**Rapport Mi-Parcours - Modélisation et Avancement**

A plane flying over a white background

Description automatically generated

**Votre prénom et nom : Yasmine Benayad**

**Votre numéro d’étudiant : 14389236330**

**Trimestre d’inscription : Automne 2024**

**Date d’envoi : 12-11-2024**

**Nom du projet sélectionné**

Satisfaction des passagers des compagnies aériennes.

**Introduction**

Ce rapport mi-parcours détaille l'avancement du projet sur la satisfaction des passagers des compagnies aériennes.

L'objectif principal est d'analyser les facteurs influençant la satisfaction afin de formuler des recommandations aux compagnies aériennes.

À ce stade, plusieurs étapes clés du projet ont été réalisées, et le plan pour les prochaines étapes est en place.

**Étapes Clés Réalisées**

1. **Collecte et Préparation des Données**

Source de données : Le dataset "Passenger Airline Satisfaction" de Kaggle a été collecté.

Les fichiers de données de train et de test ont été fusionnés pour créer un unique fichier, nommé final\_df1.

Ce fichier a ensuite été connecté à une base de données MySQL pour faciliter l'accès et la manipulation des données.

Nettoyage des données : Le dataset a été nettoyé pour traiter les données manquantes principalement par imputation et suppression des entrées incomplètes.

Les valeurs aberrantes ont été identifiées et traitées.

Encodage : Les variables catégorielles ont été converties en numériques via l'encodage one-hot encoding.

1. **Analyse Exploratoire des Données (EDA)**

Des visualisations ont été créées pour comprendre les distributions et les relations entre les variables, en utilisant matplotlib et seaborn.

Des statistiques descriptives ont mis en lumière les tendances générales du dataset.

Voici ci-dessus quelque EDA :

A graph of a bar graph

Description automatically generated with medium confidence

Interprétation du graphique de la distribution des classes :

* **Business** : Cette classe a un nombre très élevé, ce qui indique qu'un nombre significatif de passagers appartient à cette catégorie.
* **Eco** : Cette classe a également un nombre considérable, bien qu'il soit inférieur à celui de la classe "Business".
* **Eco Plus** : Cette classe a un nombre remarquablement bas, suggérant que très peu de passagers tombent dans cette catégorie.

**Déséquilibre des Données :**

La grande disparité entre les comptes pour "Business" et "Eco Plus" indique un potentiel déséquilibre des classes, cela peut affecter la performance du modèle, en particulier dans les tâches de classification, car les modèles peuvent devenir biaisés en faveur de la classe majoritaire ("Business").

**Impact sur la Modélisation :**

Lors de l'entraînement des modèles d'apprentissage automatique, il est crucial de traiter le déséquilibre des classes.

L'utilisation d'algorithmes capables de gérer les poids de classe peut être nécessaire. Des techniques telles que le suréchantillonnage de la classe minoritaire ou le sous-échantillonnage de la classe majoritaire peuvent également être envisagées pour améliorer la performance du modèle.

A graph showing two people

Description automatically generated

Interprétation du graphique :

**Distribution par Genre :**

* **Masculin(Male)** :
  + - Le nombre de passagers "neutres ou insatisfaits" est plus élevé que celui des passagers "satisfaits", ce qui indique une insatisfaction potentielle chez les passagers masculins.
* **Féminin (Female)** :
  + - Les comptes semblent être plus proches, avec les passagers "satisfaits" légèrement au-dessus des passagers "neutres ou insatisfaits".

**Niveaux de Satisfaction:**

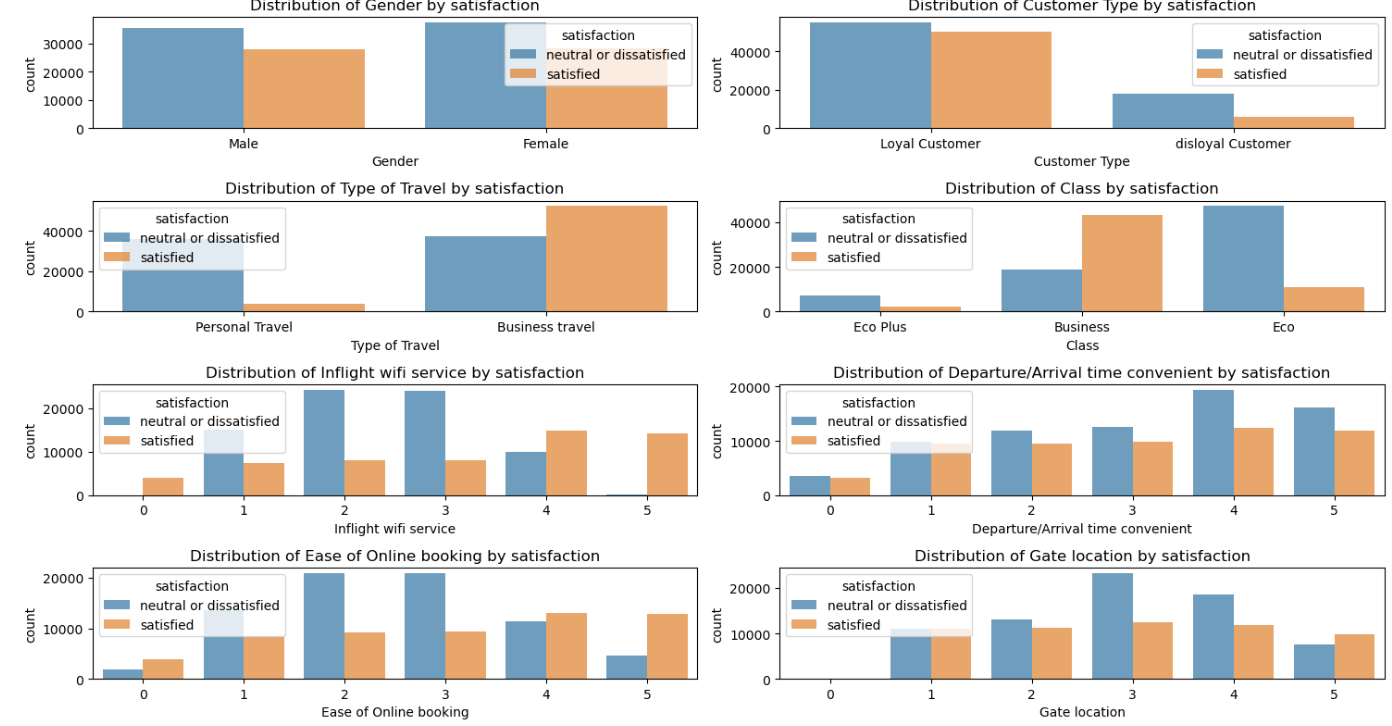
* Dans l'ensemble, les femmes semblent avoir un nombre plus élevé de passagers satisfaits par rapport aux hommes, suggérant que les passagers féminins pourraient avoir une expérience plus positive.

**Insights Potentiels :**

* Les données peuvent indiquer des différences dans la manière dont les passagers masculins et féminins perçoivent leur expérience de voyage, ce qui pourrait être utile pour les compagnies aériennes afin d’adapter leurs services ou de traiter des griefs spécifiques.

**Impact sur l'Analyse :**

* Comprendre la satisfaction par genre peut éclairer les stratégies d'amélioration de l'expérience passager. Cela pourrait inclure un marketing ciblé, des améliorations de service ou des efforts spécifiques d'engagement pour traiter les insatisfactions



Interprétation des graphiques :

1. **Distribution par genre**:

* **Masculin vs. Féminin** : Met en évidence les niveaux de satisfaction entre les genres, avec la proportion de passagers satisfaits par rapport à ceux qui sont neutres ou insatisfaits pour les deux genres.

1. **Distribution du Type de Client par Satisfaction** :

* **Client Fidèle vs. Client d'Affichage** : Fournit des informations sur la satisfaction des clients fidèles par rapport à celle des nouveaux clients ou des clients d'affichage.

1. **Distribution du Type de Voyage par Satisfaction** :

* Examine comment les différents types de voyage (par exemple, personnel, professionnel) influencent les niveaux de satisfaction, pouvant indiquer des attentes et des expériences variées.

1. **Distribution par Classe par Satisfaction** :

* Compare la satisfaction dans différentes classes de voyage (Éco, Affaires, Éco Plus) pour déterminer si des classes supérieures entraînent une satisfaction accrue.

1. **Distribution du Service en Vol par Satisfaction** :

* Évalue comment le service en vol impacte la satisfaction des passagers, où une tendance à l'insatisfaction pourrait indiquer des domaines à améliorer.

1. **Distribution de la Commodité des Horaires de Départ/Arrivée par Satisfaction** :

* Évalue si les passagers éprouvent un sentiment positif ou négatif concernant la commodité de leurs horaires de vol, ce qui est crucial pour l'expérience globale de voyage.

1. **Distribution de la Facilité de Réservation en Ligne par Satisfaction** :

* Analyse comment la facilité du processus de réservation en ligne affecte la satisfaction, un élément clé pour les interactions numériques avec les compagnies aériennes.

1. **Distribution de la Localisation des Portes d'Embarquement par Satisfaction** :

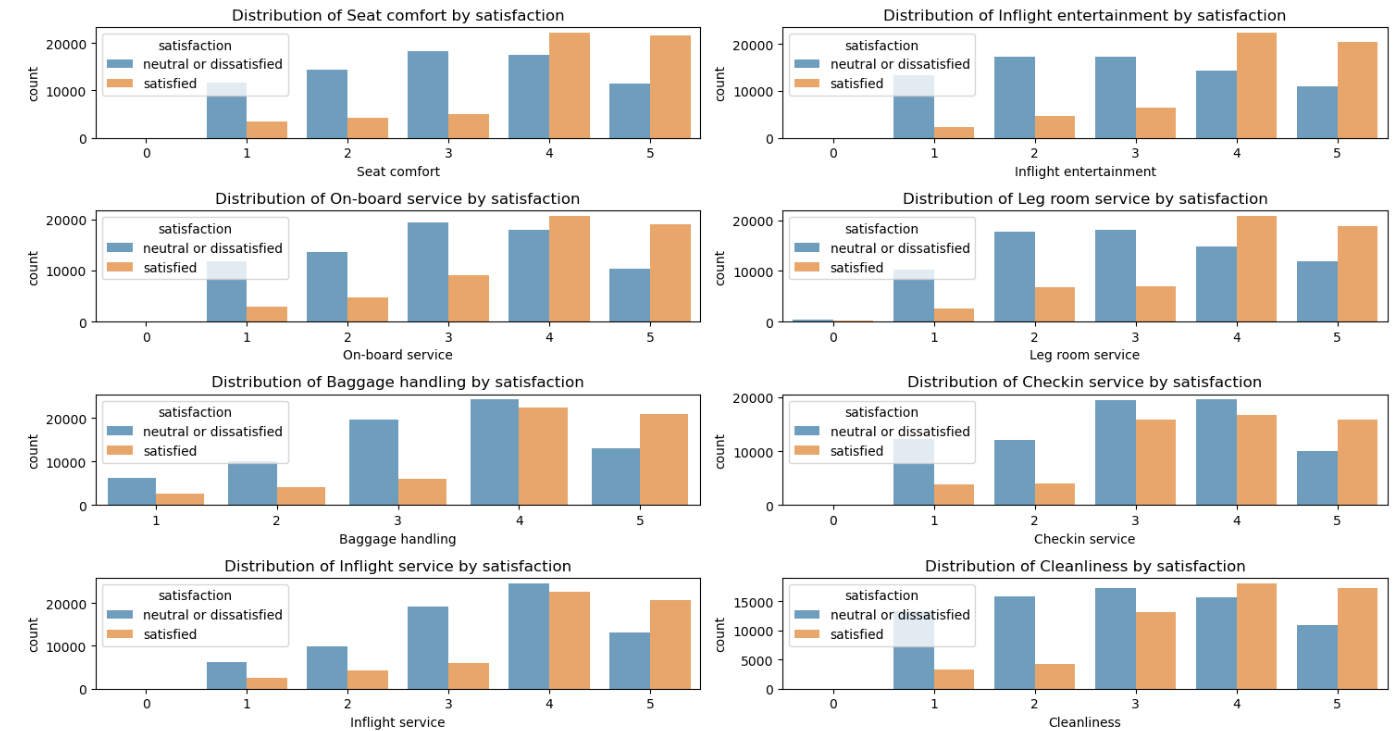
* Examine comment la commodité ou l'incommodité des emplacements des portes d'embarquement impacte les évaluations globales de satisfaction.

**Insights Potentiels :**

* **Tendances Démographiques** : Les différences de satisfaction en fonction du genre et du type de client peuvent guider des améliorations ciblées.
* **Amélioration du Service** : Des domaines comme le service en vol et la facilité de réservation en ligne peuvent nécessiter une attention immédiate pour améliorer l'expérience des passagers.
* **Satisfaction par Classe** : Comprendre comment la satisfaction varie selon la classe peut aider à adapter les services pour améliorer les expériences des différents segments de passagers.

**Impact sur l'analyse :**

* Les visualisations peuvent grandement aider à identifier les domaines qui pourraient nécessiter des améliorations ou des ajustements.
* Les compagnies aériennes pourraient utiliser ces insights pour améliorer leur service en fonction des données démographiques et des interactions de service.



Interprétation des graphiques :

1. **Distribution du Confort des Sièges par Satisfaction** :

* Examine comment les passagers notent le confort des sièges et comment cela se corrèle avec leurs niveaux de satisfaction, des notes élevées pourraient indiquer que des sièges confortables mènent à une satisfaction accrue.

1. **Distribution du Divertissement en Vol par Satisfaction** :

* Analyse la satisfaction liée aux options de divertissement en vol, une tendance vers une satisfaction plus faible pourrait suggérer que des améliorations sont nécessaires dans ce domaine.

1. **Distribution du Service à Bord par Satisfaction** :

* Évalue comment le service à bord impacte la satisfaction des passagers, une réponse positive constante pourrait indiquer une prestation de service efficace.

1. **Distribution du Service d'Espace pour les Jambes par Satisfaction** :

* Examine comment l'espace pour les jambes contribue à la satisfaction globale, si de nombreux répondants le notent mal, cela pourrait indiquer un besoin d'ajustements dans l'agencement des sièges.

1. **Distribution de la Gestion des Bagages par Satisfaction** :

* Évalue la satisfaction concernant la gestion des bagages, des niveaux élevés d'insatisfaction pourraient mettre en évidence des inefficacités dans ce processus.

1. **Distribution du Service d'Enregistrement par Satisfaction** :

* Mesure la satisfaction liée aux procédures d'enregistrement, des retours positifs ici pourraient souligner des opérations efficaces, tandis que des retours négatifs pourraient indiquer des domaines à améliorer.

1. **Distribution du Service en Vol par Satisfaction** :

* Examine la qualité du service en vol et sa relation avec la satisfaction globale des passagers, ce domaine pourrait être crucial pour améliorer l'expérience de voyage.

1. **Distribution de la Propreté par Satisfaction** :

* Évalue comment la propreté influence les niveaux de satisfaction, des notes faibles pourraient nécessiter une attention immédiate de la part de la direction de la compagnie aérienne.

**Insights Potentiels :**

* **Aspects Clés de la Satisfaction** : Les graphiques aident à identifier quels aspects de l’expérience de voyage influencent de manière significative la satisfaction des passagers, orientant ainsi les compagnies aériennes sur où concentrer leurs efforts d’amélioration.
* **Zones d'Amélioration Ciblées** : Les domaines avec de faibles niveaux de satisfaction peuvent nécessiter des actions immédiates pour résoudre les problèmes et améliorer l'expérience globale des passagers.

**Impact sur l'Analyse :**

* Comprendre la satisfaction des passagers à travers diverses dimensions de service permet aux compagnies aériennes d'adapter leurs services et de traiter des griefs spécifiques, contribuant ainsi à des niveaux de satisfaction globaux plus élevés.

A group of blue and white graphs

Description automatically generated

Interprétation des graphiques scatter plot :

Le graphe Âge vs. Distance de Vol : nous montre la relation entre l'âge et la distance de vol.

Le nuage de points indique qu'il n'y a pas de corrélation forte.

Le graphe Âge vs. Retard de Départ en Minutes :

Représente la relation entre l'âge et le retard au départ.

Distribution généralement éparse sans tendance claire.

Le graphe Âge vs. Retard d'Arrivée en Minutes :

Affiche la relation entre l'âge et le retard à l'arrivée.

Distribution éparse similaire aux précédents graphiques.

Le graphe Distance de Vol vs. Retard d'Arrivée en Minutes :

Montre la distance de vol par rapport au retard à l'arrivée.

Le nuage de points indique une certaine variabilité mais pas de corrélation forte.

Le graphe Distance de Vol vs. Retard de Départ en Minutes :

Illustre la relation entre la distance de vol et le retard au départ.

Les points de dispersion suggèrent une large distribution.

Le graphe Retard de Départ en Minutes vs. Retard d'Arrivée en Minutes :

Montre une corrélation positive claire ; à mesure qu'un retard augmente, l'autre tend à augmenter aussi.

Ce graphique suggère une forte relation entre les retards de départ et d'arrivée.

A graph of blue and orange bars

Description automatically generated

Graphique : Compte des ID par Groupe d'Âge et Satisfaction

Axes :

L'axe des x représente les groupes d'âge, divisés en tranches (par exemple, 18-30, 31-40, etc.).

L'axe des y indique le nombre d'individus dans chaque groupe d'âge, en fonction de leur satisfaction.

Barres :

Les barres sont colorées pour représenter différents niveaux de satisfaction :

Satisfait (orange)

Neutre ou insatisfait (bleu)

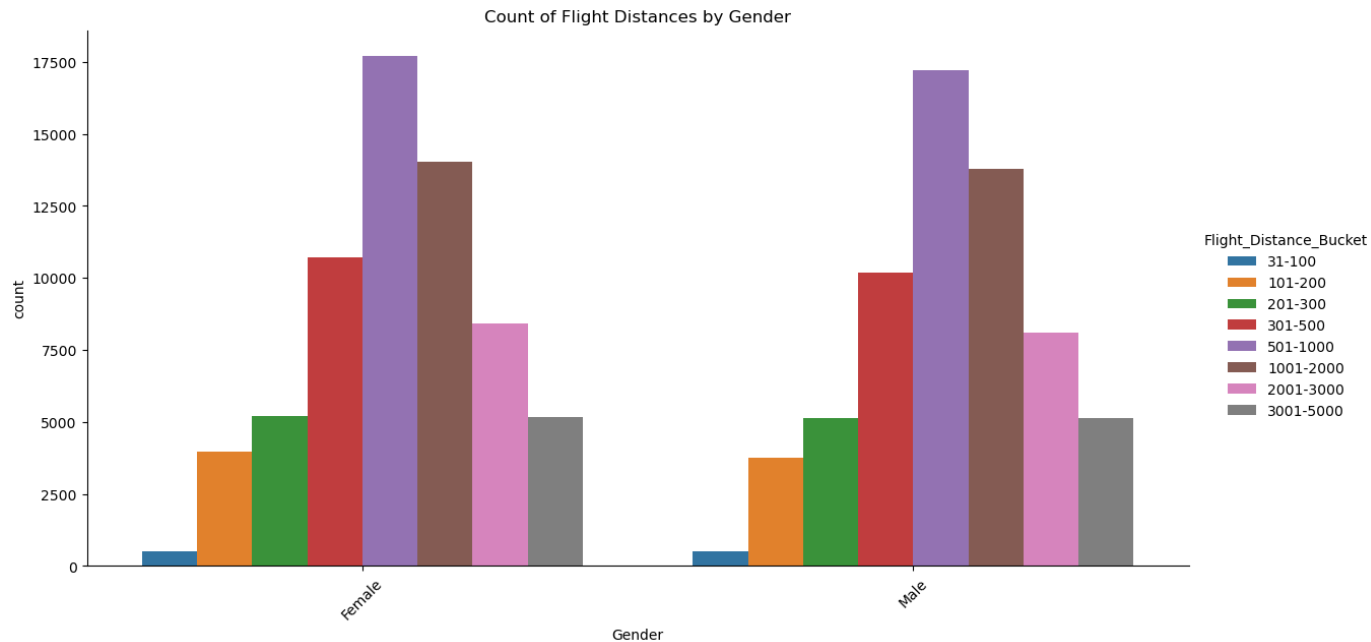
Observations :

Distribution de Satisfaction :

* Les groupes d'âge 19-30 et 31-40 semblent avoir un nombre élevé de personnes satisfaites par rapport aux autres groupes.
* Les groupes plus âgés (61-70 et 71-80) ont généralement moins d'individus, et une proportion plus élevée émerge dans la catégorie neutre ou insatisfaite.

Tendances Générales :

* La satisfaction semble diminuer chez les groupes d'âge plus avancés, tandis que les jeunes adultes montrent une plus grande satisfaction.
* Ce type d'analyse peut fournir des informations précieuses sur les perceptions de satisfaction selon les différents groupes d'âge



Graphique : Compte des Distances de Vol par Genre

Axes :

L'axe des x représente le genre (Féminin et Masculin).

L'axe des y indique le nombre de vols dans chaque catégorie de distance de vol.

Barres :

Les barres sont colorées pour représenter différentes plages de distance de vol, avec des catégories telles que :

31-100

101-200

201-300

301-500

501-1000

1001-2000

2001-3000

3001-4000

Observations :

Distribution par Genre :

* La distribution des distances de vol semble assez similaire entre les genres, avec des hauteurs de barres comparables pour chaque catégorie.
* Les hommes et les femmes semblent tous deux avoir des vols fréquents dans les catégories de courtes à moyennes distances (31-500).

Catégories de Distance :

* On observe peut-être un nombre légèrement plus élevé pour certaines catégories de distance pour l'un ou l'autre genre.
* Il peut y avoir une tension sur les distances plus longues (au-delà de 1000 km), ce qui pourrait varier selon les préférences de voyage des deux genres.

Conclusion :

Ce graphique permet d'analyser comment les distances de vol varient en fonction du genre

1. **Prétraitement**

Normalisation : Les caractéristiques numériques ont été normalisées pour garantir une échelle comparable.

Les caractéristiques numériques ont été normalisées en utilisant la méthode de normalisation min-max, où chaque valeur est mise à l'échelle pour se situer entre 0 et 1.

Cela garantit que toutes les caractéristiques contribuent de manière équitable à la distance calculée par des algorithmes sensibles à l'échelle, tels que KNN ou les réseaux de neurones.

La normalisation aide également à améliorer la convergence des algorithmes d'optimisation lors de l'entraînement des modèles.

Division des Données : Le dataset a été divisé en ensembles d'entraînement (75%) et de test (25%) pour évaluer la performance des modèles.

Cette division permet de former le modèle sur un sous-ensemble des données tout en réservant une portion pour évaluer la performance finale du modèle sur des données non vues.

Cette pratique prévient le surajustement et garantit que les performances du modèle puissent être généralisées à de nouvelles données.

La division des données peut être réalisée de manière aléatoire pour assurer une représentation équilibrée des classes dans chaque ensemble.

1. **Modélisation Initiale**

Des modèles de machine learning tels que la régression logistique, les arbres de décision, les forêts aléatoires, ainsi que des algorithmes comme SVM (Machines à Vecteurs de Support), KNN (k-Nearest Neighbors) et Gradient Boosting ont été appliqués.

Les modèles ont été évalués sur des métriques de base comme la précision, le rappel, et le score F1.

**Résultats Préliminaires**

**Régression Logistique :**

A démontré un bon équilibre entre complexité et performance, avec des résultats solides en termes de précision et de généralisation.

**Arbres de Décision :**

Offrent une interprétabilité facile, mais présentent une légère baisse de précision, ce qui peut limiter leur efficacité sur des ensembles de données variés.

**Forêts Aléatoires :**

Ont montré la meilleure performance en termes de précision et de score F1, en surmontant le problème de surajustement grâce à l'agrégation de multiples arbres, ce qui renforce leur capacité de généralisation.

**SVM (Machines à Vecteurs de Support) :**

Les SVM, notamment avec les noyaux linéaire et RBF, affichent un bon compromis entre précision et rappel, ils se révèlent efficaces pour des ensembles de données équilibrés et gèrent le surajustement tout en maintenant une haute précision.

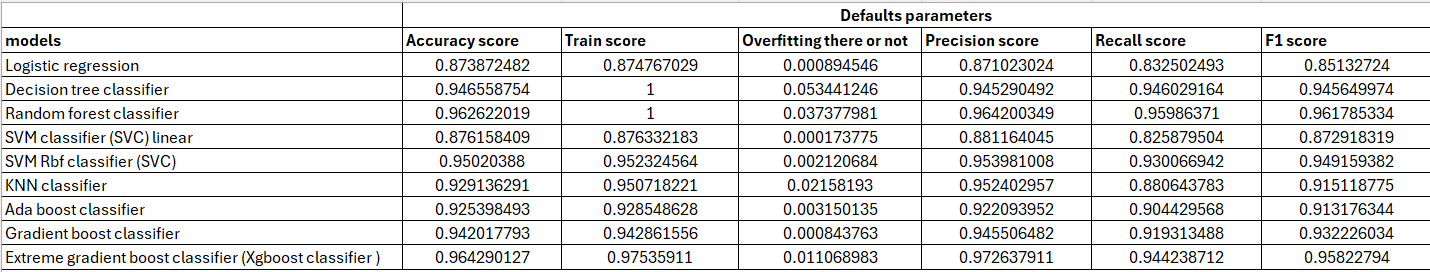
**KNN (K-Nearest Neighbors) :**

Le classificateur KNN offre également une haute précision, mais présente un risque accru de surajustement, ce qui pourrait affecter sa capacité à généraliser sur de nouvelles données.

**Gradient Boosting :**

A démontré de solides performances avec une précision élevée et un excellent score F1, maintient un bon équilibre entre précision et rappel tout en limitant le surajustement, ce qui en fait une option idéale pour des ensembles de données complexes et non linéaires.

Les modèles évalués montrent chacun des avantages distincts, permettant de choisir le meilleur en fonction des spécificités et des exigences du problème à résoudre

****

Le tableau Excel ci-dessus fournit diverses métriques pour différents modèles d'apprentissage automatique pour notre projet

* Accuracy Score (Taux de Précision) : Indique comment bien le modèle classe correctement les instances.
* Train Score (Taux d'Entraînement) : Indique la précision sur l'ensemble de formation.
* Overfitting Score (Score de Surapprentissage) : Une valeur plus basse indique que le modèle généralise bien à de nouvelles données.
* Precision Score (Taux de Précision) : Mesure l'exactitude des prédictions positives.
* Recall Score (Taux de Rappel) : Mesure dans quelle mesure le modèle capture toutes les instances positives.
* F1 Score : Moyenne harmonique de la précision et du rappel, fournissant une mesure équilibrée.

**Résultats des parametres par defaut du projet (Default Parameters):**

* **La régression logistique** **(Logistic Regression):**
* Accuracy score : 0.873
* Train Score: 0.874
* Overfitting: Non
* Precision Score: 0.868064609
* Recall Score: 0.906090189
* F1 Score: 0.887956133

Cela indique que le modèle prédit correctement le résultat environ 87,34 % du temps. C'est une performance solide dans l'ensemble, suggérant que le modèle est efficace pour distinguer les classes.

Le fait que le score d'apprentissage soit très proche du taux de précision indique que le modèle fonctionne de manière constante bien sur les données d'apprentissage sans différence significative. C'est un signe positif, suggérant que le modèle ne surajuste pas les données d'entraînement.

La précision mesure la proportion de prédictions positives vraies par rapport au total des prédictions positives. Un score de 86,81 % indique que lorsque le modèle prédit une classe positive, il a raison environ 86,81 % du temps, ce qui suggère un faible nombre de faux positifs.

Le rappel mesure la proportion de prédictions positives vraies par rapport aux véritables cas positifs dans les données. Un rappel de 90,61 % suggère que le modèle identifie correctement 90,61 % de tous les cas positifs réels, indiquant qu'il est efficace pour capturer les instances positives, bien qu'il puisse encore y avoir quelques faux négatifs

Le score F1 est la moyenne harmonique de la précision et du rappel, fournissant un équilibre entre les deux. À 88,80 %, il indique un bon équilibre entre précision et rappel, cela est particulièrement important dans des scénarios où les faux positifs et les faux négatifs peuvent entraîner des conséquences significatives.

Le modèle de régression logistique présente de solides performances, avec une bonne précision et un bon équilibre entre précision et rappel, sa capacité à généraliser sans surajustement le rend adapté pour des applications pratiques.

* **Arbre de décision** (**Decision Tree Classifier):**
* Accuracy Score: 0.946850744
* Train Score: 1
* Overfitting surajustement: oui
* Précision Score: 0.947198501
* Recall Score: 0.941775303
* F1 Score: 0.944484678

Ce taux de précision score indique que le modèle prédit correctement environ 94,69 % des cas. C'est une performance très élevée, suggérant que le modèle est efficace pour distinguer les classes

Un score d'apprentissage parfait (1) signifie que le modèle a parfaitement appris à partir des données d'entraînement, cependant, cela soulève des préoccupations concernant le surajustement.

L'indication de surajustement signifie que le modèle a mémorisé les données d'entraînement au lieu de généraliser ces connaissances à des données non vues, cela peut entraîner de mauvaises performances sur des données réelles, car le modèle pourrait ne pas bien réagir à des cas qu'il n'a pas déjà vus.

Un score de precision 94,72 % indique que parmi toutes les prédictions positives faites par le modèle, 94,72 % étaient correctes. Cela suggère que le modèle a un bon contrôle sur les faux positifs.

Avec un score de rappel de 94,18 %, le modèle est capable d'identifier environ 94,18 % des cas positifs réels. Cela montre une bonne capacité à capturer les instances positives, avec une proportion relativement faible de faux négatifs.

Un score F1 de 94,45 % indique un bon équilibre entre la précision et le rappel. Cela suggère que le modèle est à la fois précis dans ses prédictions positives et efficace pour détecter des cas positifs.

L’Arbre de Décision montre une excellente performance en termes de précision, rappel et score F1. Toutefois, le surajustement observé soulève des préoccupations quant à sa capacité à généraliser sur des données nouvelles, pour corriger cela, des techniques comme la validation croisée, la régularisation ou la réduction de la complexité du modèle pourraient être envisagées..

* **Random Forest Classifier:**
* Accuracy Score: 0.962210910
* Train Score: 1
* Overfitting: oui
* Precision Score: 0.964682056
* Recall Score: 0.94305986
* F1 Score: 0.953747099

Le classificateur Random Forest présente une excellente performance avec un taux de précision, de rappel et un score F1 très élevés.

Ce score indique que le modèle prédit correctement environ 96,22 % des cas, cela reflète une performance exceptionnelle, montrant que le modèle est très efficace pour distinguer les classes.

Un score d'apprentissage parfait (1) signifie que le modèle a parfaitement appris à partir des données d'entraînement, cela indique également un risque de surajustement.

L'indication de surajustement signifie que le modèle a trop mémorisé les données d'entraînement, ce qui pourrait nuire à sa capacité à généraliser sur de nouvelles données, cela peut entraîner une dégradation de la performance sur des cas non vus.

Avec un score de précision de 96,47 %, le modèle est très précis dans ses prédictions positives. Cela signifie que parmi toutes les prédictions positives, 96,47 % étaient effectivement correctes, réduisant ainsi les faux positifs.

Le score de rappel de 94,31 % indique que le modèle est capable d'identifier environ 94,31 % des cas positifs réels. Cela montre une bonne capacité à détecter les instances positives, bien qu'il puisse y avoir quelques faux négatifs.

Un score F1 de 95,37 % reflète un bon équilibre entre la précision et le rappel, cela signifie que le modèle est à la fois efficace dans la détection des cas positifs et fiable dans ses prédictions.

Pour améliorer la généralisation du modèle, des techniques telles que la validation croisée, la réduction du nombre d'arbres, ou la régularisation peuvent être envisagées.

* **SVM Classifier (Linear):**
* Accuracy Score: 0.965990421
* Train Score: 0.999376514
* Overfitting: oui
* Precision Score: 0.966626788
* Recall Score: 0.958020091
* F1 Score: 0.962308347

Le taux de precision indique que le modèle prédit correctement environ 96,59 % des cas. Cela démontre une performance très élevée, montrant que le modèle est efficace pour discriminer les différentes classes.

Un score d'apprentissage très proche de 1 (0,9994) signifie que le modèle a presque parfaitement appris à partir des données d'entraînement, cependant, cela soulève des préoccupations quant à un éventuel surajustement.

L'indication de surajustement signifie que le modèle a tendance à mémoriser les données d'entraînement plutôt que d'apprendre à généraliser à de nouvelles données, cela peut nuire à sa performance sur des instances non vues.

Avec un score de précision 96,66 %, le modèle est très précis dans ses prédictions positives. Cela signifie qu'environ 96,66 % des prédictions positives faites par le modèle sont correctes, ce qui indique un faible nombre de faux positifs.

Un score de rappel de 95,80 % suggère que le modèle identifie correctement environ 95,80 % des cas positifs réels, cela montre une bonne capacité à détecter les instances positives, bien qu'il puisse toujours y avoir quelques faux négatifs.

Avec un score F1 de 96,23 %, le modèle montre un excellent équilibre entre précision et rappel. Cela reflète sa compétence à la fois dans la détection des cas positifs et dans la minimisation des faux positifs.

Ce modèle SVM a un bon équilibre entre la précision sur les données d'entraînement et les données de test, le classificateur SVM linéaire présente une excellente performance avec un taux de précision, de rappel et un score F1 très élevés.

Cependant, le surajustement est une préoccupation, car il peut affecter la capacité du modèle à généraliser sur des données nouvelles, pour améliorer la généralisation, des techniques telles que la validation croisée, l'ajustement des paramètres, ou la régularisation peuvent être envisagées.

* **SVM Rbf Classifier (SVC):**
* Accuracy Score: 0.963746325
* Train Score: 0.999205769
* Overfitting: oui
* Precision Score: 0.966566511
* Recall Score: 0.953775077
* F1 Score: 0.960085095

Ce score indique que le modèle prédit correctement environ 96,37 % des cas. Ceci montre que le modèle est très performant dans la classification des instances.

Un score d'apprentissage de 0,9992 indique que le modèle a presque parfaitement appris les données d'entraînement, cela soulève des préoccupations quant à un potentiel surajustement du modèle.

La présence de surajustement signifie que le modèle a tendance à mémoriser les données d'entraînement plutôt qu'à apprendre à généraliser, cela peut nuire à sa performance sur de nouvelles données qu'il n'a pas rencontrées.

Avec un score de 96,66 %, le modèle est très précis dans ses prédictions positives, cela signifie qu'environ 96,66 % des cas que le modèle prédit comme positifs sont effectivement corrects, ce qui indique un faible taux de faux positifs.

Le score de rappel de 95,38 % indique que le modèle identifie correctement environ 95,38 % des cas positifs réels, cela montre une bonne capacité à capturer les instances positives, bien qu'il puisse y avoir quelques faux négatifs.

Avec un score F1 de 96,01 %, le modèle exhibe un bon équilibre entre les mesures de précision et de rappel, indiquant qu'il est efficace à la fois pour détecter les cas positifs et pour réduire les faux positifs.

Le classificateur SVM avec noyau RBF démontre une excellente performance avec un taux de précision, de rappel et un score F1 très élevés.

Toutefois, le surajustement est une préoccupation, ce qui peut affecter la capacité du modèle à généraliser sur des données non vues, pour améliorer cette généralisation, il serait utile d'explorer des techniques telles que la régularisation, l'ajustement des hyperparamètres ou l'utilisation de la validation croisée.

* **KNN Classifier:**
* Accuracy Score: 0.949219474
* Train Score: 0.900990701
* Overfitting: Non
* Precision Score: 0.947531367
* Recall Score: 0.955229364
* F1 Score: 0.951358591

Ce score indique que le modèle prédit correctement environ 94,92 % des cas, cela démontre une bonne performance globale, suggérant que le classificateur est efficace pour distinguer les classes.

Un score d'apprentissage de 90,10 % signifie que le modèle fonctionne assez bien sur les données d'entraînement, mais il est légèrement inférieur au taux de précision global, cela pourrait indiquer une légère chance de sous-ajustement.

L'absence de surajustement signifie que le modèle ne mémorise pas les données d'entraînement, mais apprend plutôt à généraliser, cela est encourageant pour la capacité du modèle à performer sur de nouvelles données.

Avec un score de 94,75 %, cela indique que, parmi toutes les prédictions positives du modèle, 94,75 % sont correctes, cela montre un bon contrôle sur le nombre de faux positifs.

Un score de rappel de 95,52 % indique que le modèle identifie correctement environ 95,52 % des cas positifs réels, cela montre que le classificateur est efficace pour détecter les instances positives, avec une faible proportion de faux négatifs.

Avec un score F1 de 95,14 %, cela indique un bon équilibre entre précision et rappel, signifiant que le modèle est compétent à la fois pour détecter les cas positifs et pour minimiser les faux positifs.

Le classificateur KNN montre de bonnes performances, avec un bon taux de précision, de rappel et un score F1 solide, l'absence de surajustement est un point positif, suggérant que le modèle généralise bien à de nouvelles données.

* **Ada Boost Classifier:**
* Accuracy Score: 0.949219474
* Train Score: 0.93546758
* Overfitting: Non
* Precision Score: 0.947458084
* Recall Score: 0.949360578
* F1 Score: 0.948409329

Ce score indique que le modèle prédit correctement environ 94,92 % des cas, cela suggère que le classificateur est efficace pour distinguer entre les différentes classes.

Un score d'apprentissage de 93,55 % montre que le modèle fonctionne bien sur les données d'entraînement, mais il est légèrement inférieur au taux de précision globale, cela pourrait indiquer un bon ajustement sans sous-ajustement significatif.

Le fait qu'il n'y ait pas de surajustement indique que le modèle généralise bien à de nouvelles données, ce qui est un indicateur positif pour sa robustesse.

Avec un score de 94,75 %, cela signifie que parmi toutes les prédictions positives du modèle, 94,75 % étaient correctes, cela montre que le modèle a un bon contrôle sur les faux positifs.

Un score de rappel de 94,94 % indique que le modèle réussit à identifier environ 94,94 % des cas positifs réels, cela signifie qu'il est efficace pour détecter les instances positives, tout en maintenant un faible nombre de faux négatifs.

Avec un score F1 de 94,84 %, cela indique un bon équilibre entre précision et rappel, suggérant que le classificateur est compétent à la fois dans la détection des cas positifs et dans la minimisation des faux positifs.

Le classificateur Ada Boost démontre de bonnes performances globales, avec un taux de précision, de rappel et de score F1 solides, l'absence de surajustement est positive, renforçant la capacité du modèle à généraliser sur des données inconnues.

* **Gradient Boost Classifier:**
* Accuracy Score: 0.946850744
* Train Score: 0.996371475
* Overfitting: Oui
* Precision Score: 0.944368431
* Recall Score: 0.941775303
* F1 Score: 0.943068647

Ce score indique que le modèle prédit correctement environ 94,69 % des cas, c'est une performance élevée, ce qui suggère que le modèle est efficace pour classer les données.

Un score d'apprentissage de 99,64 % signifie que le modèle a presque parfaitement appris à partir des données d'entraînement.

Le surajustement indique que le modèle a mémorisé les détails des données d'entraînement, ce qui peut nuire à sa capacité à généraliser sur des données non vues, cela pourrait entraîner une performance dégradée lors de l'application sur de nouvelles instances.

Avec un score de 94,44 %, cela signifie qu'environ 94,44 % des prédictions positives faites par le modèle étaient correctes, cela montre que le modèle a un bon contrôle sur les faux positifs.

Un score de rappel de 94,18 % indique que le modèle identifie correctement environ 94,18 % des cas positifs réels, cela suggère une bonne capacité à détecter les instances positives, bien qu'il y ait une proportion légèrement plus élevée de faux négatifs.

Avec un score F1 de 94,31 %, cela indique un bon équilibre entre précision et rappel, cela montre que le modèle est à la fois efficace dans la détection des cas positifs et en mesure de minimiser les faux positifs.

Le classificateur Gradient Boost présente de bonnes performances, mais le surajustement est une préoccupation à prendre en compte, bien que les taux de précision, de rappel et de score F1 soient solides, la capacité du modèle à généraliser sur de nouvelles données pourrait être compromise. Pour améliorer la généralisation, des techniques comme la validation croisée, la réduction de la complexité du modèle ou la régularisation peuvent être envisagées.

* **Extreme Gradient Boost Classifier (XGBoost):**
* Accuracy Score: 0.965136428
* Train Score: 0.999476177
* Overfitting: Oui
* Precision Score: 0.965088636
* Recall Score: 0.959012554
* F1 Score: 0.962036482

Ce score indique que le modèle prédit correctement environ 96,51 % des cas, cela démontre une performance très élevée, montrant que le classificateur est efficace pour distinguer les classes.

Un score d'apprentissage de 99,95 % signifie que le modèle a quasi parfaitement appris à partir des données d'entraînement, cela soulève des préoccupations liées au surajustement.

L'indication de surajustement signifie que le modèle a mémorisé les données d'entraînement plutôt que de généraliser, cela peut réduire sa capacité à bien performer sur des données nouvelles et non vues.

Avec un score de 96,51 %, cela signifie que parmi toutes les prédictions positives faites par le modèle, 96,51 % étaient correctes, cela indique un excellent contrôle sur le nombre de faux positifs.

Un score de rappel de 95,90 % montre que le modèle identifie correctement environ 95,90 % des cas positifs réels, cela suggère une très bonne capacité à capturer les instances positives, bien qu'il puisse y avoir quelques faux négatifs.

Avec un score F1 de 96,20 %, le modèle montre un bon équilibre entre précision et rappel, reflétant son efficacité à la fois dans la détection des cas positifs et dans la minimisation des faux positifs.

Le classificateur XGBoost présente d'excellentes performances, avec des scores de précision, de rappel et F1 très élevés. Toutefois, le surajustement est une préoccupation, ce qui pourrait limiter la capacité du modèle à généraliser sur de nouvelles données, pour améliorer la généralisation, il pourrait être utile d'appliquer des techniques telles que la régularisation, l'ajustement des hyperparamètres, ou l'utilisation de la validation croisée

**En résumé :**

**Logistic Regression** : Score d'accuracy de 0.8738, léger écart avec le score d'entraînement, suggérant une bonne généralisation.

**Decision Tree Classifier** : Score d'entraînement parfait de 1, mais un score d'accuracy plus bas, indiquant un surapprentissage (overfitting) probable.

**Random Forest Classifier** : Haut score d'accuracy, équilibre entre scores d'apprentissage et de test.

**SVM (lineaire et Rbf)** : Bonnes performances, mais un écart dans le modèle Rbf suggère un potentiel surapprentissage.

**KNN Classifier** : Excellent équilibre entre les scores, faible indicateur de surapprentissage.

**AdaBoost Classifier** : Performance robuste, faible tendance au surapprentissage.

**Gradient Boost Classifier** : Élevé sur le train mais relativement inférieur sur le test, possibilité de surapprentissage.

**Extreme Gradient Boost (Xgboost)** : Très bon score d'accuracy avec un léger écart dans le surapprentissage.

**Surapprentissage** : Notamment présent dans des modèles comme le Decision Tree et le Gradient Boost. Il pourrait être atténué par la régularisation.

**Bonnes Généralisations** : Modèles comme le Logistic Regression et le KNN montrent peu de surapprentissage, suggérant une performance stable.

Les modèles comme XGBoost (précision de 96,51 %) et SVM (SVC Linéaire) (précision de 96,59 %) montrent une excellente précision avec un surajustement gérable, suggérant qu'ils sont robustes et capables de généraliser efficacement sur de nouvelles données.

**Performance Équilibrée :**

Les modèles tels que Gradient Boost et Ada Boost affichent une performance équilibrée, combinant de bonnes valeurs de précision et de rappel (scores F1 autour de 94,3 % et 94,8 %, respectivement) tout en présentant des signes de surajustement.

**Compromis entre Complexité et Précision** :

Les résultats montrent que chaque modèle a ses propres avantages, avec des compromis entre complexité, précision, et capacité de généralisation. Les modèles plus complexes comme Random Forest et Gradient Boost offrent parfois une meilleure précision, mais au risque de surajustement.

**Synthèse**

Sur la base de ces résultats, XGBoost se distingue comme le meilleur modèle globalement en raison de sa haute précision, de son excellent score F1 (96,20 %) et de son surajustement relativement faible.

Les résultats préliminaires indiquent que le choix du meilleur modèle dépendra souvent du contexte spécifique et des exigences du problème à résoudre, en prenant en compte à la fois la précision et la capacité de généralisation.

Prochaines Étapes

* 1. **Optimisation des Modèles**

Affiner les modèles avec la validation croisée et la recherche de grille pour ajuster les hyperparamètres, en particulier pour les forêts aléatoires.

* 1. **Interprétation des Modèles**

Utiliser les mesures d'importance intégrées permet d'évaluer la contribution de chaque caractéristique dans les modèles basés sur des arbres, tels que les forêts aléatoires et les arbres de décision.

Ces mesures fournissent une indication de l'impact de chaque caractéristique sur les prédictions du modèle, en déterminant dans quelle mesure elles aident à réduire l'impureté (comme l'entropie) lors des séparations effectuées, cela aide à identifier les caractéristiques les plus influentes et à mieux comprendre le fonctionnement du modèle.

* 1. **Formulation de Recommandations**

Basé sur l’analyse des résultats des modèles, développer des recommandations pratiques que les compagnies aériennes peuvent implémenter pour améliorer la satisfaction des passagers.

* 1. **Documenter et Valider les Découvertes**

Consolider les conclusions et s'assurer de leur validité par des tests additionnels ou l'utilisation de méthodes statistiques.

**Conclusion**

À mi-parcours, le projet avance conformément au plan initial, les résultats préliminaires indiquent la direction à suivre pour optimiser les modèles et finaliser les recommandations.

Les prochaines étapes se concentreront sur l’optimisation, l'interprétation des modèles, et la formulation de recommandations axées sur l'amélioration de la satisfaction des passagers.

Ce rapport démontre un avancement satisfaisant vers l'atteinte des objectifs fixés.

Ce projet vise à fournir des insights exploitables pour les compagnies aériennes, renforçant ainsi leur capacité à satisfaire et fidéliser leurs clients dans un environnement commercial compétitif.