

DOSSIER

Titre du dossier : Rapport Projet Outils et Méthodes pour la simulation

Matière : Outils et Méthodes pour la simulation

Niveau : Master 2 Mention : IMA Spécialité ou Parcours : IAD

A l’attention de Malinge Lucie

Session : Janvier 2023

HUANG Chen et LAMBERT Clara

Faculté de Scicnes

IMA

Année Universitaire : 2022-2023



**CHARTE DE NON PLAGIAT**

**Protection de la propriété intellectuelle**

Tout travail universitaire doit être réalisé dans le respect intégral de la propriété intellectuelle d’autrui. Pour tout travail personnel, ou collectif, pour lequel le candidat est autorisé à utiliser des documents (textes, images, musiques, films etc.), celui-ci devra très précisément signaler le crédit (référence complète du texte cité, de l’image ou de la bande-son utilisés, sources internet incluses) à la fois dans le corps du texte et dans la bibliographie. Il est précisé que l’UCO dispose d’un logiciel anti-plagiat dans lms.uco.fr, aussi est-il demandé à tout étudiant de remettre à ses enseignants un double de ses travaux lourds sur support informatique.

*Cf. « Prévention des fraudes à l’attention des étudiants »*

Nous soussignons, Huang Chen et Lambert Clara, étudiantes en Master MIASHS s’engageons à respecter cette charte.

Fait à Angers, le 02/01/2023

Sommaire ou table des matières

Table des matières

[Introduction III](#_Toc123655956)

[1.Statisiques descriptives IV](#_Toc123655957)

[2. Graphiques et analyses IX](#_Toc123655958)

[3. Modèles de prédiction XVII](#_Toc123655959)

[Conclusion XIX](#_Toc123655960)

Introduction

Dans le cadre de notre master nous avions à réaliser un projet dans la matière « Outils et Méthodes pour la simulation ».  
Ce projet consiste à réaliser une **prédiction** sur la base de données du trafic Irigo. Pour cela il nous a été demandé de faire des statistiques descriptives, graphiques et analyses et enfin la prédiction du retard des bus avec des données allant de mi 2019 à fin 2019.

Nous avons choisi de **prédire les écarts secondes** des bus. Donc nos modèles sont des **modèles de régression**.

# 1.Statisiques descriptives

Une image contenant texte

Description générée automatiquement

Notre base de données Irigo comporte 743 821 lignes et 28 colonnes. La première colonne ne compte pas car ce sont les index donc on est plutôt à 27 colonnes.

On a 11 variables quantitatives pour 16 qualitatives.

Une image contenant texte

Description générée automatiquementValeurs Nulles

On constate qu’il y a des données manquantes. Par exemple pour les colonnes fields.numarret, fields.nomarret et fields.mnemoarret il y a 908 lignes avec des données vides. On sait que les 3 colonnes sont liées. Si on vérifie on peut confirmer cela car fields.numarret est l’identifiant de l’arret de bus, fields.nomarret son nom et fields.mnemoarret son MNE.

On va donc **supprimer** toutes les lignes avec des données manquantes pour avoir une meilleure base pour la prédiction.

Colonnes : tri et renommage

Une image contenant texte

Description générée automatiquementDans un premier temps il est important de comprendre ce que représente chaque colonne. Pour cela nous nous sommes rendues sur le site d’irigo et on a pu faire une description brève de chaque colonne :

On supprime les colonnes qui ne sont pas corrélées à notre de variable de prévision « ecart ».

Cependant on garde les colonnes « id » des colonnes utiles : fields.idligne, l’identifiant de fields.nomligne par exemple.

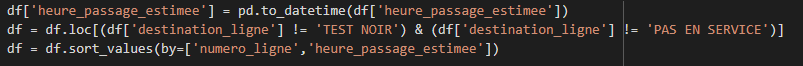
On va supprimer fields.y, fields.x et geometry.coordonnees car en vérifiant elles ne sont pas correctes et mènent à des zones géographiques non cohérentes.

On va également renommer nos colonnes pour qu’elles soient plus faciles et rapides à manipuler. Exemple sur 2 colonnes :

Modifications sur la base

On va faire quelques modifications sur notre dataframe.

Tout d’abord en regardant plus en détail on constate qu’il y a des bus « PAS EN SERVICE » et « TEST NOIR » de référencés. Ces derniers ne nous intéressent pas car on cherche à prédire les retards des bus en service. On va donc supprimer les lignes de ces bus et retrier le dataframe avec les dates en ordre croissant pour une meilleure lecture.



Ensuite on va ajouter des colonnes afin de faciliter la création de nos graphiques et la compréhension de notre base.

On va ajouter une colonne ‘Jour’ qui indiquera le jour de la semaine, une colonne ‘Calendrier’ qui reprendra la date au format mois/jour/année et une colonne ‘Heure’ qui contiendra uniquement les heures de passage.



Ensuite on ajoute une colonne désignant l’état de la date : Vacances ou période scolaire. On regarde nos dates min et max afin de savoir précisément sur quelle période on réalise nos prédictions.

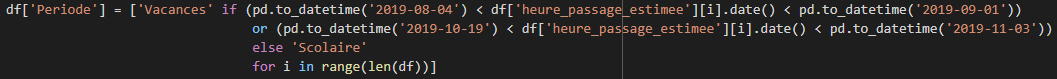
Une image contenant texte

Description générée automatiquement

On aura 2 périodes de vacances :

* Du 2019-08-04 au 2019-09-01 : Vacances d’été
* Du 2019-10-19 au 2019-11-03 : Vacances de la Toussaint

On peut créer notre nouvelle colonne :



Quelques chiffres…

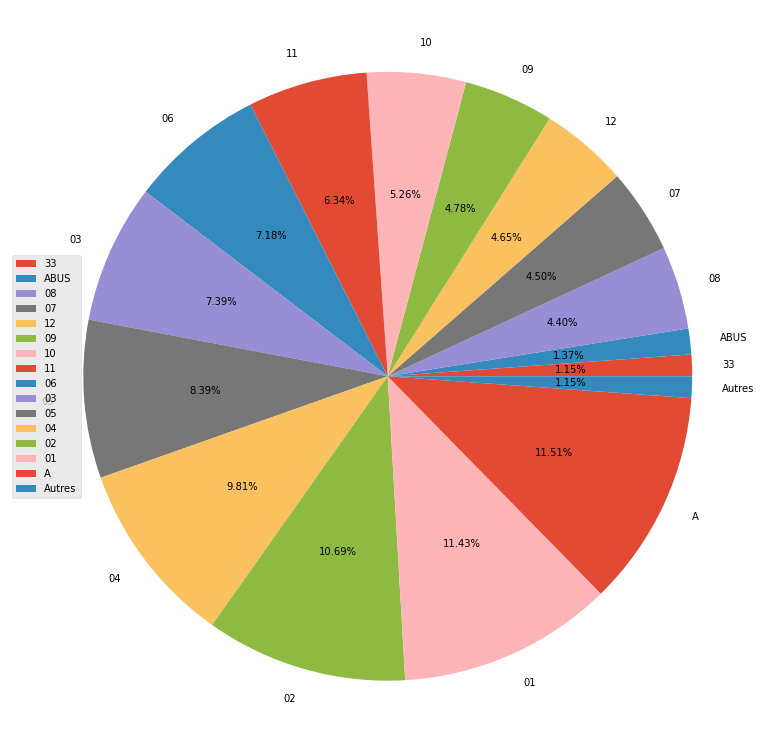
Après modification sur le dataframe on a 691 376 lignes et 21 colonnes.

On constate qu’en moyenne l’ecart du temps de passage est de 178 secondes. C’est-à-dire qu’en moyenne un bus est en retard d’environ 3 minutes.

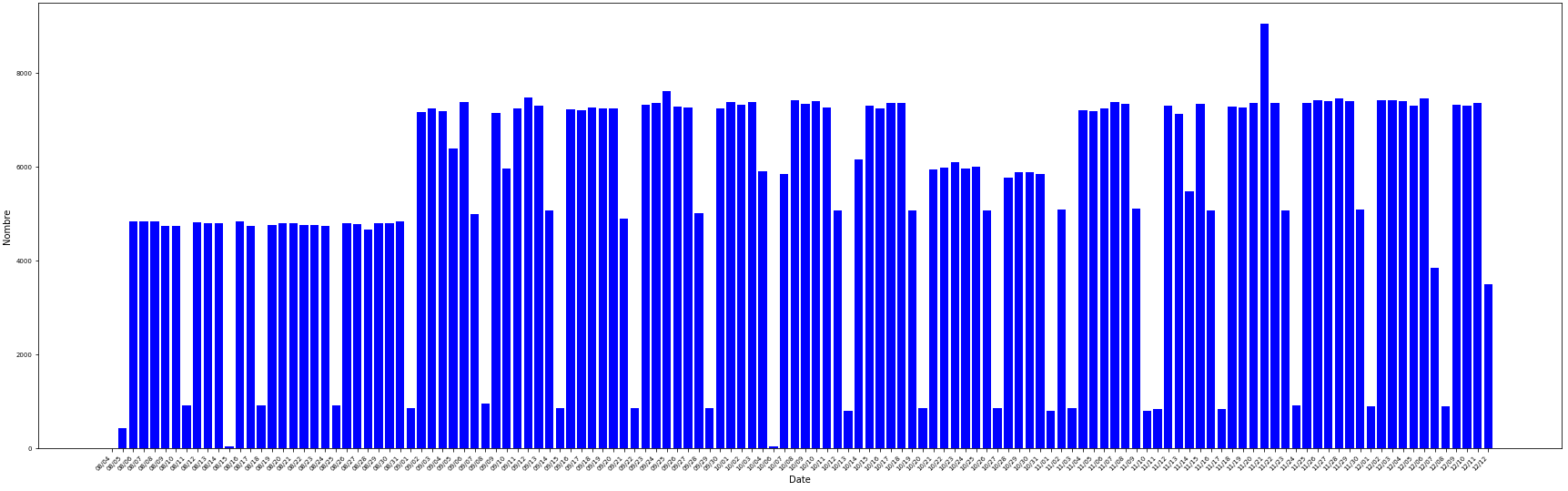
# 2. Graphiques et analyses

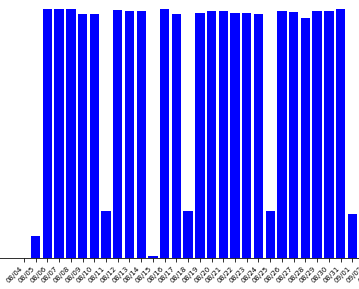
Tous les graphiques sont basés sur la période de mi 2019 à fin 2019.

Graphique 1 : Pourcentages des lignes en circulation de mi 2019 à fin 2019

Ce graphique représente la répartition des lignes en service. Il y a plus de 40 lignes sur le réseau. Il y a les lignes de villes (les bus), les lignes de campagnes (les cars) et le tramway (il n’existait que la ligne A à l’époque).  
On voit directement les lignes les plus importantes : le tramway A, le 01 et le 02 sont tous les 3 au dessus des 10%. C’est-à-dire que c’est eux qui passent le plus souvent. Par exemple le bus 02 représente 10.69% des passages sur la période donnée.

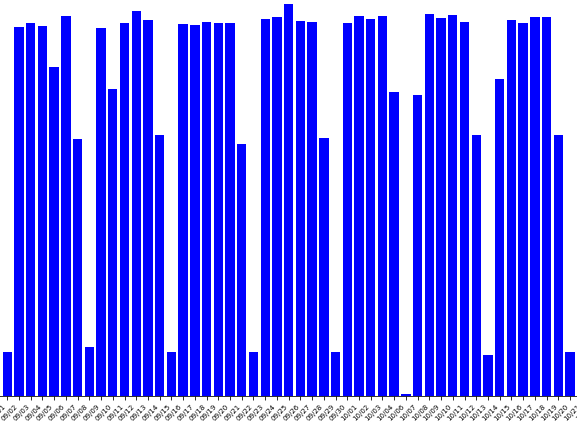
Graphique 2 : Nombre de bus en ligne cumulé par jour



On constate directement des tendances, on a une première partie très linéaire avec des valeurs « basses » :

Si on regarde les dates on est en période de vacances scolaire (du 4 aout au 1 septembre). Donc il y a moins d’horaires disponibles vu que les personnes prennent moins le bus. Les barres très basses correspondent aux Dimanche, là où les bus circulent très peu. Le 15 août on remarque une chute drastique des bus en circulation, cela s’explique puisque ce jour est férié. On a une tendance à 4900 bus en circulation en moyenne en période de vacances d’été.

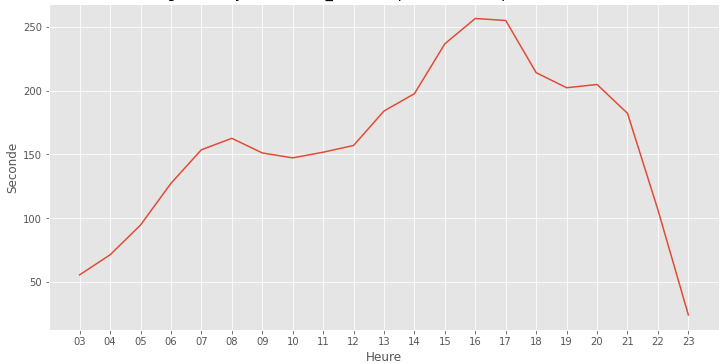
On constate un pique au 2 septembre, le jour de la rentrée scolaire :



On est à environ 7000 bus en circulation en période scolaire.

Toutes les barres « anormalement » basses, c’est-à-dire qui ne tombe pas un dimanche comme le 1er Novembre sont expliquées par des jours fériés. Sauf pour le 6 octobre, on pense qu’il s’agit d’un problème dans les données car on a aucune donnée au 5 octobre.

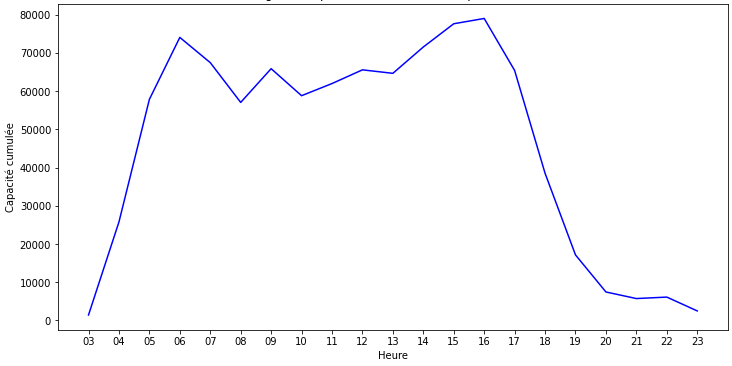
Graphique 3 : Moyenne des écarts seconde pour tous les bus dans une journée



On constate un gros pique de retard entre 16h et 18h avec une moyenne de 250 secondes environ. Cela parait logique car ce sont les heures de débauche et donc la circulation est beaucoup plus dense à ces horaires-là.

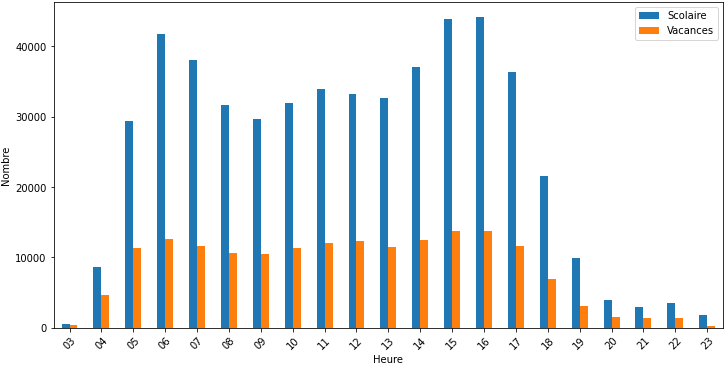
On a aussi une augmentation des retards de 7h à 9h. Encore une fois cela est cohérent puisque c’est l’heure où les gens partent au travail et où les étudiants vont en cours. Le trafic est donc ralenti car il y a plus de circulation.

Graphique 4 : Moyenne des capacités des bus par heure

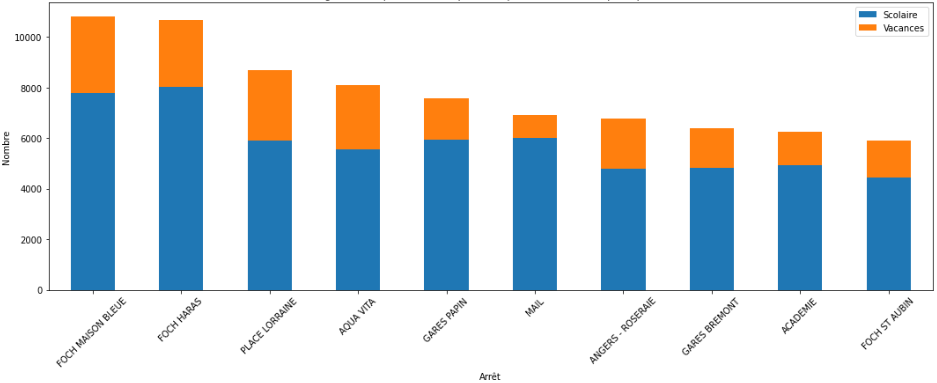


Ce graphique montre l’affluence ou la capacité cumulée des bus cumulée sur une journée.

Par exemple si on prend le point le plus haut à 16h, en tout il y a eu environ 79000 passagers cumulés sur la période de mi 2019 à fin 2019. Ce pic parait logique vu que c’est une heure de forte affluence, tout comme entre 6h et 7h.

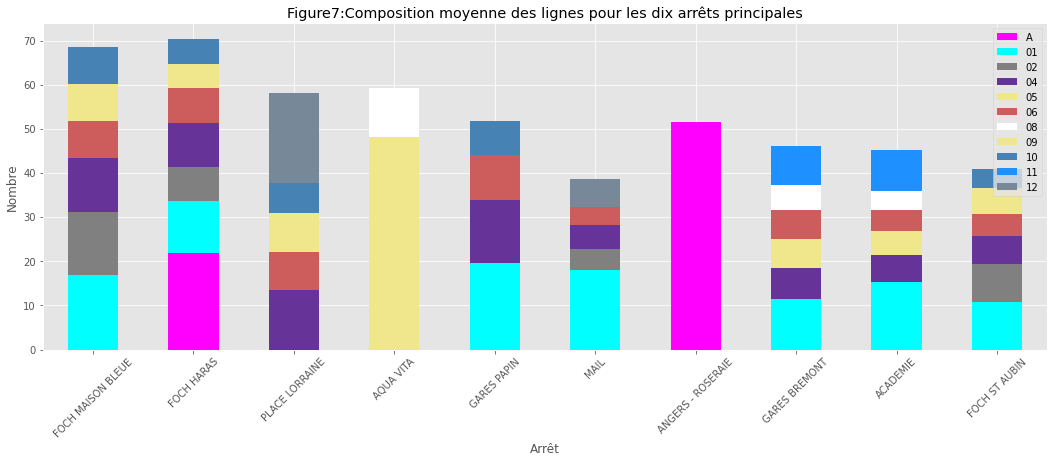
Graphique 5 : Nombre de bus cumulés (en période scolaire / en période de vacances) par heure

Ce graphique compare le nombre de transport en commun en circulation en vacances et en période scolaire. On constate que dans la période scolaire il y a des différences flagrantes entre les heures. On a des pics entre 6h-7h et 15h – 17h avec plus de 40 000 bus qui ont circulé. Cependant cette différence se fait moins ressentir en période de vacances. Au maximum on atteint les 12 000 bus en circulation à 16h.

Graphique 6 : Fréquence moyenne des bus (en période scolaire / en période de vacances) sur les 10 arrêts principaux

Ce graphique compare le nombre de bus passant aux principaux arrêts en période scolaire et de vacances. Les arrêts les plus desservis sont ceux du centre ville : foch maison bleue et foch haras, que ce soit en vacances ou non avec environ 8000 passages en période scolaire et 3000 en vacances.

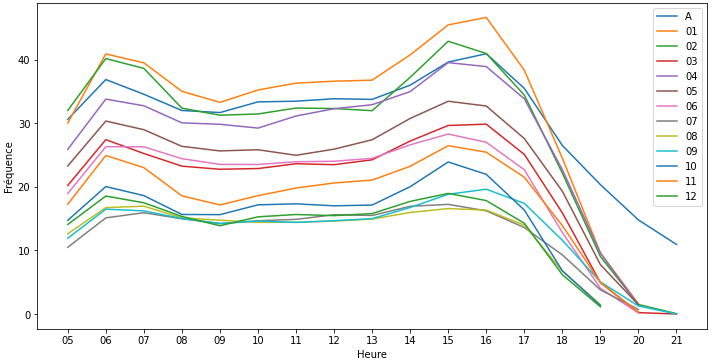
Graphique 7 : Composition des lignes sur les 10 arrêts principaux



Ce graphique montre la composition moyenne des lignes des bus principaux et le tramway A sur les arrêts principaux.

Nous avons trouvé qu’il y a en moyenne 5 bus qui dessert ces arrêts. On constate que la ligne 1 est présente dans 6 arrêts avec de nombreux passages, environ 20 000 à l’arrêt GARES PAPIN par exemple.

Graphique 8 : Nombre moyen des transports en circulation par heure

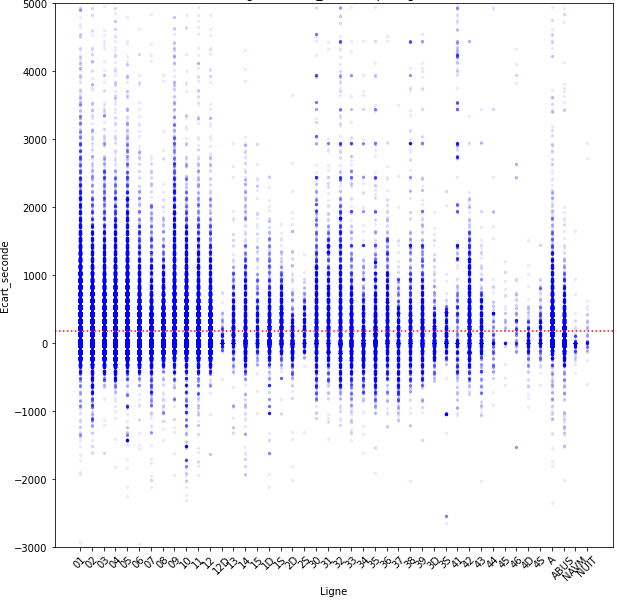


Ce graphique nous montre la moyenne du nombre de bus par heure allant des lignes 1 à 12 et le tramway A, en circulation par heure sur la période mi 2019 à fin 2019.

Le 1 se retrouve en première place, c’est le bus qui circule le plus et donc où la demande est forte. On retrouve toujours les mêmes pics entre 6h et 7h, et entre 16h et 17h pour tous les bus.

En moyenne il y a 48 bus 1 en service à 16h.

Graphique 9 : Ecart seconde par ligne



Ce graphique permet de facilement voir les retards ou avances des lignes Irigo. On constate que les retards touchent toutes les lignes à plus ou moins grande échelle. A première vu on peut voir que c’est la ligne 1qui est le plus touchée par les retards. Mais toutes les premières lignes sont touchées par des retards importants. On voit que les dernière lignes (de car notamment) on beaucoup moins de retard. La ligne 45 n’en a quasiment pas. Mais cela est logique car elle a beaucoup moins d’horaire que les lignes de bus.

Graphique 10 : Moyenne de l’écart seconde pour les lignes de bus de 1 à 12 et la ligne de tramway en période scolaire / de vacances

Une image contenant table

Description générée automatiquementUne image contenant table

Description générée automatiquementCe graphique nous montre en moyenne l’écart seconde des bus par heure en période scolaire. C’est la ligne 1 qui a le plus de retard par heure, elle est en moyenne en retard tout au long de la journée. Pour les lignes 2 et 3, c’est plutôt le soir entre 17h et 20h qu’elles sont très en retard. Ces lignes passent dans le centre-ville et sont énormément empruntées par les collégiens et lycéens, cela peut expliquer en partie les retards à ces horaires. Le tramway A n’a quasi pas de retard, ce qui est normal vu que la circulation est très réglementée et bien gérée pour son passage.

On a tout de suite moins de retard en période de vacances. Même si la ligne 2 est beaucoup impacté le soir. Cela s’explique puisque c’est une des lignes qui circulent le plus en vacances.

# 3. Modèles de prédiction

Comme dit précedement on cherche à prédire le retard donc nous allons faire des modèles de classification.

On a d’abord traduit toutes nos variables qualitatives en variables quantitatives.

Par exemple un extrait de code qui transforme les données de vehicule\_etat :

Une image contenant texte

Description générée automatiquement

On stocke chaque valeur unique dans un dictionnaire et on lui attribue un chiffre.

Ensuite on a sélectionné les colonnes pour le modèle : vehicule\_etat, nom\_ligne, numero\_ligne, vehicule\_modele, destination\_ligne, nom\_arret, jour, période, heure\_passage\_estimee.

Une image contenant texte

Description générée automatiquementOn va également créer une colonne « Retard » composé de 0 (pas en retard) et de 1 (en retard), qui sera notre variable à prédire.

On considère un bus en retart quand l’écart seconde est supérieur à 60.

Une image contenant texte

Description générée automatiquementOn peut maintenant sélectionner nos variables explicatives et notre variable à prédire.

On pense à retirer la colonne « ecart\_seconde » sinon notre précision sera toujours égale à 1.

On va ensuite séparer nos données avec un jeu d’entrainement (80%) et un jeu de test (20%)



Une image contenant texte

Description générée automatiquementOn commence avec un **Random Forrest** avec les paramètres par défaut :

Résultat

Une image contenant texte

Description générée automatiquement

On commence avec 77% de vrais positifs, une précision de 74% et un recall de 76%, c’est plutôt bien pour un premier modèle mais on va essayé de l’améliorer.

Après tunning des hyperparamètres principaux on a constaté que les résultats étaient similaires.

On a testé un modèle du **plus proche voisin** mais les résultats étaient très mauvais :

Une image contenant texte

Description générée automatiquement

Résultats **SVM** :

Une image contenant texte

Description générée automatiquement

Conclusion

Après plusieurs tests c’est le Random Forest qui reste le meilleur modèle avec un accuracy de 77%, une précision de 74% et un recall de 76%. Les 3 sont supérieur à 70% ce qui fait de notre modèle un bon modèle.

Afin d’améliorer celui-ci il faudrait procéder à un tunning complet des hyperparamètres. Cependant cela demande beaucoup de temps et une très bonne machine, ce que nous n’avons pas.