



Projet Machine Learning avec Python : Prédiction des incidents à San Francisco

ALLARD Lucas
HEDFI Houeida
MEFENZA NOUNTU Thierry

Contents

1	Introduction et problématique	2
2	Description de la base de données	2
2.1	Base de données originale	2
2.2	Nettoyage de la base	4
2.3	Création d'une nouvelle variable	4
3	Visualisation des données	6
3.1	Distribution des modalités des variables temporelles	6
3.2	Distribution des modalités des variables géographiques	8
4	Prédiction nombre d'appels à San Francisco	10
4.1	Prédiction du nombre d'appels par heure et par quartier à San Francisco	10
4.2	Prédiction du nombre d'appels à San Francisco en ajoutant d'autres variables	12
4.2.1	Prédiction avec la météo de San Francisco heure par heure	12
4.2.2	Prédiction avec les jours fériés de San Francisco heure par heure	12
4.3	Prédiction multiple du nombre d'appels par heure et par catégorie à San Francisco .	13
5	Conclusion et perspectives	14

1 Introduction et problématique

Dans ce projet, nous travaillons sur une base de données d'appels des usagers adressés à un centre d'appel de tous les unités de sapeurs pompiers de la ville de San Francisco pour déclarer un incident. Lorsqu'un incident est déclaré, les équipes des différentes unités des sapeurs pompiers s'organisent pour intervenir sur le lieu de l'incident. Si l'équipe la plus proche d'un incident déclaré ne dispose pas des ressources suffisantes pour intervenir, ceci peut alors avoir des conséquences graves.

Un modèle permettant de prédire avec précision le nombre d'appels en fonction du quartier, d'une date et d'une heure donnée permettrait aux équipes de sapeurs pompiers de mieux s'organiser afin qu'un nombre suffisants d'agents soient disponibles et capables de répondre aux incidents. Notre objectif est donc de contribuer à l'allocation des ressources des sapeurs pompiers de la ville de San Francisco en prédisant le nombre d'appels par quartier et pour une date donnée.

Notre travail est divisé en trois parties. La première partie porte sur la description de la base de données et sur le nettoyage de celle-ci, le deuxième partie présente la visualisation des données utile pour mieux comprendre et traiter nos données en vue d'une bonne prédiction, la troisième partie présente nos différents modèles de prédiction et enfin nous terminons par une conclusion.

2 Description de la base de données

2.1 Base de données originale

Nos données proviennent du site des sapeurs pompiers de la ville de San Francisco(<https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3>). Elles comprennent toutes les réponses des centres d'appels des différentes unités aux appels des usagers du 01 janvier 2001 à aujourd'hui et est mis à jour chaque jour. Le fichier que nous avons téléchargé à la date du 12 janvier 2021 comportait 49 colonnes et environ 5500000 lignes. Les 49 colonnes (ou variables) peuvent être regroupées comme suit:

- le numéro d'appel(Call Number), le numéro d'incident(Incident Number), RowID(identifiant unique de l'appel)
- l'identifiant de l'unité: Unit ID, Unit sequence in call dispatch, Unit sequence in call dispatch, Station Area

- l'adresse de l'incident(Les adresses sont associées à un numéro de bloc, à une intersection ou à une cabine téléphonique, et non à une adresse spécifique): Address(adresse d'un midbloc proche de l'incident), City(ville de l'incident), Zipcode of Incident, Location(latitude et longitude de l'adresse), Neighborhoods - Analysis Boundaries,Zip Codes, Neighborhoods (old), Neighborhoods, SF Find Neighborhoods, Neighborhoods - Analysis Boundaries(Quartier de l'incident)
- Lieu particulier: Central Market/Tenderloin Boundary Polygon - Updated
- Unités: Battalion(district de réponse d'urgence(en cas d'incendie)), Station Area(Unité la plus proche du lieu de l'incident), Box(bloc d'unités associé à l'adresse de l'incident), ALS Unit, Unit Type, Unit sequence in call dispatch(nombre indiquant la position à laquelle une unité a été assignée à un appel)
- type de priorité(urgence ou non-urgence): Original Priority(priorité appel initial), Priority, Final Priority(priorité appel final)
- la date de l'appel: Call Date, Watch Date
- les variables de temps: Received DtTm(heure de réception de l'appel),Entry DtTm(heure d'enregistrement de l'appel) , Dispatch DtTm(heure de transmission de l'appel à l'unité), Response DtTm(heure à laquelle l'unité se met en route pour le lieu de l'incident), On Scene DtTm(l'heure d'arrivée sur le lieu de l'incident), Transport DtTm(heure de départ pour l'hôpital si l'unité est une ambulance), Hospital DtTm(heure de d'arrivée à l'hôpital si l'unité est une ambulance), Available DtTm(l'heure où l'unité n'est plus assignée à l'appel),
- Type de district associé à l'adresse de l'incident: Supervisor District, Shape, Supervisor Districts, Fire Prevention, Districts, Current Police Districts, Police Districts,Civic Center Harm Reduction Project Boundary, Current Police Districts 2, Current Supervisor Districts, Fire Prevention District
- le type d'appel: Call Type Group(groupe du type d'appels(feu, alarme, potentiellement mortel, pas mortel)), Call Type(type d'incidents(ici il y en environ 32))
- le type de réponse apporté: Call Final Disposition
- nombre d'alarmes (entre 0 et 5) associées à l'incident: Number of Alarms

2.2 Nettoyage de la base

Pour nettoyer notre base de données, nous avons procédé en plusieurs étapes.

1. Sélection des variables pertinentes pour notre problématique

Parmi les 49 variables, certaines se répètent tandis que d'autres ne nous semblent pas pertinentes pour répondre à notre préoccupation. Nous avons finalement retenu les variables suivantes: Call Type, Received DtTm, Address, Zipcode of Incident, Battalion, Station Area, Box, Final Priority, Call Type Group, Neighborhoods - Analysis Boundaries, Location.

2. Suppression des doublons et des valeurs manquantes

Comme il y a plusieurs appels pour déclarer le même incident, nous ne retenons qu'un seul appel et supprimons les autres appels déclarant le même incident. Ensuite nous supprimons toutes les lignes ayant des colonnes vides.

3. Suppression des lignes enregistrées en 2021

Dans notre étude nous regardons aussi l'évolution des incidents par année afin de mieux appréhender nos données et comme l'année 2021 n'est pas assez représenté, nous supprimons les lignes enregistrées cette année.

2.3 Crédation d'une nouvelle variable

Dans la suite, la variable que nous essayerons de prédire est le type d'incidents (Call Type) et il y a environ 32 type d'incidents (voir la figure 1 ci-contre). On peut voir que le type Medical Incident a une fréquence de près de 80% (Ce qui est assez gênant car pour avoir un bon prédicteur, il suffirait de prédire Medical Incident la plupart du temps). Nous avons scindé le type Medical Incident en deux types: Medical Potentially Life-Threatning et Medical Non Life-Threatning. Aussi nous avons beaucoup trop de type d'incidents avec des fréquences quasi nulles. En regroupant certains types d'incidents et en scindant le type Medical Incident en deux, nous avons donc crée une nouvelle variable ayant six types d'incidents: Medical Potentially Life-Threatning, Medical Non Life-Threatning, Alarms, Fire, Incidents et Others. Dans la suite, nous travaillerons exclusivement avec cette nouvelle variable, 'Call Category', en lieu et place de la variable Call Type. La figure 2 présente les fréquences des modalités de la variable 'Call Category'.

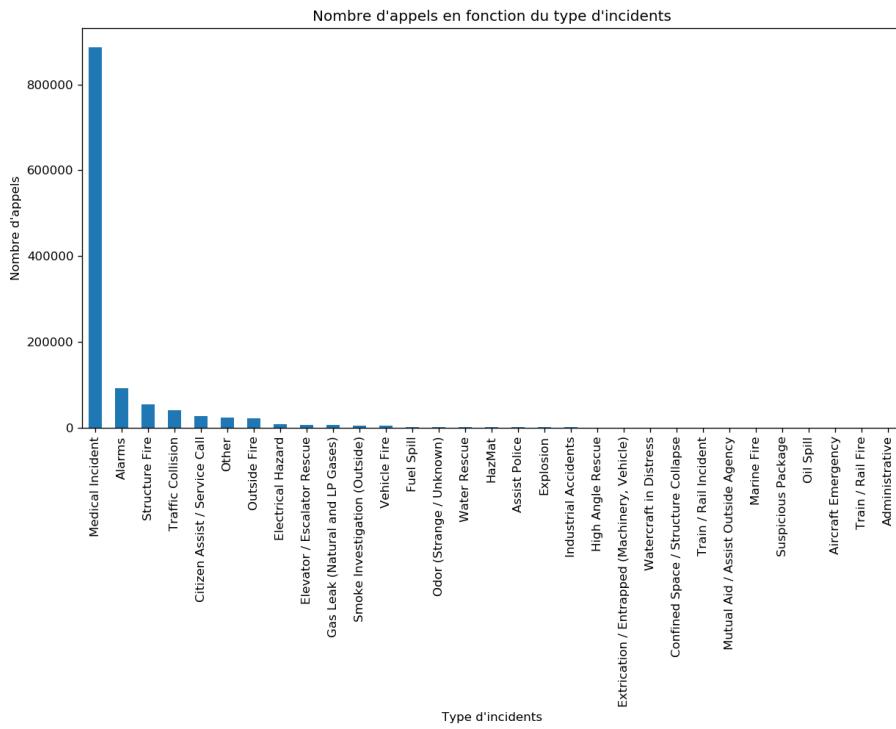


Figure 1: Nombre d'incidents en fonction du type d'incidents

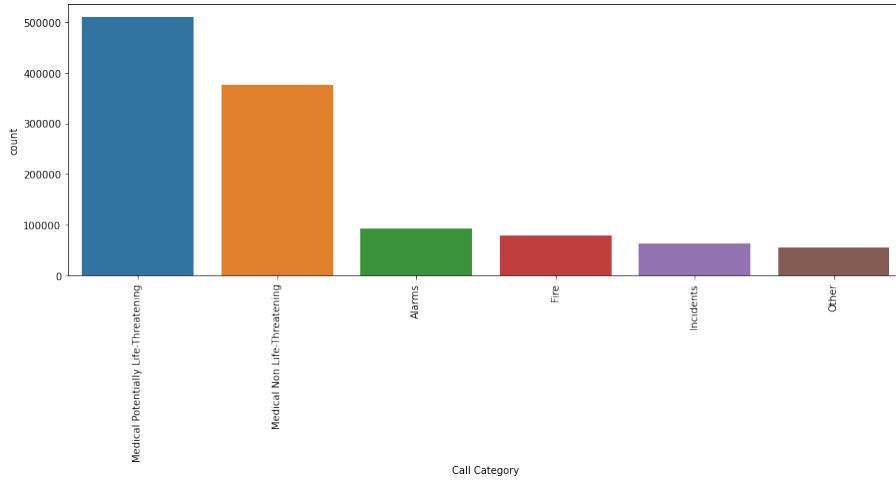


Figure 2: Nombre d'incidents en fonction de la catégorie d'incidents

3 Visualisation des données

Dans cette section, nous représentons les données afin de les visualiser, d'analyser les différents croisements et d'identifier les tendances et les effets potentiels. Cette démarche nous permettra de décider quelles données retirer ou conserver et de choisir les algorithmes qui leurs sont appropriés. Nous avons restreint notre analyse descriptive aux visualisations que nous avons jugé les plus pertinentes, toutefois, un plus grand éventail de graphiques existe dans le notebook (cf partie 1).

3.1 Distribution des modalités des variables temporelles

Les distributions des nombres d'appels par heure de chaque catégorie d'appels sont représentées à la figure 3. Malgré une tendance quasi-normale, qu'on perçoit distinctement dans les six graphiques, autour du milieu de la journée, il existe des différences substantielles dans les occurrences des différentes catégories d'urgences. La modalité la plus fréquente de la catégorie 'Alarms' est 10h tandis que celle de 'Fire' et de Incidents est à 17h. Quand aux incidents à caractère médical, la modalité la plus fréquente de la catégorie 'Medical Potentially Life-Threatening' est 13h et celle de 'Medical Non Life-Threatening' est 16h. Ainsi nous disposons d'une première idée visuelle sur la mobilisation des différents services d'urgence par heure à San Francisco.

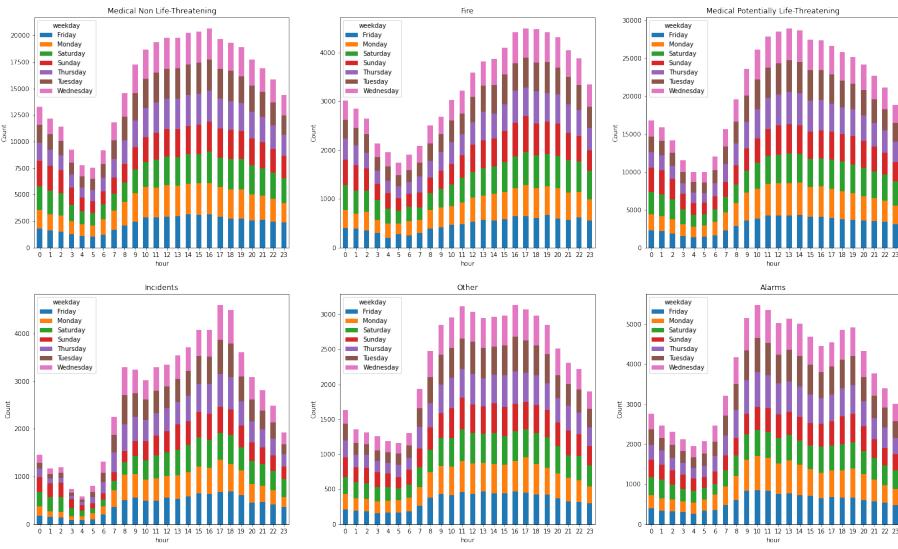


Figure 3: Catégories d'incidents en fonction de l'heure

A la figure 4 sont représentées les distributions des appels reçus par mois des différentes catégories d'incidents. L'effet du mois sur l'occurrence des incidents est notable mais moins significatif que

l'effet de la variable heure. Le pic en juillet de la catégorie 'Fire' n'est pas surprenant comparé aux incidents relativement élevés au moins de janvier, toutes catégories confondues.

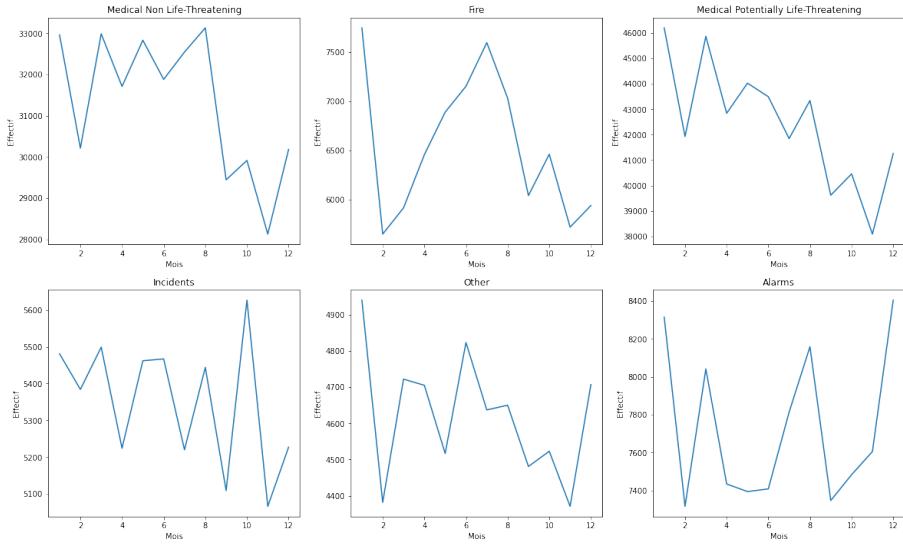


Figure 4: Catégories d'incidents en fonction du mois

A la figure 5, nous observons les distributions des appels d'urgence par jour du mois, des différentes catégories d'incidents. On remarque que les fluctuations le long du mois sont légères, ainsi, le jour du mois ne semble pas exercer un effet remarquable sur la survenance des appels d'urgences. Nous notons que les déclins observés le 31 du mois sont dues au fait que la moitié des mois ne durent que 30 jours.

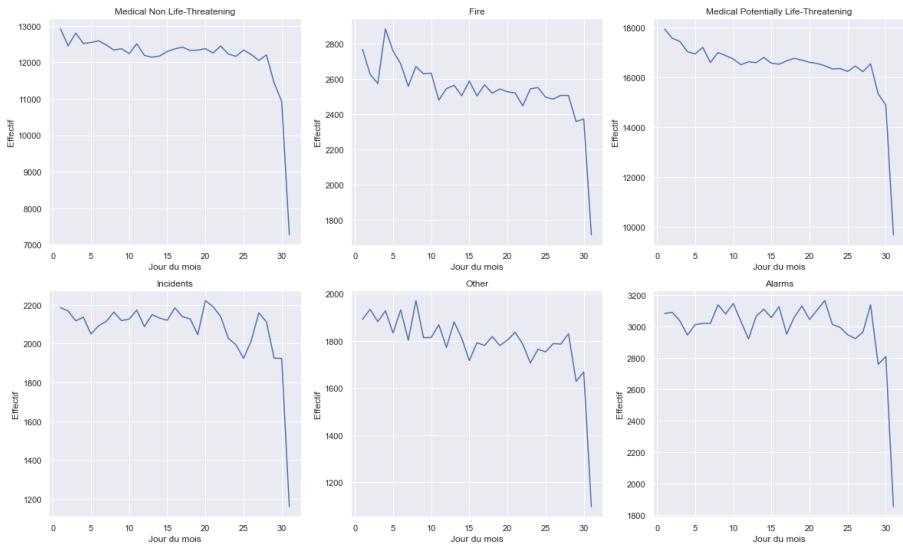


Figure 5: Catégories d'incidents en fonction du jour du mois

3.2 Distribution des modalités des variables géographiques

La distribution des appels d'urgence par quartier est représentée à la figure 7. Nous observons une grande disparité entre les quartiers, le nombre d'urgences dans les quatre quartiers les plus affectés représentent 40% de la totalité des appels qui ont eu lieu à San Francisco dans 41 différents quartiers.

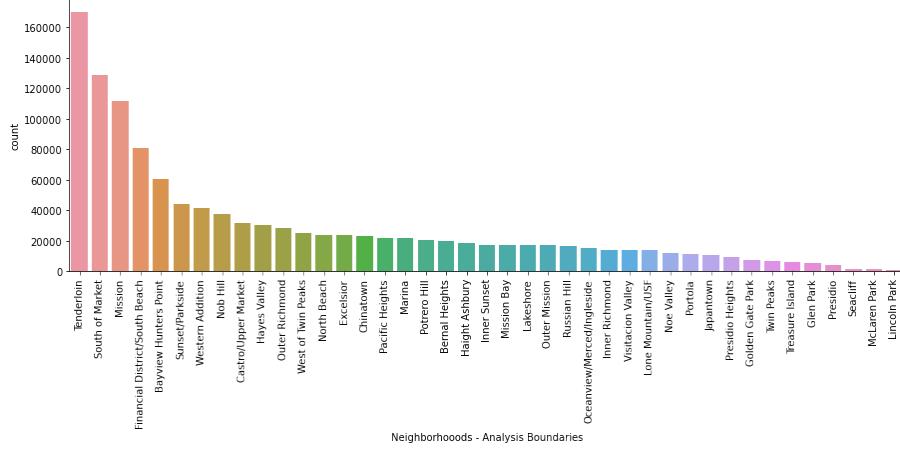


Figure 6: Nombre d'incidents en fonction du quartier

La distribution du nombre d'incidents par catégorie dans le quartier où il y a eu le plus d'appels, Tenderloin, est représentée à la figure 8. La distribution du nombre d'incidents par catégorie dans le quartier où il y a eu le moins d'appels, Lincoln Park, est représentée à la figure 9. Nous remarquons que malgré que la catégorie d'urgence prédominante dans les deux quartiers est 'Medical Potentially Life-Threatening', la différence des natures d'urgence entre les deux quartiers est notable.

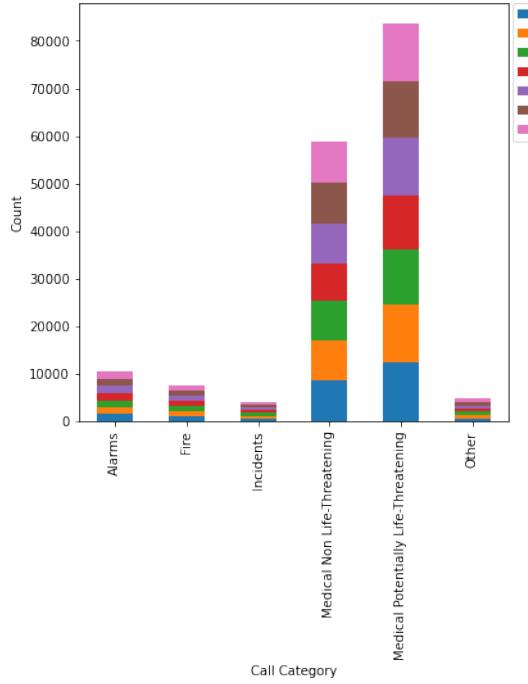


Figure 7: Incidents à Tenderloin

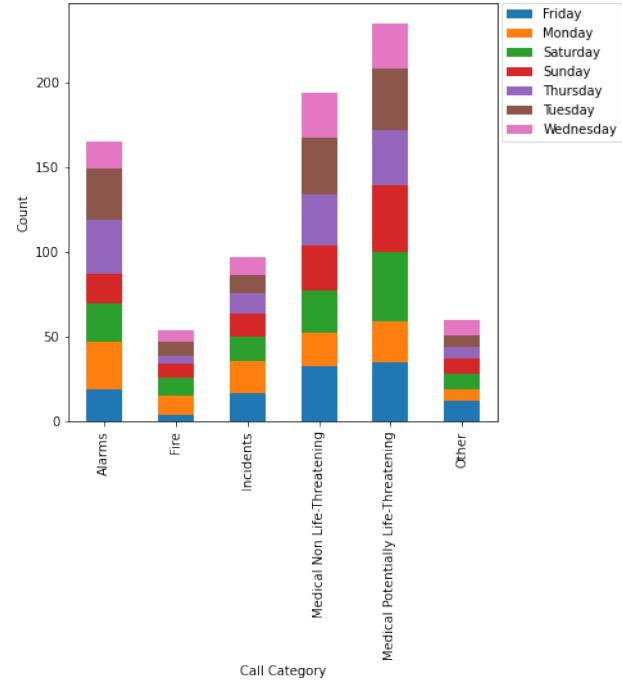


Figure 8: Incidents à Lincoln Park

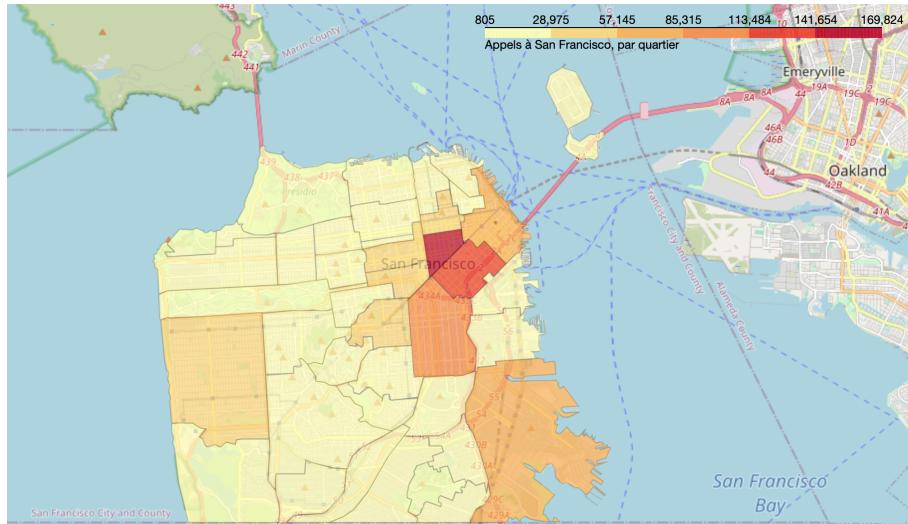


Figure 9: Cartographie du nombre d'incidents en fonction du quartier

A la figure 9, nous observons une cartographie du nombre d'incidents en fonction du quartier. Dans les figures 10 à 15, nous observons les cartographies par nombre d'appels pour chaque type d'incidents, des différences fondamentales sont notées quant à la dispersion des incidents par catégorie à San Francisco.

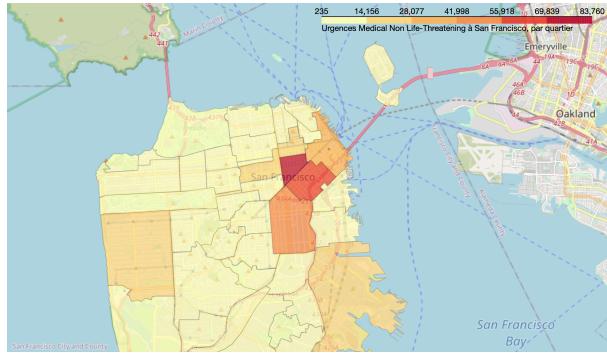


Figure 10: Cartographie de Medical Non Life-Threatening

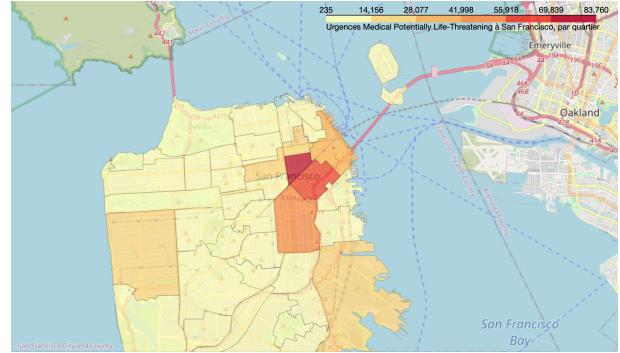


Figure 11: Cartographie de Medical Potentially Life-Threatening

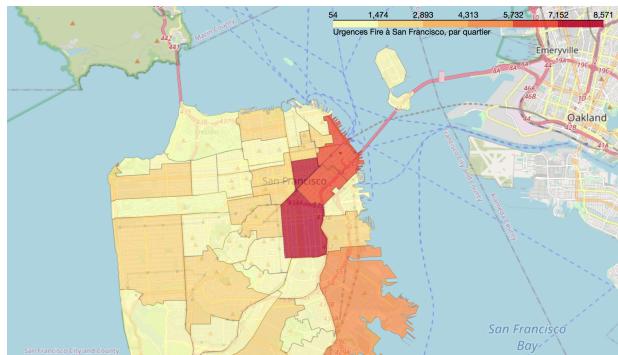


Figure 12: Cartographie de Fire

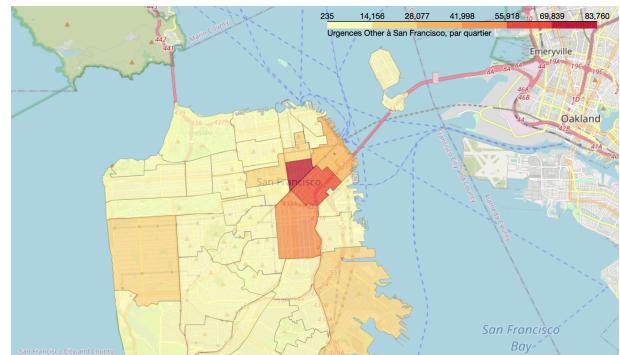


Figure 13: Cartographie de Other

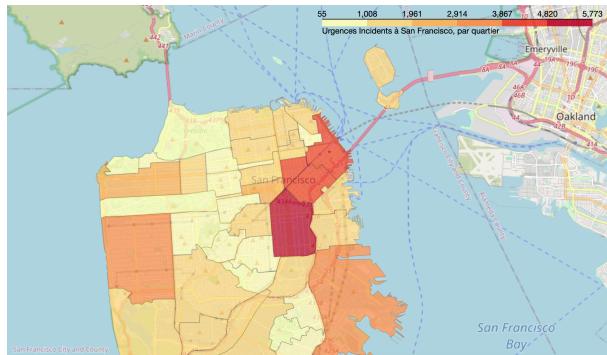


Figure 14: Cartographie de Incidents

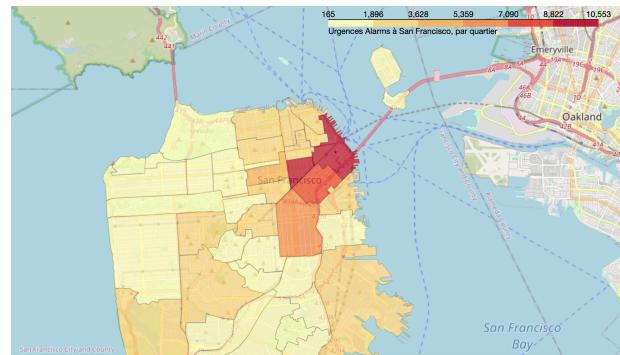


Figure 15: Cartographie de Alarms

4 Prédiction nombre d'appels à San Francisco

4.1 Prédiction du nombre d'appels par heure et par quartier à San Francisco

Notre objectif est de prédire le nombre d'appels par quartier et par heure afin de donner une prévision du nombre d'appels au centre d'appels d'urgence de San Francisco. Dans notre base de

données, nous avions des heures manquantes (celles durant lesquelles aucun appel n'a été émis) que nous avions ajouté afin que toutes les heures soient représentées.

Dans un premier temps, nous allons utiliser les variables qui semblent pertinentes d'après les analyses de visualisation le mois, la date, l'heure, le jour et enfin le quartier pour essayer de prédire le nombre d'appel dans l'heure demandée. Pour cela, nous allons utiliser des techniques de Machine Learning (Random Forest et KNN) et de Deep Learning (Réseau de Neurones) afin de détecter des patterns et confirmer les premières analyses descriptives faites.

Les trois modèles ont été utilisé pour de la classification puisque nous avons considéré 5 classes:

- la classe 0: 0 appel recensé durant l'heure pour le quartier
- la classe 1: 1 appel recensé durant l'heure pour le quartier
- la classe 2: 2 appels recensés durant l'heure pour le quartier
- la classe 3: 3 appels recensés durant l'heure pour le quartier
- la classe 4: 4 appels ou plus recensés durant l'heure pour le quartier

Pour des raisons de taille importantes de notre base de données, nous avons utilisé uniquement les dates du 01/01/2017 au 31/12/2019. Nous pouvons alors observer les résultats suivants:

Table 1: Modèles et Performance

Modèles	Accuracy	F1-score	F1-score par classe
Réseau de Neurones	0.7283	0.6138	(0.8428, 0, 0, 0, 0)
KNN	0.7123	0.6477	(0.8413, 0.1469, 0.0796, 0.0658, 0.149)
Random Forest	0.706	0.6674	(0.847, 0.197, 0.143, 0.136, 0.246)

Notre meilleur modèle en terme d'accuracy est le RN mais en regardant de plus près (cf notebook partie 2), notre modèle ne fait que prédire 0 et comme notre output contient 73% de 0 c'est pour cela que nous avons une accuracy correcte mais qui ne veut rien dire. Ainsi, nous préférions regarder le F1-score (ou la matrice de confusion) qui est plus représentatif de la performance de notre modèle. Par conséquent, le RF est le plus performant mais il n'est pas très performant pour prédire les classes autre que 0 comme nous pouvons le voir dans le tableau (le vecteur F1-score par classe).

Pour améliorer notre modèle, nous allons partir sur le RF mais en y incluant d'autres variables explicatives du nombre d'appels la météo de San Francisco pour chaque heure.

4.2 Prédiction du nombre d'appels à San Francisco en ajoutant d'autres variables

Dans la perspective de l'amélioration de la performance de nos modèles, nous avons élargi notre base de données en incorporant des features additionnelles.

4.2.1 Prédiction avec la météo de San Francisco heure par heure

Nous avons trouvé une base de données de la météo heure par heure de San Francisco que nous allons combiner avec notre base de données existante. Etant donné que la base de donnée liée à la météo n'est pas complète, nous avons utilisé les années 2014 à 2016 pour enrichir notre modèle.

Notre modèle s'est légèrement amélioré ($F1\text{-score} = 0.6868$) mais nous avons toujours des difficultés à prédire les classes autre que 0 (donc au moins un appel) comme le montre le $F1\text{-score}$ par composante : [0.858 , 0.186 , 0.1453, 0.1074, 0.1623].

Ainsi, il semblerait que la météo n'ait pas un grand impact sur le nombre d'appels d'urgence. Nous allons essayer une autre variable: les jours fériés. En effet les jours fériés peuvent avoir un impact (par exemple le 1 janvier beaucoup de personnes font la fête ce qui peut engendrer plus de risques et par conséquent plus d'appels au centre).

4.2.2 Prédiction avec les jours fériés de San Francisco heure par heure

Pour des raisons encore une fois de données, nous limitons notre base de données aux années 2017 à 2019.

Notre modèle ne s'est pas amélioré ($F1\text{-score} = 0.674$) mais il semble prédire un peu mieux les classes avec au moins un appel (classes 1 à 4) puisque que le $F1\text{-score}$ par composante est: [0.851, 0.216, 0.1509, 0.129, 0.262].

Le fait de restreindre notre base de données à 3 années d'historiques peut avoir un impact sur notre modèle puisque cela fait seulement 30 données de jours fériés. Même en ayant une base de données plus consistante, notre modèle ne pourrait toujours pas être fiable.

4.3 Prédiction multiple du nombre d'appels par heure et par catégorie à San Francisco

Dans cette partie, nous prédisons le nombre d'appels par quartier, par heure et respectivement pour chacune des catégories d'appels Alarms, Fire, Incidents, Medical Non Life-Threatening, Medical Potentially Life-Threatening et Other. Chaque label est donc représenté par un vecteur de longueur 6. Nos variables features sont month, day, hour, Neighborhoods - Analysis Boundaries(quartier) et le jour de la semaine weekday. Dans notre base de données, nous avions des heures manquantes (celles durant lesquelles aucun appel n'a été émis) que nous avions ajouté afin que toutes les heures soient représentées.

Nous avons utilisé trois modèles avec la fonction de perte par défaut ainsi qu'avec la fonction de perte mean squared error pour faire nos prédictions (cf notebook partie 3) et les performances de chaque modèle sont rappelées dans Table 2. L'accuracy représente la proportion de bonnes prédictions sur les données de test et L'accuracy par composante représente la proportion de bonnes prédictions sur les données de test pour chaque catégorie d'appel.

Nous avons évalué nos résultats uniquement sur les données de l'année 2019 à cause de la taille importante de notre base. Le modèle qui nous fournit les meilleures performances est la *Regression Logistique*. Les taux de bonnes prédictions sont meilleures pour les catégories d'appel autres que Medical Non Life-Threatening et Medical Potentially Life-Threatening car les appels pour ces catégories ne sont pas aussi récurrents que ceux des deux catégories prédominantes suscitées. Cependant, l'accuracy multiple (qui consiste à l'évaluation de l'ensemble du vecteur) n'est pas élevée (0.724) et ceci est du essentiellement au fait que le modèle multioutput prédit chaque composante du label d'une manière indépendante (il prédit chaque colonne et ensuite il effectue la concaténation de toutes les prédictions).

Table 2: Modèles et Performance

Modèles	fonction de perte	Accuracy	Accuracy par composante
KNN	perte par défaut	0.348	(0.889, 0.922, 0.914, 0.647, 0.600, 0.928)
	Mean squared error	0.575	(0.939, 0.961, 0.955, 0.807, 0.776, 0.963)
Random Forest	perte 0/1	0.711	(0.964, 0.978, 0.975, 0.872, 0.849, 0.98)
	Mean squared error	0.701	(0.963, 0.977, 0.973, 0.867, 0.843, 0.979)
Regression Logistique	perte 0/1	0.724	(0.966, 0.979, 0.976, 0.882, 0.858, 0.981)
	Mean squared error	0.724	(0.966, 0.979, 0.976, 0.882, 0.858, 0.981)

5 Conclusion et perspectives

L'objectif ce travail était de prédire le nombre d'appels de San Francisco afin de permettre une meilleure allocation des ressources au sein des services responsables de porter secours à la population. La première étape était d'organiser notre bases de données à travers des opérations préliminaires. La seconde étape, nous avons exploré et visualisé les données temporelles et géographiques afin de nous approprier les données de notre base et de déceler les variables pertinentes à notre problématique. Nous avons conclus que les variables pertinentes à retenir étaient la catégorie d'appel (une variable que nous avons créé qui résume le type d'appel ainsi que la catégorie du type d'appel) le mois, le jour, le jour de la semaine, l'heure et le quartier. A la troisième étape, nous avons prédit le nombre d'appels en utilisant plusieurs modèles de Machine Learning (KNN, Random Forest, Regression Logistique, Réseaux Neurones) mais ceux-ci ne sont pas en général très performants et ceci malgré que nous avons ajouté à nos données d'autres features (météo, jours fériés) que nous avons jugé avoir un impact sur le nombre d'appels.

Dans une perspective d'améliorer la performance de notre modèle, nous suggérons l'ajout d'autres features telles que les dates des festivités sportives (basket et football américain) importantes à San Francisco, l'âge des personnes prises en charge pour les incidents médicaux. Ces features pourraient ajouter des informations pertinentes qui contribuerait à l'explication de l'occurrence des urgences.