UNIVERSITÉ
CÔTE D'AZUR

BUT SD - $3^{ème}$ année

Rapport projet Data Mining

Triage des patients aux urgences

Algassimou Diallo

Décembre 2023

SAE: Mise en oeuvre d'un processus de Datamining
Enseignant : Celia Da Costa Pereira

Contents

1 Introduction

La gestion efficace du triage des patients aux urgences demeure un défi central dans l'amélioration des systèmes de santé actuels, cherchant à optimiser les ressources médicales en fonction de la gravité des cas.
Notre projet, influencé par l'article scientifique : "Predicting hospital admission at emergency department triage using machine learning" de Woo Suk Hong, Adrian Haimovich, et R. Andrew Taylor, se positionne comme une analyse approfondie du triage hospitalier aux urgences. En tirant profit de leurs résultats et leur dataset, notre objectif principal est d'effectuer des analyses pertinentes sur le triage aux urgences et élaborer un modèle de machine learning robuste, visant à améliorer la réactivité et l'efficacité des services d'urgence.

Ce rapport détaille de manière exhaustive le déroulement de notre projet, depuis le prétraitement des données et l'analyse exploratoire des données jusqu'à l'implémentation des modèles de machine learning et à l'analyse approfondie des résultats obtenus.

2 Prise en main du dataset

La première étape de notre projet a consisté à prendre en main le jeu de données initialement fourni au format R.data, et à le convertir pour son utilisation dans l'environnement Python. Le dataset regroupe l'ensemble des visites aux services d'urgence pour des adultes (18-108 ans) de mars 2014 à juillet 2017. Ces visites proviennent d'un établissement académique et de deux urgences communautaires, se soldant soit par une admission soit par un renvoi. Chaque visite de patient a généré l'extraction de **972** variables, fournissant ainsi une richesse d'informations pour l'analyse et le développement de notre modèle.

Notre démarche a débuté par une exploration minutieuse de l'article, visant à comprendre en profondeur la nature des variables extraites, les critères d'admission ou de renvoi, ainsi que les résultats obtenus par les auteurs. Cette phase préliminaire, a été fondamentale pour orienter notre méthodologie pour ce projet. Face à la complexité du dataset, comprenant **plus de 900 variables et plus 500000 enregistrements par visite de patient**, nous avons pris une décision stratégique en vue de réaliser une analyse exploratoire efficace.

Nous avons opté pour une **division du dataset en catégories** celles présentées dans l'article. Ce choix a permis de cibler les dimensions importantes et de faciliter les analyses pour notre étude tout en éliminant le besoin d'analyser chaque variable individuellement.

Category	Number of Variables	Only Triage	Only History	Full
Response variable (Disposition)	1	X	X	X
Demographics	9	X	X	X
Triage evaluation	13	X		X
Chief complaint	200	X		X
Hospital usage statistic	4		X	X
Past medical history	281		X	X
Outpatient medications	48		X	X
Historical vitals	28		X	X
Historical labs	379		X	X
Imaging/EKG counts	9		X	X
Total	972	223	759	972

Les catégories privilégiées pour notre EDA sont :**Demographics, Triage evaluation, Chief complaint, Hospital usage statistic et Outpatient medications.**

3 Analyse exploratoire des données(EDA)

Suite à la division du dataset en catégories pertinentes, notre analyse exploratoire des données (EDA) a révélé des observations intéressantes dans chacune des catégories du dataset.

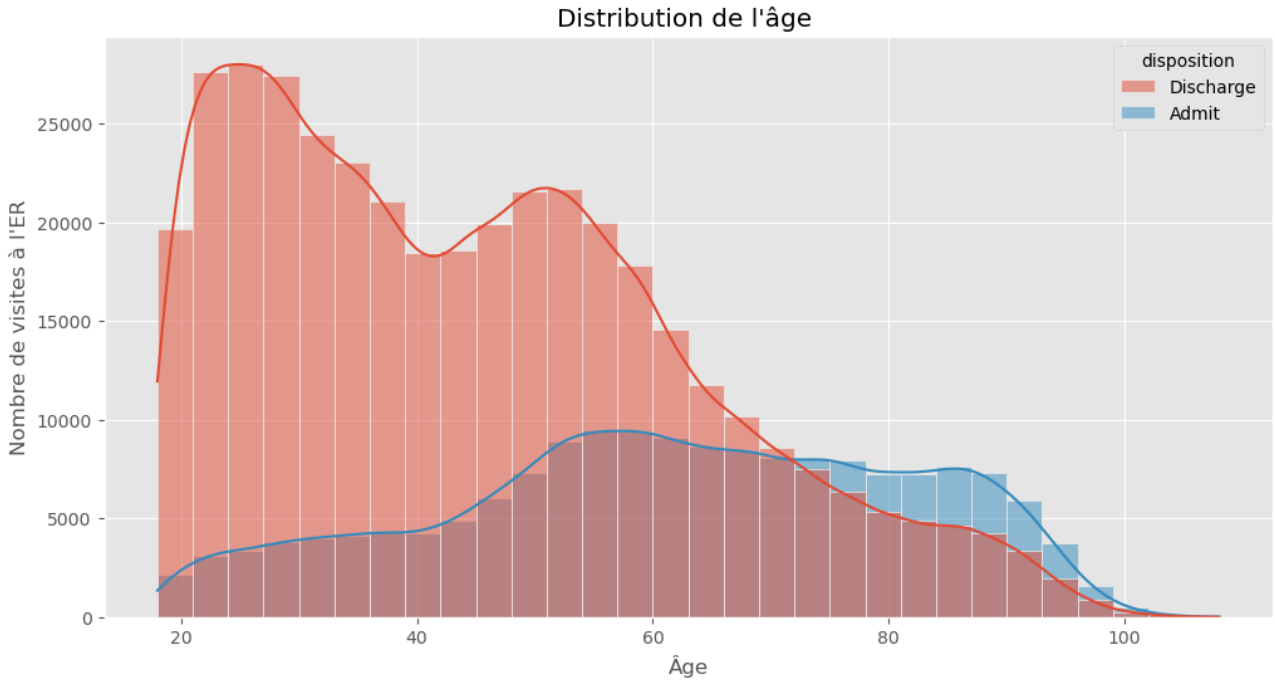
3.1 Demographics

"L'Age": :Nous avons examiné la relation entre l'âge des patients et la variable cible: la "disposition". En traçant la distribution de cette variable en fonction de l'âge, des différences significatives ont émergé entre les patients classés comme "discharge" et "admit".

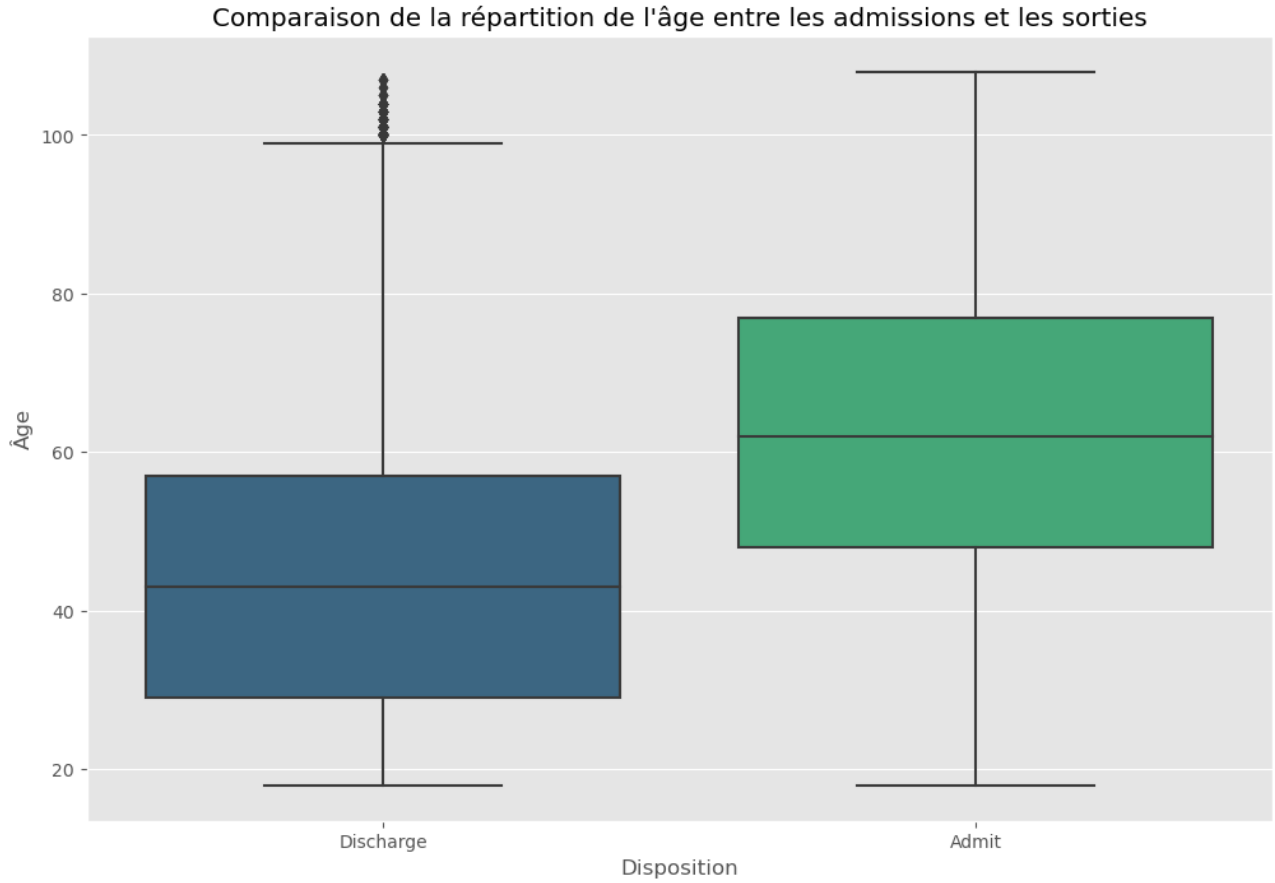
Pour les patients classés comme "discharge", nous avons observé une **distribution bimodale**. Deux pics distincts apparaissent lors de l'analyse en fonction de l'âge. Le premier pic, aux environs de **25 ans**, regroupe une population importante de plus de 25 000 patients. À cet âge, la distribution des patients "admit" est notablement basse, avec moins de 5 000 patients. Cette disparité souligne une tendance où **les patients plus jeunes sont fortement associés à la catégorie "discharge"**.

Une évolution significative se produit vers l'âge de **50 ans**. À ce stade, la distribution des patients "admit" atteint son sommet, regroupant près de 10 000 patients, et maintient une stabilité relative par la suite. En contraste, la distribution des patients "discharge" diminue de manière drastique, suggérant une transition vers une prédominance des admissions au-delà de cet âge.

La présence de ces tendances distinctes dans la distribution des patients "discharge" et "admit" en fonction de l'âge suggère fortement une corrélation significative entre ces deux variables. L'analyse exploratoire révèle que les patients plus jeunes, caractérisés par un premier pic aux environs de 25 ans, sont nettement associés à la catégorie "discharge". En revanche, la prévalence des admissions augmente de manière notable après l'âge de 50 ans, signalant une transition vers une population plus âgée plus susceptible d'être "admit". Cette observation initiale nous conduit à supposer que **l'âge joue un rôle crucial dans la détermination de la disposition des patients aux urgences.**



Pour une visualisation plus approfondie de la différence de distribution en fonction de l'âge entre les patients "discharge" et "admit", nous avons construit un boxplot. Ce graphique met en évidence de manière claire les variations significatives dans les tranches d'âge des deux groupes. En observant le boxplot, il devient manifeste que la population des patients "discharge" est nettement plus jeune que celle des patients "admit". Les boîtes délimitant le premier et le troisième quartile pour les "discharges" s'étendent vers des valeurs d'âge inférieures, tandis que pour les "admis", ces boîtes s'étendent vers des valeurs d'âge supérieures. Les médianes, représentées par les lignes à l'intérieur des boîtes, confirment cette tendance, soulignant une différence notable dans les âges médians des deux groupes. Ce boxplot renforce visuellement nos conclusions précédentes, confirmant que l'âge exerce une influence significative sur la disposition des patients aux urgences.



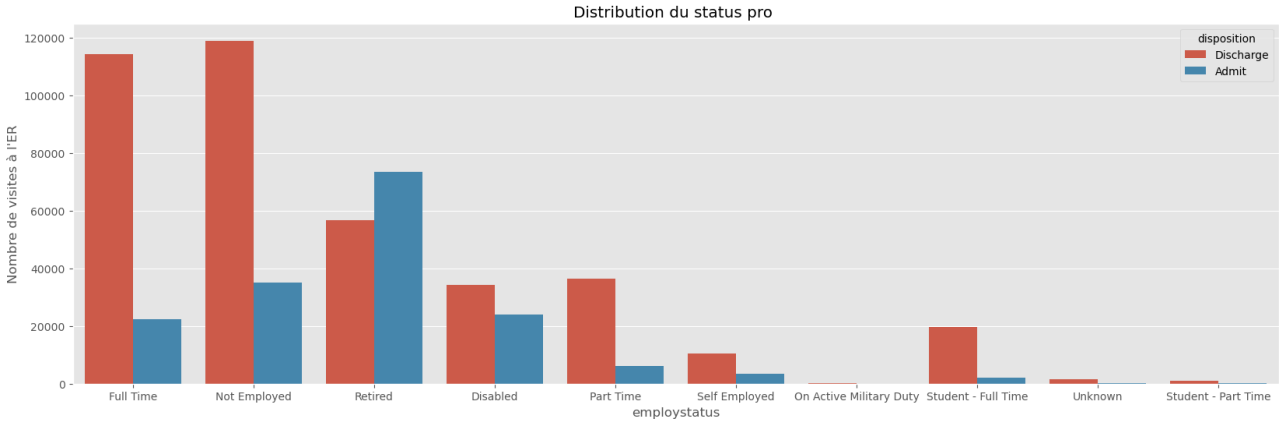
Pour consolider nos observations visuelles, nous avons entrepris un test statistique afin de vérifier la significativité de la différence d'âge entre les patients classés comme "admit" et "discharge". Nous avons opté pour **le test t de Student**, un test fréquemment utilisé pour comparer les moyennes de deux groupes. Ce test examine si la différence observée entre les moyennes est statistiquement significative ou si elle pourrait être attribuée au hasard.

Les résultats obtenus indiquent une **T-statistique de 294**, une valeur extrêmement élevée, et **une p-value très proche de zéro**. Dans le contexte de notre test d'hypothèse, la T-statistique représente l'écart entre les moyennes observées des deux groupes par rapport à ce que l'on pourrait attendre par simple variabilité aléatoire. Les hypothèses formulées sont les suivantes :

- Hypothèse nulle (H0) : Il n'y a aucune différence significative d'âge entre les admissions et les sorties à l'urgence.
- Hypothèse alternative (H1) : Il existe une différence significative d'âge entre les admissions et les sorties à l'urgence.

Au seuil de 1%, les résultats de notre test d'hypothèse renforcent de manière significative nos conclusions quant à l'influence de l'âge sur la disposition des patients aux urgences. **Avec une p-value pratiquement nulle, nous pouvons donc rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.** Ces résultats confirment qu'il existe une différence significative d'âge entre ces deux groupes de patiente et donc que l'âge est un facteur influençant énormément l'admission aux urgences. Cette concentration plus élevée d'admissions dans la population plus âgée est cohérente avec le fait que les individus plus âgés ont généralement un risque médical plus élevé, nécessitant une admission plus fréquente. Ainsi, l'association entre l'âge et la disposition des patients aux urgences reflète la réalité médicale où les jeunes, en règle générale, présentent des symptômes moins graves ou des problèmes de santé moins complexes, tandis que les personnes âgées sont plus susceptibles de nécessiter des soins hospitaliers approfondis.

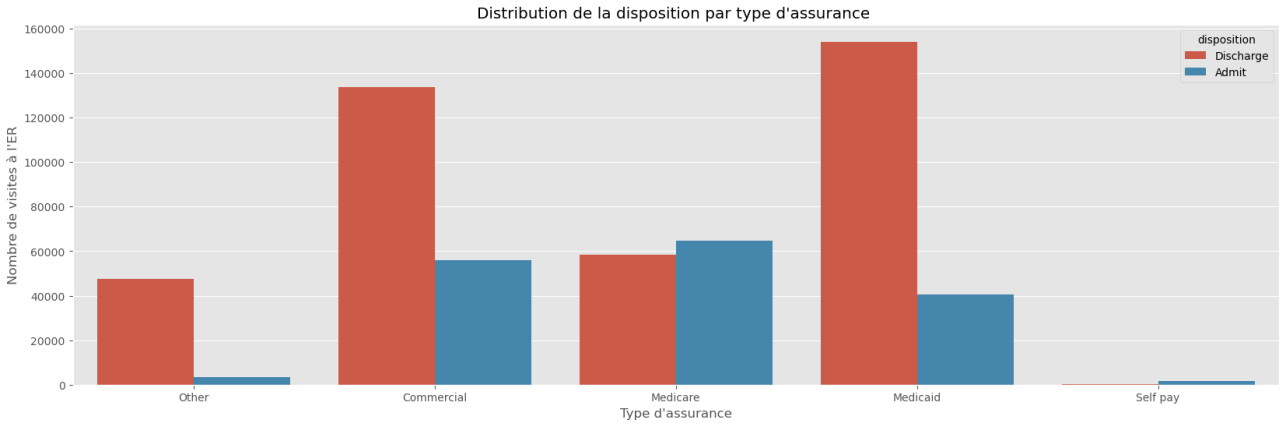
Notre exploration des données s'est étendue aux variables : **statut professionnel et type d'assurance des patients**, révélant des tendances significatives en termes de disposition aux urgences.
Il est intéressant de noter que la majorité des patients, indépendamment de leur statut professionnel, sont plus fréquemment classés comme "discharge" que "admit". Cependant, une exception notable émerge parmi les patients retraités, qui présentent une tendance inverse. Les données indiquent que les patients exerçant une activité professionnelle, qu'ils soient employés, indépendants, ou dans d'autres catégories actives, sont plus susceptibles d'être libérés après leur visite aux urgences. En revanche, les patients retraités, bien qu'ils constituent une proportion relativement faible de l'échantillon, sont plus fréquemment admis.



Quant à la relation entre le type d'assurance des patients et leur admission aux urgences, il est important de souligner que le système d'assurance médicale aux États-Unis est diversifié, avec différents programmes destinés à différentes catégories de la population.

Notre analyse a révélé des variations significatives dans la disposition en fonction du type d'assurance. Les patients bénéficiant de l'assurance **Medicaid** sont plus fréquemment classés comme "discharge", indiquant une tendance à une libération après la visite aux urgences. À l'inverse, ceux couverts par l'assurance **Medicare**, principalement destinée aux personnes âgées, sont plus souvent admis.

Ces résultats peuvent être expliqués par les caractéristiques propres à chaque type d'assurance. **L'assurance Medicaid** cible principalement les personnes à faible revenu, tandis que **l'assurance Medicare** est destinée aux personnes âgées de 65 ans et plus.



Dans l'exploration des variables telles que le "genre", la "race", "l'ethnicité" et la "religion", nos observations ne révèlent pas de tendances significatives ou d'associations marquées avec la disposition des patients aux urgences.
En définitive, notre analyse approfondie des facteurs démographiques influençant la disposition des patients aux urgences révèle des tendances significatives. L'âge émerge comme un élément déterminant, avec une propension plus élevée à l'admission pour les personnes âgées. Les facteurs socioéconomiques, tels que le statut professionnel et le type d'assurance, influencent également les décisions de triage, mettant en lumière l'impact des réalités financières sur le parcours des patients.

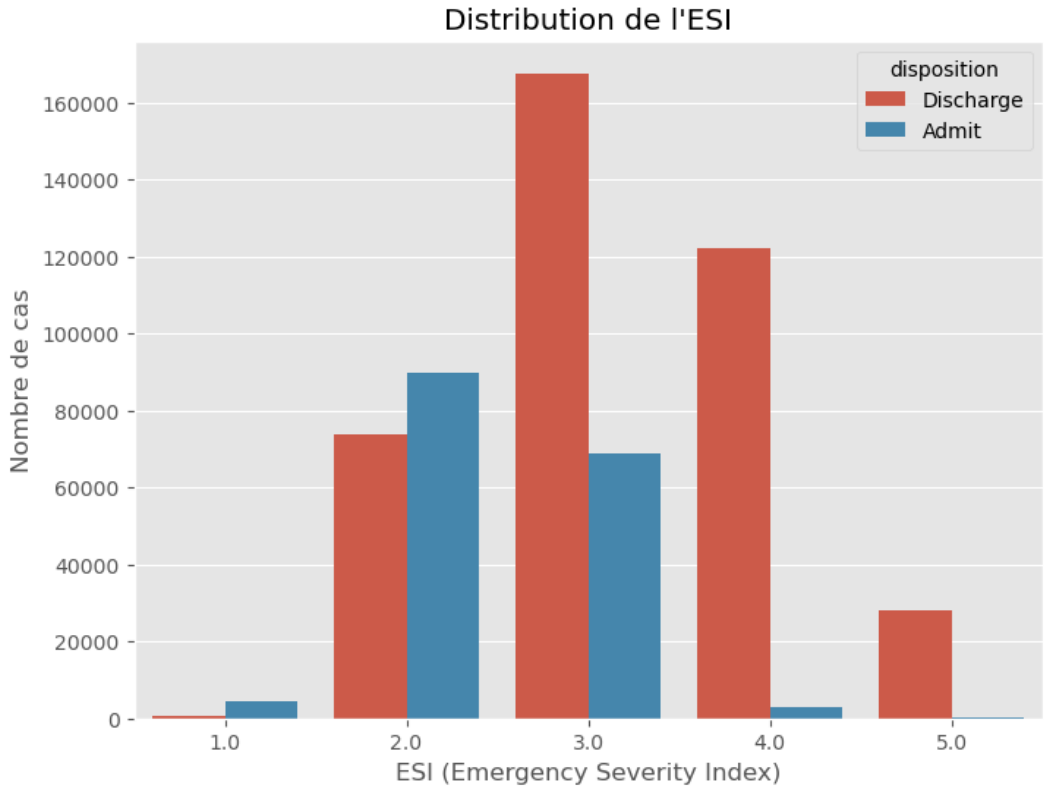
3.2 Triage Evaluation

Après avoir analyser les variables démographiques nous nous sommes intéressés au variables liées au triage à savoir les variables représentant les informations recueillises par le personnel médicale lors de l'arrivé du patient et le Emergency Severity Index(ESI) attribué.

L'Emergency Severity Index (ESI) est un outil de triage largement utilisé pour évaluer la gravité des cas aux services d'urgence. Il catégorise les patients en cinq niveaux (ESI 1 à ESI 5) en fonction de la sévérité de leur état de santé, permettant ainsi une allocation efficace des ressources médicales en fonction des besoins immédiats.

Dans notre analyse exploratoire, nous avons examiné la relation entre l'ESI et la variable cible, à savoir la disposition des patients aux urgences. Les résultats, sans surprise, mettent en évidence des tendances significatives. Les patients classés avec un ESI de 1 ou 2 sont plus fréquemment admis, reflétant la sévérité élevée de leurs conditions médicales. Ces cas nécessitent généralement une intervention médicale immédiate, ce qui se traduit par une faible probabilité d'être libérés après la visite aux urgences.

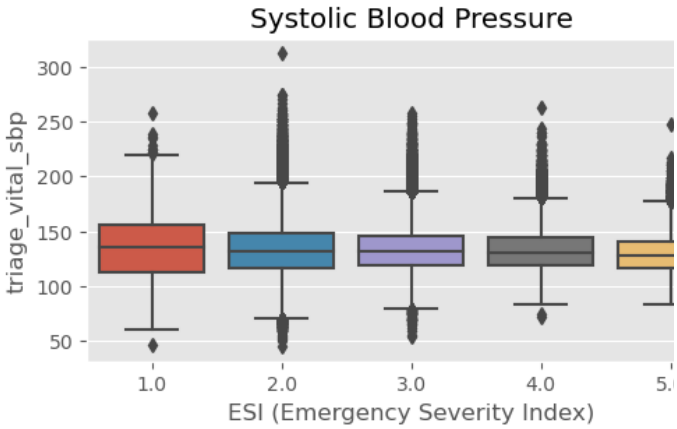
D'autre part, les patients avec un ESI de 4 ou 5, catégorisés comme moins graves, sont plus souvent libérés, soulignant la prédominance des cas moins urgents dans ces catégories. Cette observation confirme la capacité de l'ESI à trier efficacement les patients en fonction de la gravité de leurs conditions. Cette première analyse établit une corrélation claire entre l'ESI et la disposition des patients, démontrant l'utilité de cet indice de triage dans la prise de décision rapide et précise aux urgences.

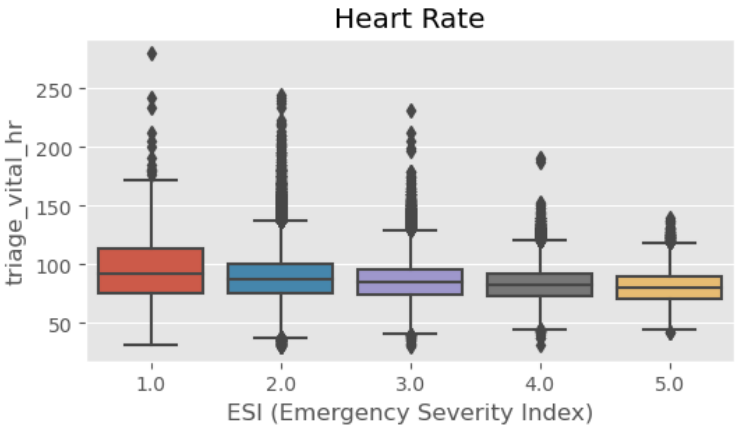


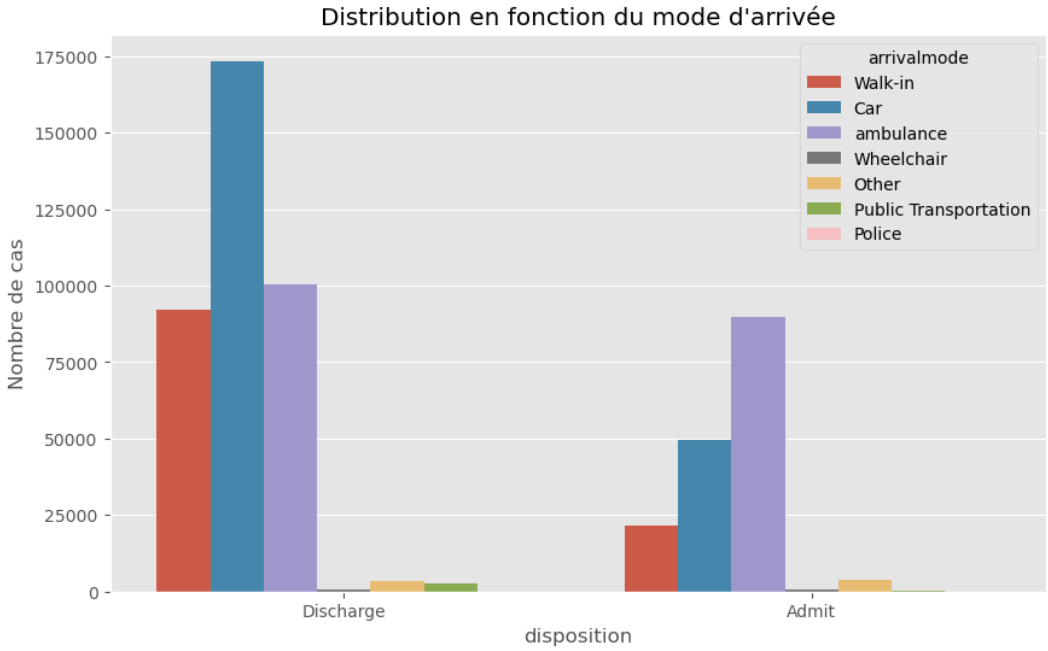
L'analyse détaillée des relations entre l'Emergency Severity Index (ESI) et diverses autres variables, telles que l'âge, les services d'urgence, le mode d'arrivée, et les signes vitaux, offre des éclairages supplémentaires sur la processus de triage aux urgences.

En ce qui concerne l'âge, nos résultats confirment l'intuition que les ESI les plus graves sont associés aux groupes de personnes âgées, reflétant la fréquence accrue de conditions médicales graves dans cette population.

Les variations observées dans les services d'urgence (A, B, C) et les modes d'arrivée (voiture, ambulance, police, etc.) soulignent l'impact de la rapidité d'intervention. Les patients arrivant en ambulance, indiquant souvent des situations médicales plus urgentes, sont plus fréquemment classés comme "admit", tandis que ceux arrivant en voiture sont plus souvent libérés, soulignant des cas potentiellement moins urgents. Cependant, la révélation la plus notable se situe dans la corrélation significative entre les signes vitaux (pressions diastoliques et systoliques, fréquence cardiaque) et l'ESI. Les valeurs plus élevées de ces signes vitaux sont associées à des ESI plus graves, indiquant une corrélation directe entre la sévérité des signes vitaux et la gravité perçue de l'état du patient.



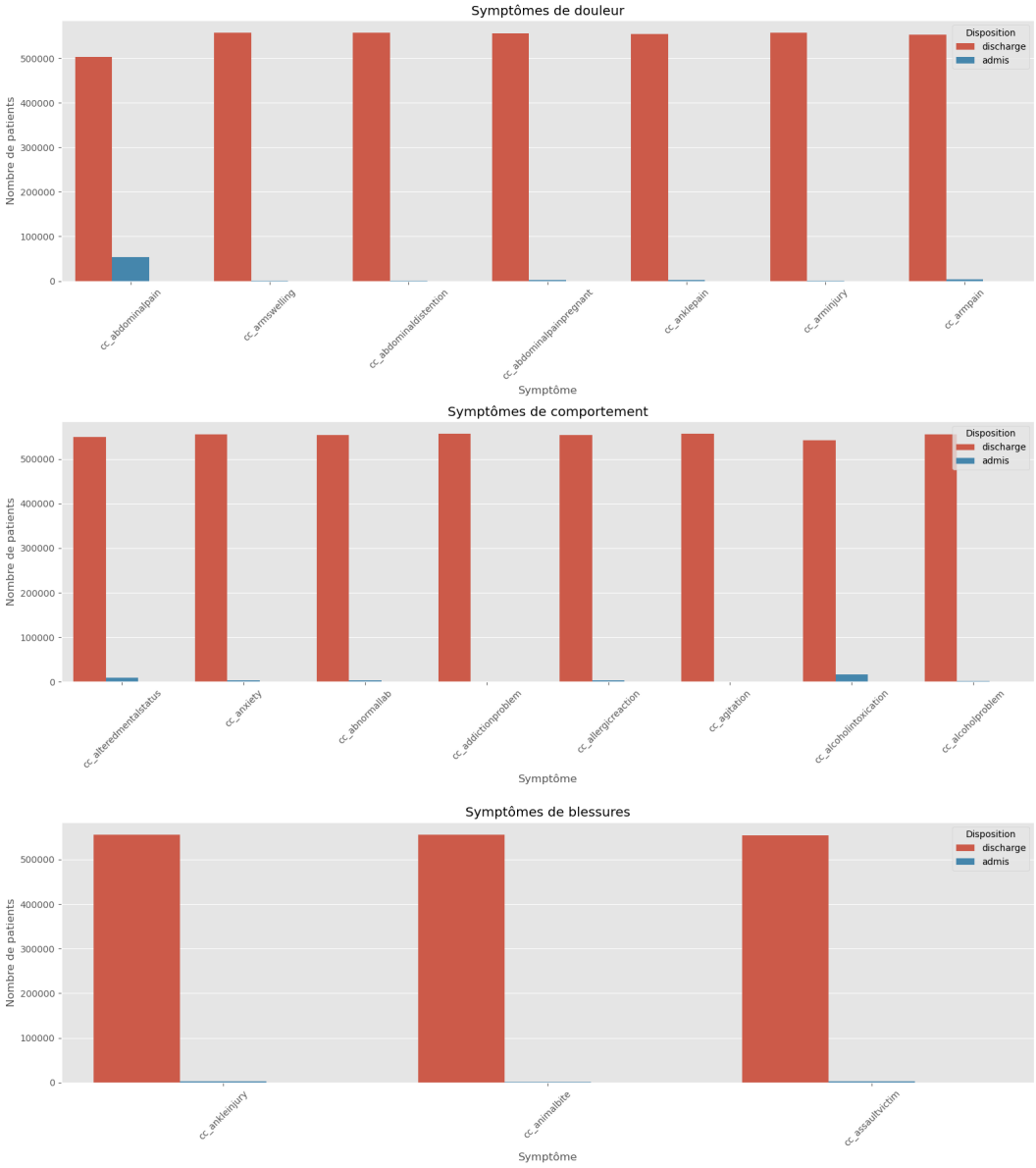




3.3 Chief Complaint

L'analyse des variables du chief complaint, représentant les symptômes des patients, soulève des observations intéressantes sur la relation entre les motifs de consultation et la disposition finale aux urgences.

Dans l'ensemble, la plupart des symptômes ne semblent pas exercer une influence significative sur l'admission, avec une tendance générale à la libération ("discharge"). Cependant, quelques symptômes spécifiques, tels que "abdominal pain", "altered mental status", et "alcohol intoxication", se démarquent en étant associés plus fréquemment à des cas d'admission.

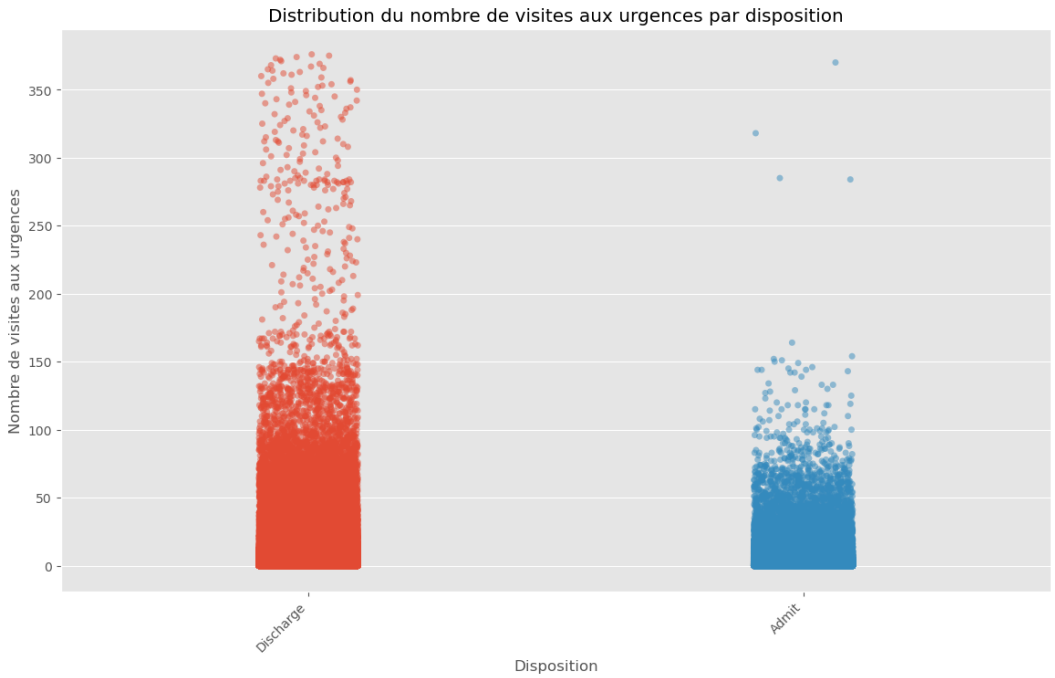


3.4 Patient Hospital History

Nous avons ensuite entrepris une analyse approfondie de l'historique médical des patients en explorant plusieurs variables clés. Ces données révélatrices comprennent la disposition du patient lors de la dernière visite ("previous-dispo"), le nombre total de visites aux urgences ("n_edvisits"), le nombre d'admissions antérieures ("n_admissions"), et le nombre de chirurgies précédentes ("n_surgeries"). Ces informations offrent un éclairage précieux sur la fréquence des visites, les tendances en matière d'admissions, et les antécédents chirurgicaux des patients.

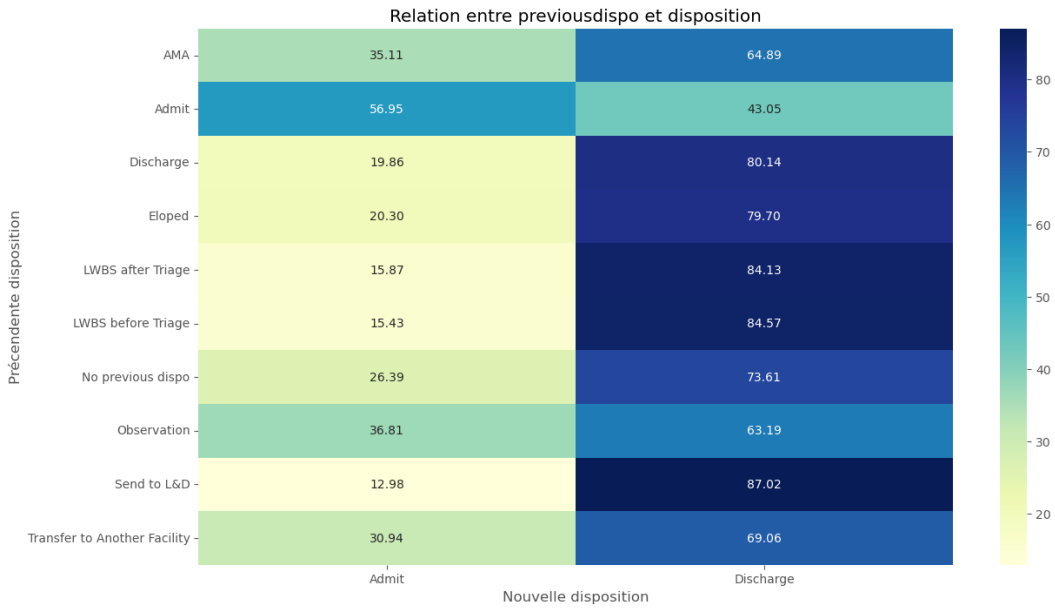
Cette exploration approfondie nous a conduit à des observations particulièrement intéressantes qui apportent des nuances significatives à notre compréhension des parcours médicaux individuels.

L'analyse de la variable "n_edvisits", représentant le nombre total de visites aux urgences, apporte une observation significative. Nous constatons une corrélation inverse entre le nombre de visites et la disposition actuelle du patient. En d'autres termes, plus un patient a fréquemment visité les urgences par le passé, plus il est susceptible d'être classé en "discharge" lors de la visite actuelle. En revanche, les patients qui sont admis ont tendance à avoir un historique de visites aux urgences moins fréquent.



La variable "previousdispo", qui indique la disposition du patient lors de sa dernière visite aux urgences, a révélé des informations particulièrement riches et pertinentes. En construisant un heatmap, nous avons pu visualiser la relation entre la disposition précédente du patient et sa nouvelle disposition. Le heatmap met en évidence une observation cruciale : la disposition antérieure du patient exerce une influence significative sur sa nouvelle disposition aux urgences. Plus spécifiquement, pour les patients dont la dernière disposition était "admit", 57% d'entre eux sont à nouveau admis lors de leur visite actuelle. Cette constatation suggère une tendance à la récurrence des admissions pour les individus ayant déjà nécessité une hospitalisation antérieure.
Tandisque la majorité des patients ayant une disposition antérieure autre que "admit" ont une chance élevée d'être "discharge" lors de leur visite actuelle, dépassant souvent 70% d'entre eux.

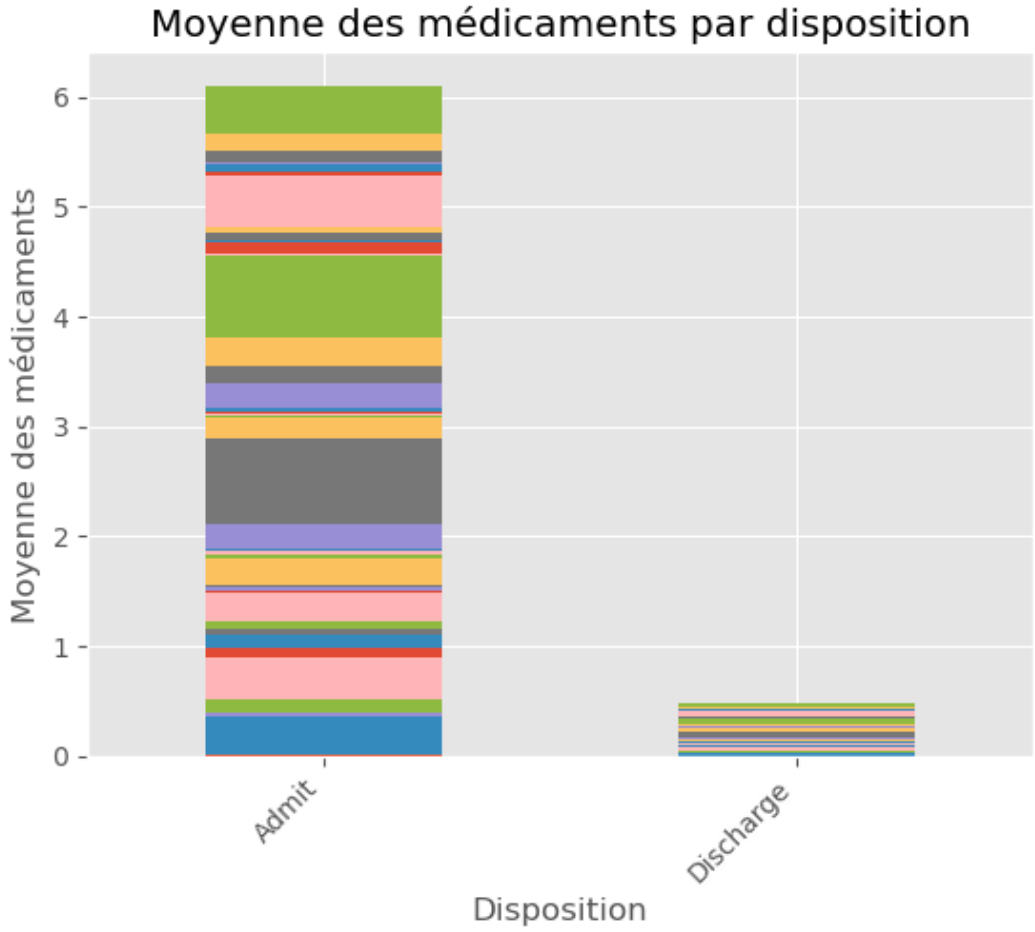
Cependant, certaines dispositions antérieures présentent des caractéristiques particulières. Pour les patients ayant une disposition antérieure "AMA" (Against Medical Advice, c'est-à-dire que le patient choisit de partir contre les recommandations et conseils du personnel médical.), "transfer to another facility", et "observation", la probabilité de la nouvelle admission est plus élevée par rapport aux autres dispositions antérieures. Ces catégories montrent des pourcentages d'admission de 30% indiquant que ces situations spécifiques peuvent augmenter la probabilité d'une nouvelle admission aux urgences.



3.5 Oupatient Medications

L'analyse des variables liées aux "outpatient medications", représentant le nombre de médicaments utilisés par les patients, révèle des observations significatives. Une corrélation notable se dégage entre le nombre de médicaments et la disposition finale du patient aux urgences.

En effet, les patients classés en "admit" ont tendance à consommer un nombre significativement plus élevé de médicaments par rapport aux patients libérés ("discharge"). Cette observation indique une association entre une polypharmacie, c'est-à-dire la prise de plusieurs médicaments, et une probabilité plus élevée d'admission aux urgences.



En conclusion de l'EDA approfondie du processus de triage aux urgences, chaque facette explorée a contribué à affiner notre compréhension des facteurs influant sur la disposition des patients. L'intégration de variables telles que les variables démographiques(age,statut pro,..) l'Emergency Severity Index (ESI), l'historique médical, les symptômes, et la médication des patients a permis de dégager des tendances significatives.
L'âge a émergé comme un facteur crucial influençant la disposition des patients aux urgences. La distribution observée a souligné une corrélation significative entre l'âge du patient et sa probabilité d'admission.
De plus l'ESI s'est avéré être un outil crucial dans la stratification des patients en fonction de la sévérité de leurs conditions médicales, tandis que l'historique médical, en particulier la disposition antérieure et la fréquence des visites, a révélé des liens forts avec la disposition actuelle. Les symptômes des patients ont apporté des nuances importantes, soulignant que certains motifs de consultation sont plus susceptibles d'entraîner des admissions. Et enfin l'analyse des médicaments a mis en lumière l'impact de la polypharmacie sur la probabilité d'admission.
Ces analyses et conclusions nous ont énormément aidé et influencé dans nos décisions lors du prétraitement des données et du choix des modèles.

4 Prétraitement des données

Suite à notre Analyse Exploratoire des Données (EDA), nous avons entamé le prétraitement des données en considérant attentivement nos découvertes de l'EDA. En examinant le dataset, nous avons constaté un nombre important de valeurs manquantes dans diverses catégories telles que les données démographiques, les symptômes et l'Emergency Severity Index (ESI).

Guidés par les constatations de l'EDA, où nous avons identifié l'influence significative de chaque variable manquante sur notre variable cible (par exemple, l'âge et l'ESI), **nous avons décidé de ne pas supprimer ou remplacer simplement ces valeurs par des moyennes ou médianes**. Une telle approche aurait été trop simpliste et risquée pour la qualité de notre modèle.

Nous avons plutôt opté pour une méthode plus nuancée :**l'imputation KNN (K-Nearest Neighbors)**.
L'imputation KNN remplit les valeurs manquantes en se basant sur la similarité entre observations. Pour chaque donnée manquante, l'algorithme identifie les K voisins les plus proches en termes de variables disponibles, puis estime la valeur manquante en prenant la moyenne de ces voisins.
Nous avions décidé d'utliser cette méthode en choisissant **k=2** pour les variables de démographies(car nous avions remarqué deux groupes : jeunes et vieux) et **k = 5** pour les varibles sur le triage Evalution(car nous avons 5 niveaux différents d'ESI) mais l'imputation KNN, bien qu'efficace, peut être gourmande en termes de puissance de calcul et de temps, surtout avec des ensembles de données très vastes(plus de 500 milles lignes dans notre cas). Ainsi en raison de ces contraintes de puissance de calcul et de temps, nous avons décidé de ne pas utiliser la méthode d'imputation KNN pour notre dataset. Les exigences computationnelles élevées de cette approche ont dépassé les capacités de notre machine, entraînant des temps d'exécution trop longs.(nous avons laissé le script s'éxécuté pendant des heures puis à ça a planté).
Nous avons donc exploré des méthodes alternatives. Après avoir envisagé plusieurs approches, notre choix s'est finalement porté sur l'imputation directe des valeurs manquantes à l'aide des modèles de machine learning en passant par **XGBoost**. Cette idée a été inspirée par les informations tirés de l'article scientifique. Nous avons décidé d'utiliser le xgboost pour construire un modèle de **régression logistique et un DecisionTree.**
Cette approche offre une solution pragmatique, combinant la puissance des modèles avec la nécessité de traiter efficacement les valeurs manquantes dans notre dataset.
Nous avons également effectué un **one-hot-encoding** pour toutes les valeurs catégorielles du dataset.

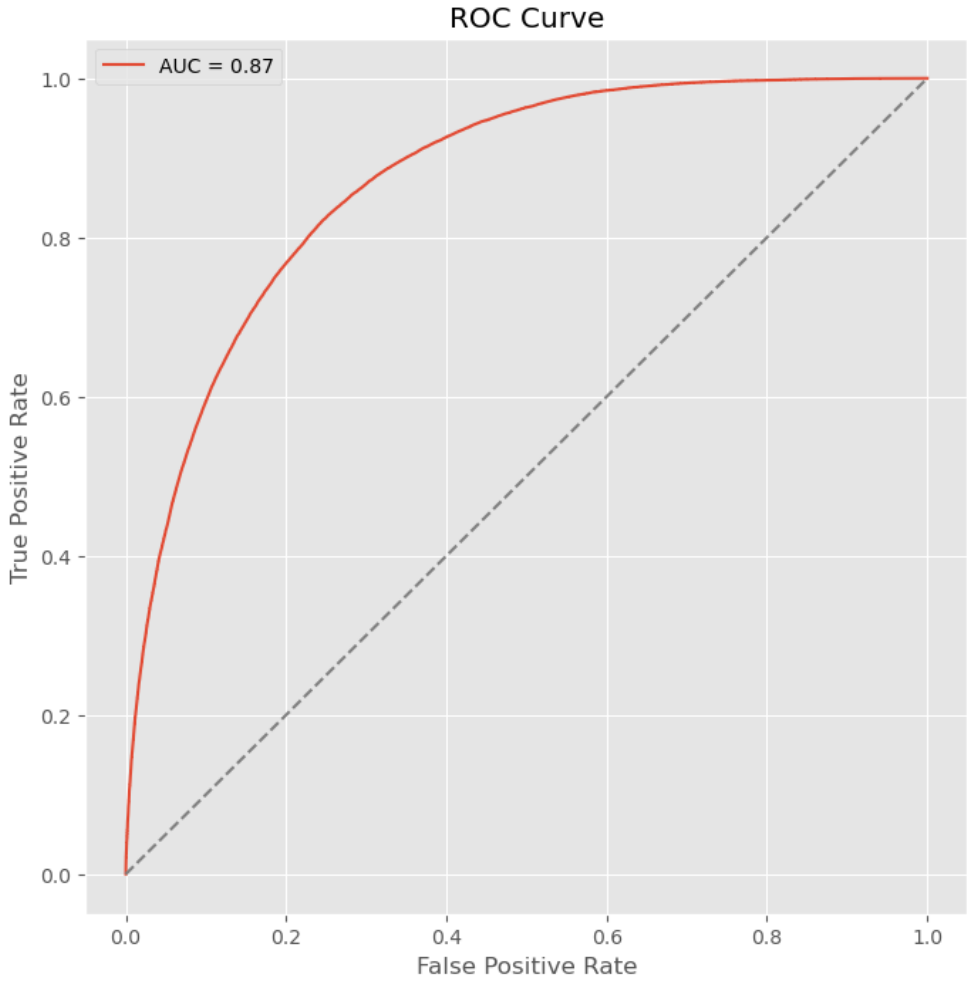
5 Modèles

Après avoir effectué l'EDA et mis en œuvre le prétraitement de nos données, nous avons entamé la phase cruciale de construction de nos modèles. Forts des résulats obtenus lors de l'EDA et des décisions prises lors du prétraitement pour les variables manquantes, nous avons choisi de construire deux modèles principaux : **une regression logistique avec XGBoost et un DecisionTree.**
XGBoost a été choisi en raison de sa capacité à gérer efficacement les valeurs manquantes et à traiter des ensembles de données complexes. Sa capacité à gérer des relations non linéaires et à ajuster automatiquement les poids des observations fait de XGBoost un choix idéal, en particulier dans notre contexte de classification binaire. L'arbre de décision a également été choisi pour sa simplicité, son interprétabilité et aussi sa capacité à gérer les valeurs manquantes.
En parallèle à la sélection de ces modèles, nous avons décidé de focaliser notre analyse sur **un sous-ensemble du dataset**. Cette décision repose sur les catégories spécifiques que nous avons minutieusement étudiées lors de l'EDA , révélant une forte corrélation avec la variable cible(toutes les catégories à part chief complaint qui s'est averrée être la catégorie la moins impactante). En concentrant nos efforts sur ces catégories, nous visons à maximiser la pertinence de nos modèles et à cibler les caractéristiques les plus influentes pour notre problème de classification. Par ailleurs les variables de ces catégories sont, d'après l'article scientifique, les variables avec "information gain" les plus importantes.

5.1 Modèles avec le sous-ensemble du dataset

5.1.1 Regression Logistique avec XGBoost

Le premier modèle que nous avons construit est le modèle de régression logistique binaire.
Le modèle de régression logistique basé sur XGBoost se révèle être un outil solide dans notre étude du triage aux urgences. Avec une **accuracy de 81% et une Aire sous la Courbe (AUC) de 87%**, il montre une bonne capacité à distinguer entre les patients "admit" et ceux "discharge". La matrice de confusion met en lumière une bonne reconnaissance des vrais positifs et négatifs, mais signale aussi quelques erreurs de classification des admissions. Le rapport de classification offre une vue détaillée des performances du modèle pour chaque catégorie, soulignant l'équilibre nécessaire pour prédire correctement les admissions et les sorties. Ces résultats globaux témoignent de l'efficacité de notre modèle en prenant en compte tout simplement le sous-ensemble du dataset.




```
Confusion Matrix:
[[105388  12667]
 [ 19401  30690]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.89      0.87    118055
           1       0.71      0.61      0.66     50091

    accuracy                           0.81    168146
   macro avg       0.78      0.75      0.76    168146
weighted avg       0.80      0.81      0.81    168146
```

5.1.2 Descision Tree

Le deuxième modèle que nous avons construit est le modèle de Decision Tree.
Le modèle offre de bonnes performances avec **une accuracy de 77%** et **une AUC de 76%**. Il a correctement identifié près de **80% des sorties et 70% des admissions** . Cependant, le recall pour les admissions est relativement bas, à 40%, indiquant que le modèle a tendance à manquer certains cas d'admission.

```
Accuracy: 0.7710204227278674
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.93      0.85    118055
           1       0.70      0.40      0.51     50091

    accuracy                           0.77    168146
   macro avg       0.74      0.66      0.68    168146
weighted avg       0.76      0.77      0.75    168146
```

En synthèse, nos deux modèles, la régression logistique basée sur XGBoost et le Decision Tree, ont dévoilé de bonnes performances sur le sous-ensemble du dataset.

Le modèle de régression logistique avec XGBoost s'est révélé très efficace, affichant une précision globale de 81% et une AUC de 87%. Ses capacités à traiter efficacement les valeurs manquantes et à bien distinguer entre les classes en font pour le moment notre meilleur modèle pour la prédiction du triage aux urgences.

5.2 Extension des Modèles au Dataset complet

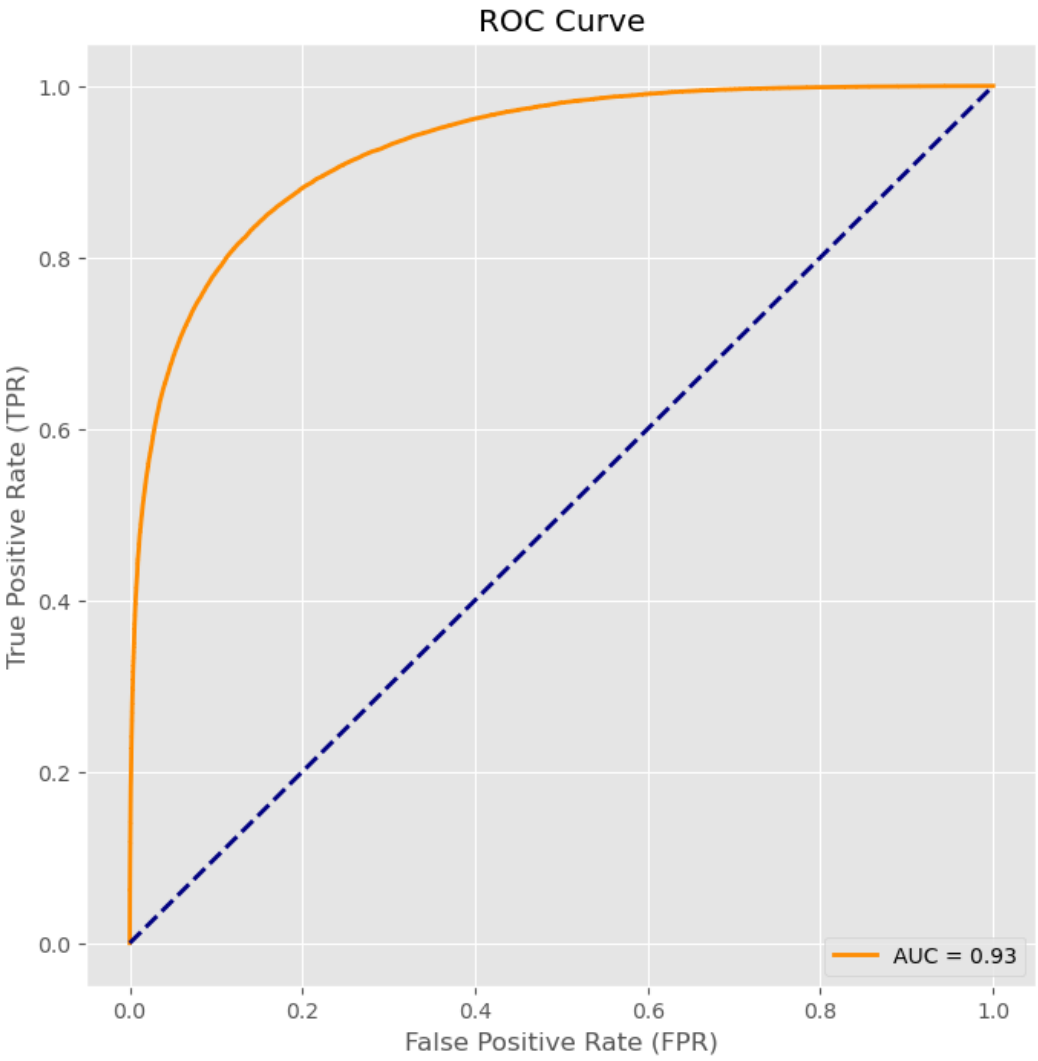
Suite à la conception initiale de nos modèles sur un sous-ensemble du dataset et nos bons résultats obtenus, nous avons voulu les étendre au dataset complet et comparer la différence de performance. Allons-nous obtenir de bien meilleurs résultats avec l'ensemble des variables du dataset?

Lors de l'extension du modèle de régression logistique à l'ensemble du dataset, nous observons une amélioration notable de la performance par rapport aux résultats initiaux sur le sous-ensemble. L'accuracy a legèrement augmenté passant de **81% à 87%**, pareil pour l'AUC qui passe de **87% à près de 93%** soit une amélioration globale de **6%** par rapport à notre modèle sur le sous-ensemble du dataset.

```
Accuracy: 0.8713142150274167
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91    118055
           1       0.84      0.70      0.77     50091

    accuracy                           0.87    168146
   macro avg       0.86      0.82      0.84    168146
weighted avg       0.87      0.87      0.87    168146
```



Quant à l'extension du modèle de Decision Tree à l'ensemble du dataset, il maintient aussi de bonnes performances, avec **une meilleure accuracy de 83%**. Le modèle excelle aussi dans la prédiction des sorties avec **une précision de 85% et un recall de 91%**, confirmant sa compétence à identifier correctement ces cas. Pour les admissions , le modèle présente une précision de 75% et un rappel de 62%, avec un F1-score global de 68%.
L'Aire sous la Courbe (AUC) de 82% souligne la capacité du modèle de Decision Tree à classer correctement les échantillons "admit" et "discharge". Bien que légèrement inférieure à l'AUC obtenue lors de l'analyse du sous-ensemble (AUC de 87‰).
Nous avons obtenons ici aussi une amélioration globale de 6% par rapport au modèle précédent.

```
Accuracy: 0.8261272941372378
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.91      0.88    118055
           1       0.75      0.62      0.68     50091

    accuracy                           0.83    168146
   macro avg       0.80      0.77      0.78    168146
weighted avg       0.82      0.83      0.82    168146
```

En bref, l'extension de nos modèles au dataset complet nous a permis d'obtenir une amélioration globale de 6% de la performance. Ainsi on se rend compte que les 4 catégories analysées lors de l'EDA :**Demographics, Triage evaluation, Hospital usage statistic et Outpatient medications**, sont les catégories influençant le plus la disposition des patients.

6 Conclusion

Notre exploration approfondie du triage aux urgences, combinant une analyse approfondie des données, des modèles de machine learning sur un sous-ensemble du dataset et une extension à l'ensemble du dataset, offre une perspective globale sur l'efficacité de nos approches. Les deux modèles, la Régression Logistique avec XGBoost et le Decision Tree, ont démontré de très bons résultats dans la prédiction de la disposition des patients aux urgences.**La Régression Logistique** sur le dataset complet a émergé comme étant **le meilleur de nos modèles** avec une précision globale de 87.1% et une AUC de 92%. Le Decision Tree, bien que moins complexe, a maintenu des performances solides avec une précision de 82.6% et une AUC de 82%. Il excelle particulièrement dans la prédiction des sorties.