65street,variante 1.

```
Confusion Matrix and Statistics

          TL TTD TD TM
  TL       0   0  0  1
  TTD      0   0  0  1
  TD       0   0  1  0
  TM       0   1  0  0

Overall Statistics

                Accuracy : 0.25
                  95% CI : (0.0063, 0.8059)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.9375

                   Kappa : 0

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TL Class: TTD Class: TD Class: TM
Sensitivity                 NA      0.0000       1.00     0.0000
Specificity               0.75      0.6667       1.00     0.5000
Pos Pred Value              NA      0.0000       1.00     0.0000
Neg Pred Value              NA      0.6667       1.00     0.3333
Prevalence                0.00      0.2500       0.25     0.5000
Detection Rate            0.00      0.0000       0.25     0.0000
Detection Prevalence      0.25      0.2500       0.25     0.2500
Balanced Accuracy           NA      0.3333       1.00     0.2500
```

Figure III.31:Résultats d'Implémentation pratique de Naïve Bayes de65 street,variante 1.

Le rapport de précision est de 25% le pourcentage est faible.

5.3.3 108street (62 avenue-apex place):

Jour	Semaine	Mois	Traffic
Samedi	1	Avril	TD
Dimenche	2	Avril	TTD
Lundi	2	Avril	TD
Mardi	2	Avril	TTD
Mercredi	2	Avril	TD
Jeudi	2	Avril	TD
Vendredi	2	Avril	TD
Samedi	2	Avril	TD
Dimenche	3	Avril	TD
Samedi	2	Octobre	TM
Dimenche	3	Octobre	TL
Lundi	3	Octobre	TM
Mardi	3	Octobre	TM
Mercredi	3	Octobre	TL
Jeudi	3	Octobre	TM
Vendredi	3	Octobre	TM

Table III.4: La classification pour 108 street**,**variante 1.

```
Confusion Matrix and Statistics

          TD TTD TM
  TD       1   0  0
  TTD      1   0  0
  TM       0   0  1

Overall Statistics

               Accuracy : 0.6667
                 95% CI : (0.0943, 0.9916)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.7407

                  Kappa : 0.5

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TD Class: TTD Class: TM
Sensitivity             0.5000          NA    1.0000
Specificity             1.0000      0.6667    1.0000
Pos Pred Value          1.0000          NA    1.0000
Neg Pred Value          0.5000          NA    1.0000
Prevalence              0.6667      0.0000    0.3333
Detection Rate          0.3333      0.0000    0.3333
Detection Prevalence    0.3333      0.3333    0.3333
Balanced Accuracy       0.7500          NA    1.0000
```

Figure III.32:Résultats d'Implémentation pratique de Naïve Bayes de 108 street, variante 1.

Le taux de précision est de 66%, C'est aussi un bon pourcentage.

5.4 Variante 02

5.4.1 avenue 1

Jour	Semaine	Mois	Traffic
Samedi	1	Octobre	TM
Dimenche	2	Octobre	TM
Lundi	2	Octobre	TL
Mardi	2	Octobre	TL
Mercredi	2	Octobre	TL
Jeudi	2	Octobre	TL
Vendredi	2	Octobre	TM
Samedi	2	Octobre	TM
Dimenche	3	Octobre	TM
Samedi	1	Novembre	TTD
Dimenche	2	Novembre	TTD
Lundi	2	Novembre	TD
Mardi	2	Novembre	TD

Table III.5: La classification pour l'avenue 1, variante 2

```
Confusion Matrix and Statistics


      TL TM TD TTD
  TL   1  0  0   0
  TM   0  1  0   0
  TD   0  0  1   1
  TTD  0  0  0   0

Overall Statistics

               Accuracy : 0.75
                 95% CI : (0.1941, 0.9937)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : 0.05078

                  Kappa : 0.6667

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TL Class: TM Class: TD Class: TTD
Sensitivity               1.00      1.00    1.0000       0.00
Specificity               1.00      1.00    0.6667       1.00
Pos Pred Value            1.00      1.00    0.5000        NaN
Neg Pred Value            1.00      1.00    1.0000       0.75
Prevalence                0.25      0.25    0.2500       0.25
Detection Rate            0.25      0.25    0.2500       0.00
Detection Prevalence      0.25      0.25    0.5000       0.00
Balanced Accuracy         1.00      1.00    0.8333       0.50
```

Figure III.33:Résultats d'Implémentation pratique de Naïve Bayes de l'avenue 1, variante 2.

Le taux de précision est de %75, c'est un bon pourcentage pour la prédiction.

5.4.2 (65 street 10 avenue-11 avenue)

Samedi	3	Octobre	TD
Dimanche	3	Octobre	TL
Mercredi	3	Septembre	TM
Lundi	3	Octobre	TM
Mardi	3	Octobre	TM
Jeudi	3	Septembre	TTD
Vendredi	3	Septembre	TTD
Mercredi	3	Octobre	TTD
Lundi	4	Septembre	TD
Jeudi	4	Septembre	TD
Mercredi	4	Septembre	TL
Lundi	4	Octobre	TL
Mardi	4	Septembre	TM
Dimanche	4	Octobre	TM

Table III.6: La classification pour 65 street , variante 2.

```
Confusion Matrix and Statistics


      TD TL TM TTD
  TD   1  0  0   0
  TL   0  0  1   0
  TM   0  0  0   1
  TTD  0  0  0   0

Overall Statistics

               Accuracy : 0.3333
                 95% CI : (0.0084, 0.9057)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.7037

                  Kappa : 0.1429

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TD Class: TL Class: TM Class: TTD
Sensitivity             1.0000        NA    0.0000     0.0000
Specificity             1.0000    0.6667    0.5000     1.0000
Pos Pred Value          1.0000        NA    0.0000        NaN
Neg Pred Value          1.0000        NA    0.5000     0.6667
Prevalence              0.3333    0.0000    0.3333     0.3333
Detection Rate          0.3333    0.0000    0.0000     0.0000
Detection Prevalence    0.3333    0.3333    0.3333     0.0000
Balanced Accuracy       1.0000        NA    0.2500     0.5000
```

Figure III.34:Résultats d'Implémentation pratique de Naïve Bayes de65 street, variante 2.

Le taux de précision est de 33%, reste un peu faible.

5.4.3 (108street 62 avenue-apex place)

Jour	Semaine	Mois	Traffic
Samedi	1	Avril	TL
Dimanche	2	Avril	TL
Lundi	2	Avril	TM
Mardi	2	Avril	TL
Mercredi	2	Avril	TM
Jeudi	2	Avril	TM
Vendredi	2	Avril	TM
Samedi	2	Avril	TM
Dimanche	3	Avril	TL
Samedi	3	Septembre	TM
Dimenche	4	Septembre	TM
Lundi	4	Septembre	TL
Mardi	4	Septembre	TTD
Mercredi	4	Septembre	TL

Table III.7: La classification pour 108 street, variante 2.

```
Confusion Matrix and Statistics


          TL TM TD
  TL       3  0  0
  TM       0  2  0
  TD       0  0  1

Overall Statistics

               Accuracy : 1
                 95% CI : (0.5407, 1)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.01563

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TL Class: TM Class: TD
Sensitivity                1.0    1.0000    1.0000
Specificity                1.0    1.0000    1.0000
Pos Pred Value             1.0    1.0000    1.0000
Neg Pred Value             1.0    1.0000    1.0000
Prevalence                 0.5    0.3333    0.1667
Detection Rate             0.5    0.3333    0.1667
Detection Prevalence       0.5    0.3333    0.1667
Balanced Accuracy          1.0    1.0000    1.0000
```

Figure III.35:Résultats d'Implémentation pratique de Naïve Bayes de 108street, variante 2.

5.5 Variante 03

Jour	Semaine	Mois	Traffic
Mardi	1	Janvier	TL
Mercredi	1	Janvier	TL
Jeudi	1	Janvier	TL
Vendredi	1	Janvier	TL
Samedi	1	Janvier	TL
Dimenche	1	Janvier	TL
Lundi	2	Janvier	TL
Mardi	2	Janvier	TL
Mercredi	2	Janvier	TL
Jeudi	2	Janvier	TL
Vendredi	2	Janvier	TL
Samedi	2	Janvier	TL
Dimenche	2	Janvier	TL
Lundi	3	Janvier	TL
Mardi	3	Janvier	TD
Mercredi	3	Janvier	TD
Jeudi	3	Janvier	TD

Table III.8: La classification pour la variante 3.

```
Confusion Matrix and Statistics

          TL TD TTD TM
  TL      31  0   0  0
  TD       0 28   0  0
  TTD      0  0  23  0
  TM       0  0   0 26

Overall Statistics

                Accuracy : 1
                  95% CI : (0.9664, 1)
    No Information Rate : 0.287
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TL Class: TD Class: TTD Class: TM
Sensitivity              1.000    1.0000       1.000    1.0000
Specificity              1.000    1.0000       1.000    1.0000
Pos Pred Value           1.000    1.0000       1.000    1.0000
Neg Pred Value           1.000    1.0000       1.000    1.0000
Prevalence               0.287    0.2593       0.213    0.2407
Detection Rate           0.287    0.2593       0.213    0.2407
Detection Prevalence     0.287    0.2593       0.213    0.2407
Balanced Accuracy        1.000    1.0000       1.000    1.0000
```

Figure III.36:Résultats d'Implémentation pratique de Naïve Bayes de variante 3.

5.6 Variante04 :

Jour	Semaine	Mois	Traffic
Mardi	1	Janvier	TD
Mercredi	1	Janvier	TD
Jeudi	1	Janvier	TD
Vendredi	1	Janvier	TD
Samedi	1	Janvier	TD
Dimenche	1	Janvier	TD
Lundi	2	Janvier	TD
Mardi	2	Janvier	TM
Mercredi	2	Janvier	TM
Jeudi	2	Janvier	TL
Vendredi	2	Janvier	TL
Samedi	2	Janvier	TL
Dimenche	2	Janvier	TL
Lundi	3	Janvier	TL
Mardi	3	Janvier	TM
Mercredi	3	Janvier	TM
Jeudi	3	Janvier	TD

Table III.9: La classification pour la variante 4**.**

```
Confusion Matrix and Statistics

          TD TM TL TTD
  TD      25  0  0   0
  TM       0 28  0   0
  TL       0  0 30   0
  TTD      0  0  0  25

Overall Statistics

                Accuracy : 1
                  95% CI : (0.9664, 1)
    No Information Rate : 0.2778
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: TD Class: TM Class: TL Class: TTD
Sensitivity             1.0000    1.0000    1.0000     1.0000
Specificity             1.0000    1.0000    1.0000     1.0000
Pos Pred Value          1.0000    1.0000    1.0000     1.0000
Neg Pred Value          1.0000    1.0000    1.0000     1.0000
Prevalence              0.2315    0.2593    0.2778     0.2315
Detection Rate          0.2315    0.2593    0.2778     0.2315
Detection Prevalence    0.2315    0.2593    0.2778     0.2315
Balanced Accuracy       1.0000    1.0000    1.0000     1.0000
```

Figure III.37:Résultats d'Implémentation pratique de Naïve Bayes de variante 4.

6. Conclusions

Dans ce chapitre, nous avons utilisé deux modèles de machine learning pour la prédiction du trafic routier, à savoir : la régression et la classification bayésienne naïve.

Le tableau suivant donne une synthèse des meilleurs résultats (taux d'erreur) obtenus pour chaque variante et pour chaque modèle utilisé.

Variante	Meilleur taux d'exactitude obtenu par la Régression linéaire ou polynomiale	Meilleur taux d'exactitude obtenu par Naïve Bayes
Variante 1	**96%**	**66%**
Variante 2	**90%**	**75%**
Variante 3	6%	100%
Variante 4	12%	100%

Table III.10: Taux d'exactitude obtenus pour chaque variante et pour chaque modèle utilisé.

Les résultats obtenus montrent que la Régression donne généralement un taux d'erreur acceptable (et souvent très bon) pour les variantes 1 et 2. De même la méthode Naïve Bayes a donné pour les deux premières variantes un taux d'erreur de 66% et 75%.

Cependant, on remarque que la régression n'a pas abouti à un taux acceptable pour les variantes 3 et 4. De même, la classification Bayésienne Naïve n'a pas donné de bons résultats pour les deux dernières variantes.

On peut interpréter ces résultats par le fait que la prédiction du trafic routier sur toute la journée n'est pas généralement adéquate, et que la prédiction qui peut donner de meilleurs résultats est celle liée à un instant donné (ou heure fixe).

CONCLUSION GENERALE

La ville moderne se transforme progressivement en une ville intelligente. L'accélération de l'urbanisation et la rapidité la croissance de la population urbaine exerce une forte pression sur la gestion du trafic urbain. Une prévision précise du trafic est essentielle pour de nombreuses applications du monde réel. Par exemple, la prévision du volume de trafic peut aider la ville à réduire la congestion, la prédiction de la demande de voitures peut inciter au covoiturage les entreprises pré-attribuent des voitures aux régions à forte demande.

Les ensembles de données croissants disponibles sur le trafic nous offrent du potentiel de nouvelles perspectives pour explorer ce problème.

Dans ce mémoire, nous avons étudié le problème de prédiction de la congestion (ou le trafic routier) d'une ville. Cette étude est réalisée en utilisant les modèles de machine learning, à savoir : la régression et la classification Bayésienne Naïve (apprentissage supervisé). Nous avons utilisé des données réelles issues de la ville de New York durant la période de 2014 à 2019.

Notre étude a donné de bons résultats quand il s'agit de prédire le trafic routier à un instant donné, voisinant un taux d'erreur entre 90 et 96% pour la régression et entre 66 et 75% pour la classification Naïve Bayes. Cependant, nous avons remarqué que prédire le trafic routier sur toute la journée n'a pas été adéquat sur les données que nous utilisées.

Nous prétendons arriver aux objectifs tracés au début de ce projet, toutefois, nous pouvons recommander de futures recherches sur ce sujet, notamment avoir des données réelles d'une ville de notre pays et appliquer d'autres modèles de machine Learning pour arriver à des résultats pertinents.

REFERENCES BIBLIOGRAPHIQUE

[1] Bourbia, F. (2005). Le problème de la circulation et du stationnement dans le centre ville de Remerciements

[2] MESSEGUEM Chahrazed,Data classification using deep learning approach (Master), université de M'sila , 23/06/2020.

[3] Site web : https://www.infociments.fr/sites/default/files/article/fichier/CT-T32.pdf , consulté le : décembre 2020.

[4] Xu, G., Jin, H. H., & Liu, J. (2013). Traffic status prediction and analysis based on mining frequent subgraph patterns. *Advanced Materials Research*, *605–607*, 2543–2548.

[5] Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2021). Deep Learning on Traffic Prediction: Methods, Analysis and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, 1–16.

[6] Julien JACQUES, Ricco Rakotomalala ,(2014). (n.d.) . Introduction au Data Mining , Université de Lyon.

[7] Site web: https://towardsdatascience.com/5-types-of-regression-and-their- properties-c5e1fa12d55, consulté le: 11/09/2020

[8] Site web: https://analyticsinsights.io/top-5-des-types-de-regression/, consulté le: février 2020

[9] Dr. LOUNNAS Bilal,(2021), Induction, D. T. (n.d.). Information retrieval (IR) and Data mining (DM) Prediction : using features to predict unknown or, (Dm), 1–26.Université de M'sila.

[10] Ouali, Choayb,(2015),Classification automatique de textes, université de M'sila .

[11] Site web: https://www.edureka.co/blog/naive-bayes-in-r/, consulté le: 26/05/2020

[12] Ferecatu, M., & Crucianu, M,(2014), (n.d.). modèles graphiques (RCP209) Plan du cours Objectif.

[13] Site web: https://mrmint.fr/introduction-k-nearest , consulté le: 02/10/2020

[14] Li, Y., & Shahabi, C. (2018). A brief overview of machine learning methods for shortterm traffic forecasting and future directions. *SIGSPATIAL Special*, *10*(1), 3–9.

[15] Site web: https://www.lebigdata.fr/machine-learning-et-big-data , consulté le: 03/02/2021

[16] Site web: https://cours-informatique-gratuit.fr/dictionnaire/office-excel/ , consulté le:21/06/2020

[17] Citadelle, L. (2012/2013). Introduction - Qu ' est-ce que le logiciel R ? Objectifs de ce TP et commencements, 1–9.

[18] Site web: https://quanti.hypotheses.org/488 , consulté le: 07/03/2020

[19] Site web: https://cran.r-project.org/web/packages/e1071/index.html , consulté le: 23/05/2021

[20] Tanagra, T. (2018). Machine Learning avec le package « caret », 1–32.

[21] Site web: https://en.wikipedia.org/wiki/Confusion_matrix , consulté le : 13/05/2021