Rapport de Projet | Technique d'Apprentissage Artificiel
Master 1 Informatique Option Big Data

Prédiction de Prix de Billets de Train

Réalisé par

Mourad SEKKAR **Agnès AOUCHICHE**

Encadré par
Rakia DJAZIRI

Année universitaire - 2023/2024

Table des matières

Table des figures

Introduction

1 Contexte du Projet

Le secteur du transport ferroviaire en Espagne connaît une demande croissante, avec une diversité de trajets, de classes et de tarifs proposés par les compagnies ferroviaires. Dans ce contexte, la prédiction précise des prix des billets de train revêt une importance stratégique pour les voyageurs et les entreprises du secteur.

2 Objectif du Projet

L'objectif principal de ce projet est de mettre en place des modèles de prédiction capables d'estimer les prix des billets de train en fonction de diverses caractéristiques telles que l'origine, la destination, le type de train, la classe et d'autres détails temporels. Cette prédiction aidera les voyageurs à planifier leurs déplacements et les compagnies ferroviaires à ajuster leurs stratégies tarifaires.

3 Importation des librairies Python

Les outils essentiels utilisés incluent Pandas pour la manipulation des données, NumPy pour les calculs, Matplotlib et Seaborn pour la visualisation graphique, ainsi que Scikit-learn pour les modèles d'apprentissage automatique. Selon les besoins spécifiques, d'autres librairies ont été intégrées pour des analyses avancées.

4 Aperçu des Données

Les données utilisées contiennent des détails complets sur les trajets en train en Espagne, comprenant les informations sur les itinéraires, les types de trains, les classes et les tarifs, ainsi que les horaires de départ. Ces données sont centrales pour développer des modèles de prédiction des prix des billets de train.

	Unnamed	insert_date	origin	destination	start_date	end_date	train_type	price	train_class	fare
0	702	2019-04-19 05:37:35	PONFERRADA	MADRID	2019-06-02 15:00:00	2019-06-02 19:42:00	MD-AVE	59.50	Turista con enlace	Flexible
1	703	2019-04-19 05:37:35	PONFERRADA	MADRID	2019-06-02 17:15:00	2019-06-02 23:03:00	MD-AVE	34.65	Turista con enlace	Promo +
2	704	2019-04-19 05:37:35	PONFERRADA	MADRID	2019-06-02 17:15:00	2019-06-02 23:10:00	MD-LD	39.95	Turista con enlace	Promo +
3	705	2019-04-19 05:37:35	PONFERRADA	MADRID	2019-06-02 17:15:00	2019-06-02 22:14:00	MD-AVE	40.60	Turista con enlace	Promo +
4	706	2019-04-19 05:37:35	PONFERRADA	MADRID	2019-06-02 18:55:00	2019-06-02 23:03:00	ALVIA	27.90	Turista	Promo

FIGURE 1 – Visualisation des 5 premières lignes du DataSet

Exploration et Préparation des Données

Les données ont été extraites du site internet Kaggle [1].

1 Visualisation des Données

La première étape de l'analyse a consisté à visualiser les distributions des données pour mieux comprendre les origines, destinations et tarifs des trains. Les graphiques obtenus ont permis de mettre en évidence la répartition des trajets selon les différentes catégories, offrant ainsi un aperçu visuel des tendances et de la variabilité des données.

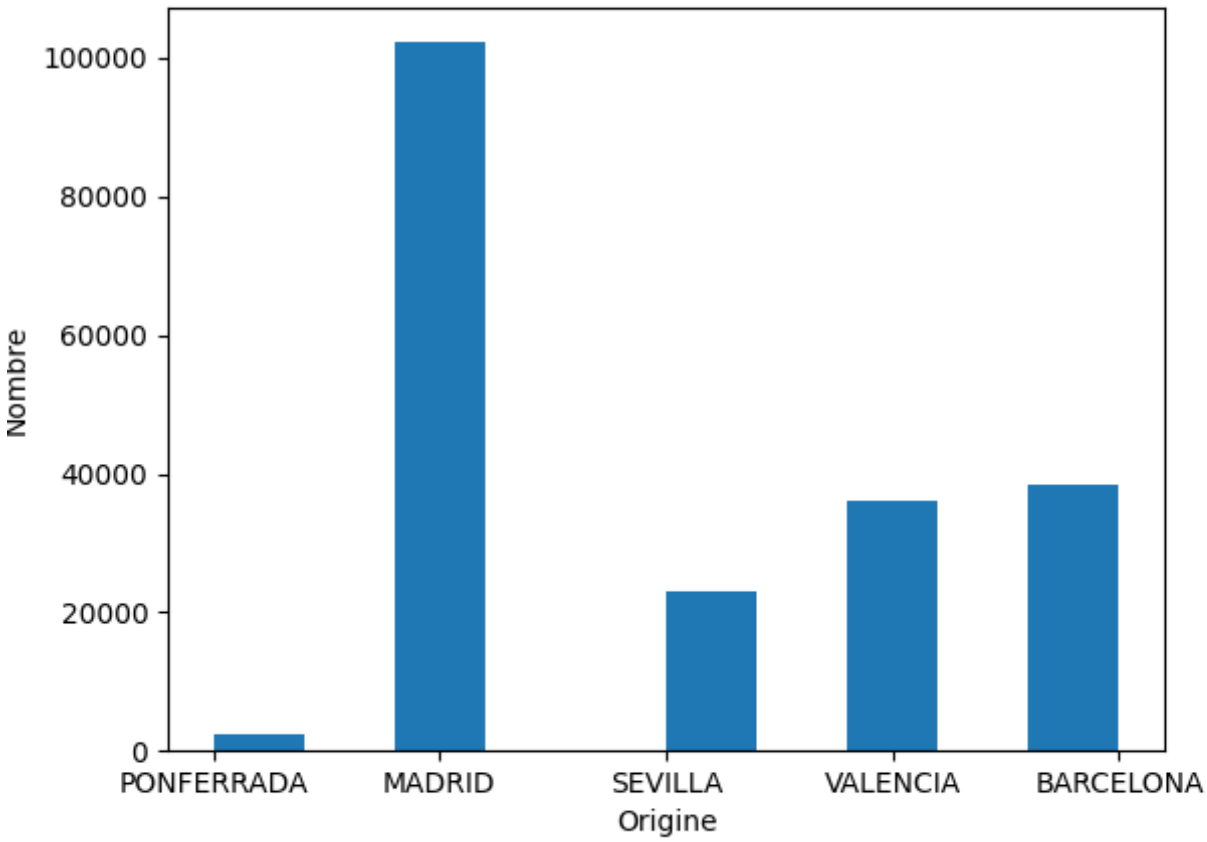

FIGURE 2 – Nombre de trains selon la ville de départ

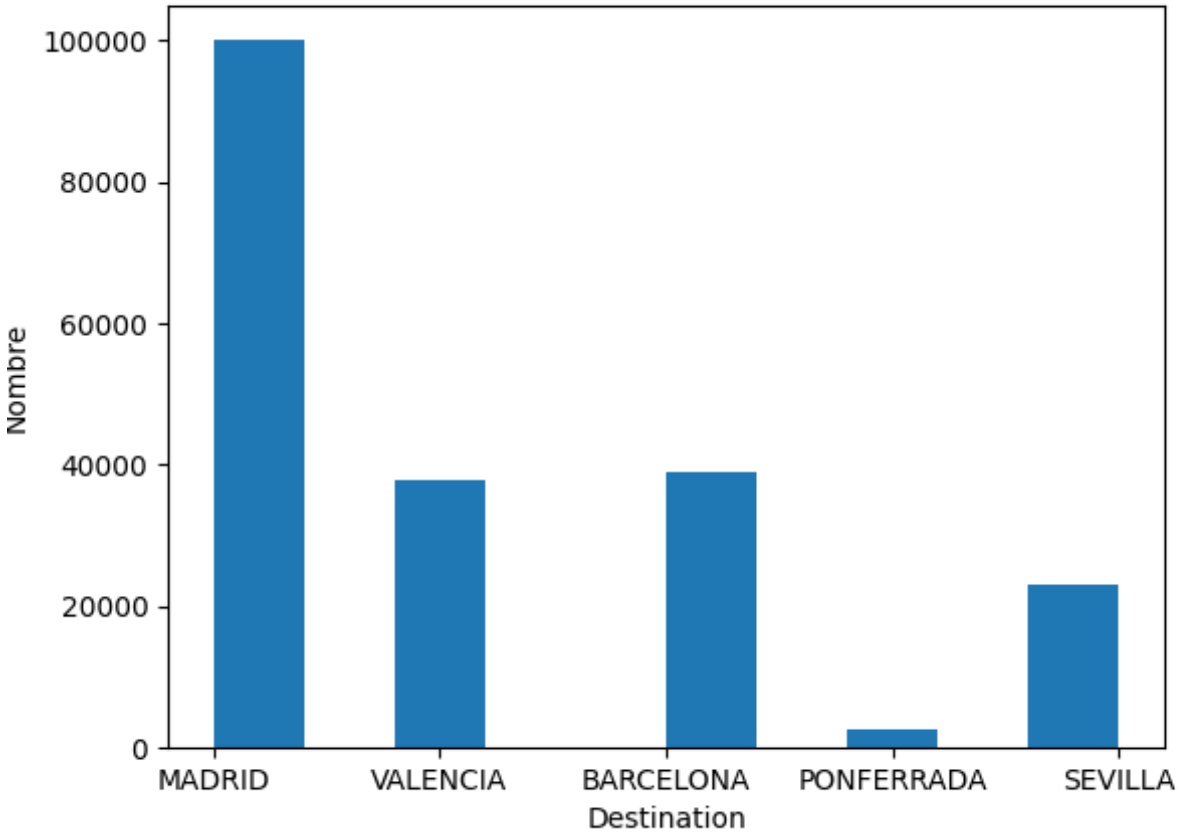

FIGURE 3 – Nombre de trains selon la ville de destination

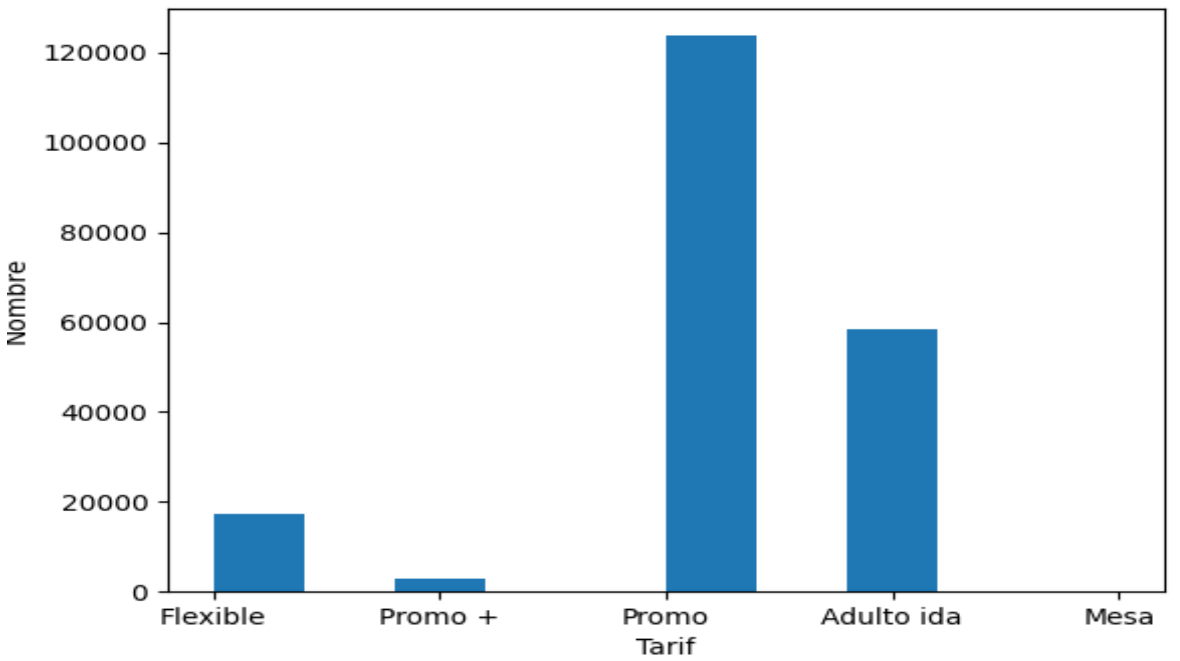

FIGURE 4 – Nombre de trains selon le tarif

2 Nettoyage et Transformation

Une phase cruciale de nettoyage des données a été réalisée pour garantir la qualité des informations utilisées dans la modélisation. Cela comprenait la suppression des colonnes inutiles, la gestion des valeurs manquantes, ainsi que la transformation des données temporelles et la conversion des variables catégoriques en formats adaptés pour l'entraînement des modèles de prédiction. Les données résultantes à ce traitement ont été enregistrées dans une nouveau fichier CSV.

3 Analyse de Corrélation

Une analyse de corrélation approfondie a été entreprise pour explorer les relations entre les différentes caractéristiques des trajets de trains.

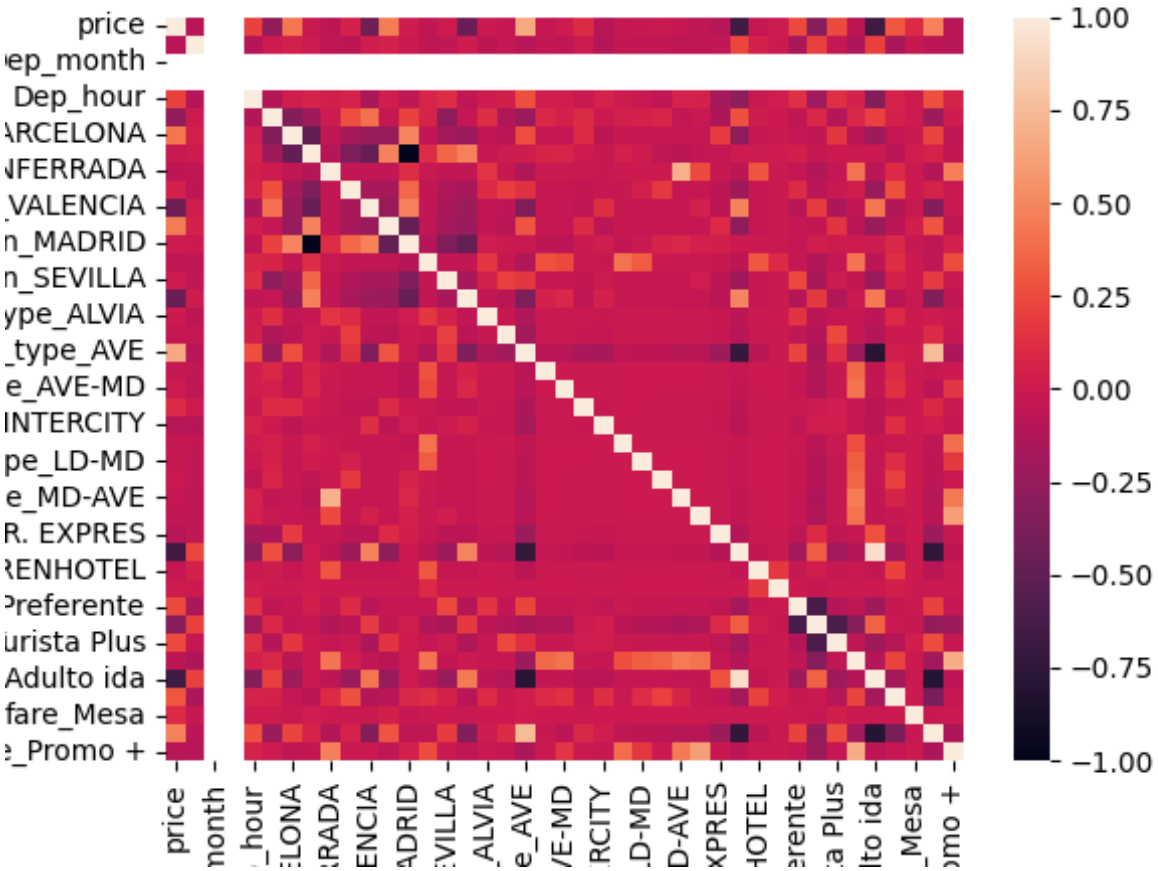

FIGURE 5 – Tableau de corrélation des données du DataSet

D'après le tableau de corrélation, voici quelques conclusions principales tirées :

— Il y a une forte corrélation positive (0.65) entre le prix (price) et le type de train AVE, ce qui indique que les trains AVE sont généralement plus chers.

— Le prix est négativement corrélé avec le type de train Régional (-0.65), suggérant que ces trains sont moins chers.

— Il y a une corrélation négative importante (-0.78) entre le prix et le tarif Adulte aller simple (Adulto ida), ce qui est logique car ce tarif est moins cher.

— Le prix est positivement corrélé avec la classe Préférente (0.24) et la classe Touriste Plus (0.25). Ces classes plus chères ont donc tendance à être associées à des prix de billets plus élevés.

— Certaines origines et destinations spécifiques sont aussi corrélées positivement ou négativement avec le prix, reflétant probablement le fait que certains trajets sont plus chers.

En résumé, le tableau montre bien que le prix du billet est lié de manière logique au type de train, à la classe et au tarif choisis par le passager. Les correlations reflètent ces relations.

4 Classification - Random Forest

4.1 Analyse de la distribution des prix

Dans un premier temps, la distribution des prix dans le jeu de données a été analysée à l'aide d'un histogramme. Ceci a permis d'observer visuellement la plage de variation des prix et la concentration de valeurs sur certains intervals.

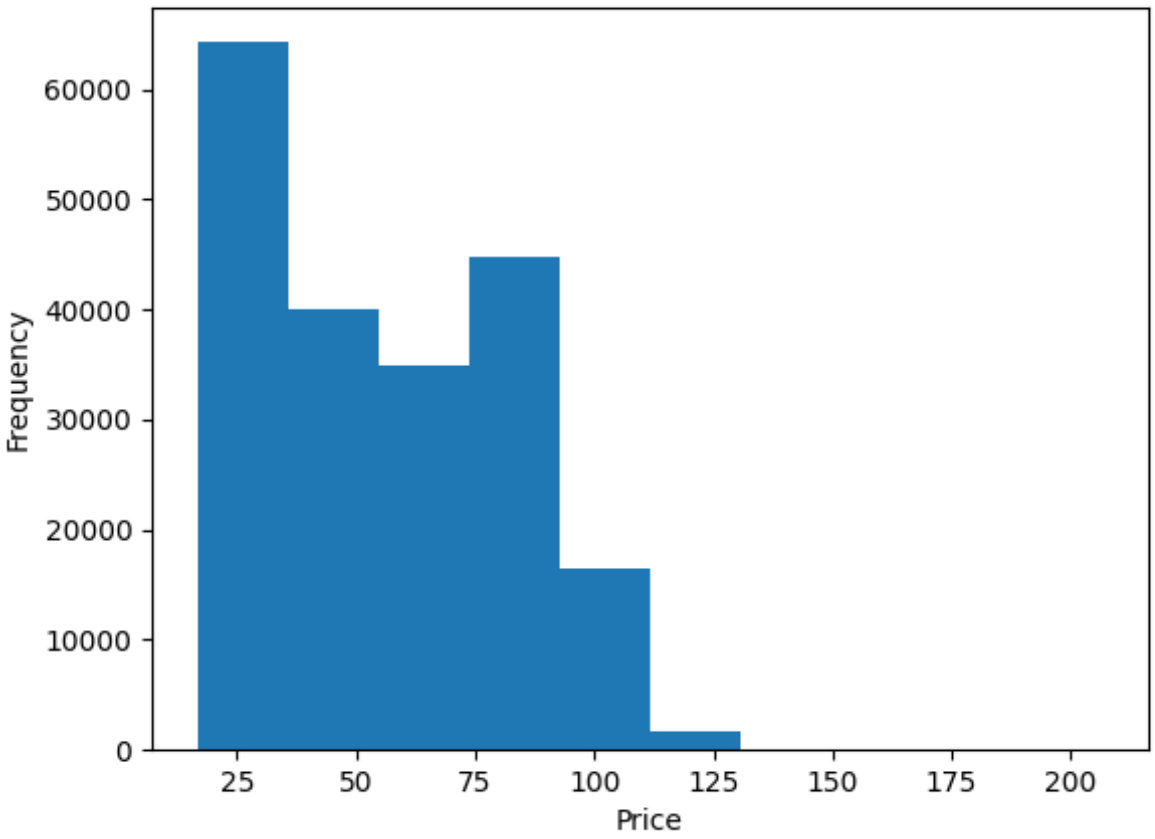

FIGURE 6 – Histogramme de distribution des prix des billets

4.2 Discrétisation des prix en classes

Sur la base de cette analyse, la variable quantitative "prix" a été discrétisée en 4 classes :

— Bas : moins de 50€

— Moyen : entre 50€ et 100€

— Élevé : plus de 100€

Un nouvel attribut catégoriel "price_category" a ainsi été créé pour permettre la classification. Les bornes de classes ont été déterminées manuellement en analysant la distribution des prix.

4.3 Modélisation Random Forest

Après préparation des données, un modèle de classification Random Forest (forêt d'arbres décisionnels) a été entrainé pour prédire la catégorie de prix des billets.

L'algorithme Random Forest combine un grand nombre d'arbres de décision, chaque arbre étant entrainé sur un sous-ensemble aléatoire des données et avec un sous-ensemble aléatoire de features. Leur prédiction finale est aggregée pour déterminer la classe majoritaire. Après entrainement sur 80% des données, le modèle atteint une précision de 98% sur l'ensemble de test. L'algorithme arrive donc à prédire correctement la catégorie de prix dans 98% des cas, ce qui est satisfaisant pour notre cas d'usage.

En conclusion, l'approche Random Forest est un premier résultat prometteur pour catégoriser automatiquement le prix des billets sur la base des autres caractéristiques.

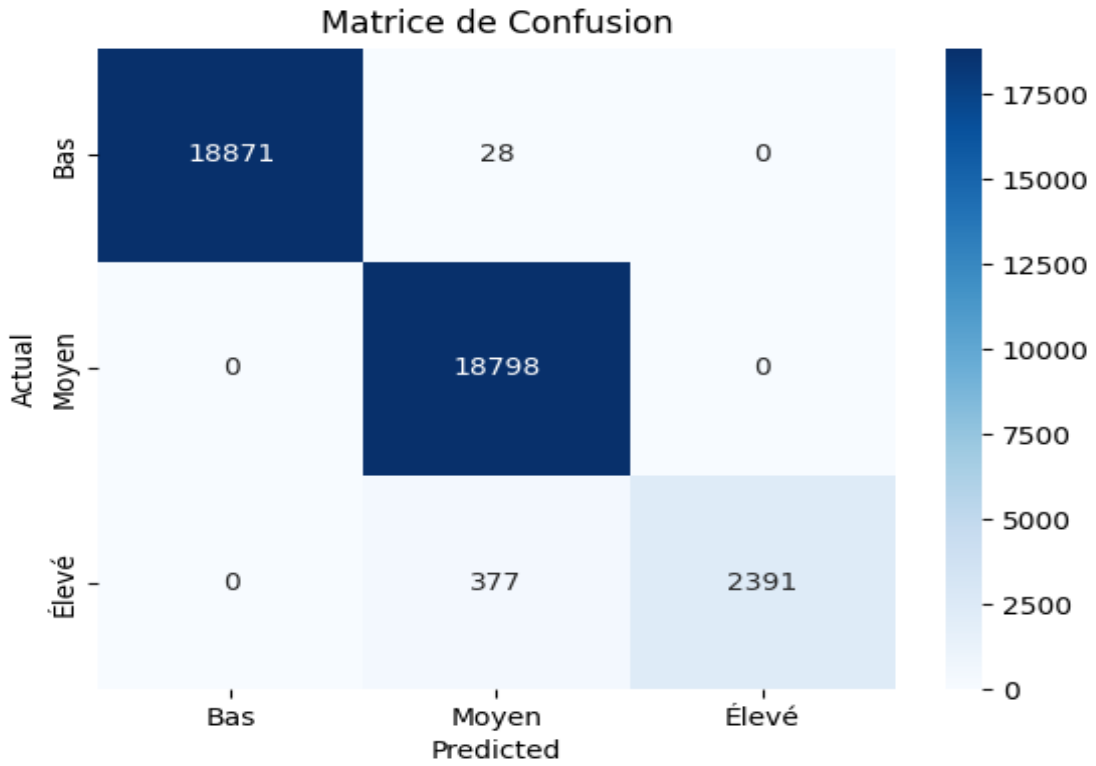

FIGURE 7 – Matrice de confusion

Les données sont maintenant prêtes pour la modélisation. Chaque classe (catégorie de prix) va être entrainée avec un modèle différent.

Modélisation et Entraînement

1 Division des Données

Les données ont été divisées en ensembles distincts pour l'entraînement et le test des modèles. Cela a été effectué avec une fonction qui prend en paramètre les données ainsi que la catégorie de prix visée par le traitement à effectuer.

2 Différents Modèles

Plusieurs modèles ont été explorés pour prédire les prix des billets de train. Parmi ceux-ci :

2.1 Régression Linéaire

Une analyse de régression linéaire a été réalisée pour la catégorie de prix "Bas" afin de prédire les tarifs des billets de train. Les étapes suivantes on été entreprises pour élaboréer le modèle et évaluer ses performances :

2.1.1 Entraînement du Modèle

Un modèle de régression linéaire a été créé en utilisant les données d'entraînement correspondant à cette catégorie de prix. Les caractéristiques pertinentes ont été sélectionnées pour prédire de manière optimale les tarifs des billets de train.

2.1.2 Évaluation du Modèle

Le modèle de régression linéaire a été évalué sur l'ensemble de test pour estimer sa performance. Les métriques suivantes ont été calculées :

— Coefficient de détermination (R^2) : 0.663

— Erreur quadratique moyenne (MSE) : 19.399

— Erreur quadratique moyenne racine (RMSE) : 4.404

le modèle semble expliquer raisonnablement bien la variance des prix des billets 'BAS' avec un R^2 de 0.663. Cependant, les erreurs moyennes (MSE et RMSE) indiquent qu'il y a encore une marge d'amélioration pour affiner les prédictions et réduire l'écart entre les valeurs prédites et réelles.

2.1.3 Visualisation des Résultats

Pour illustrer les performances du modèle, une dispersion des prédictions par rapport aux valeurs réelles a été tracée.

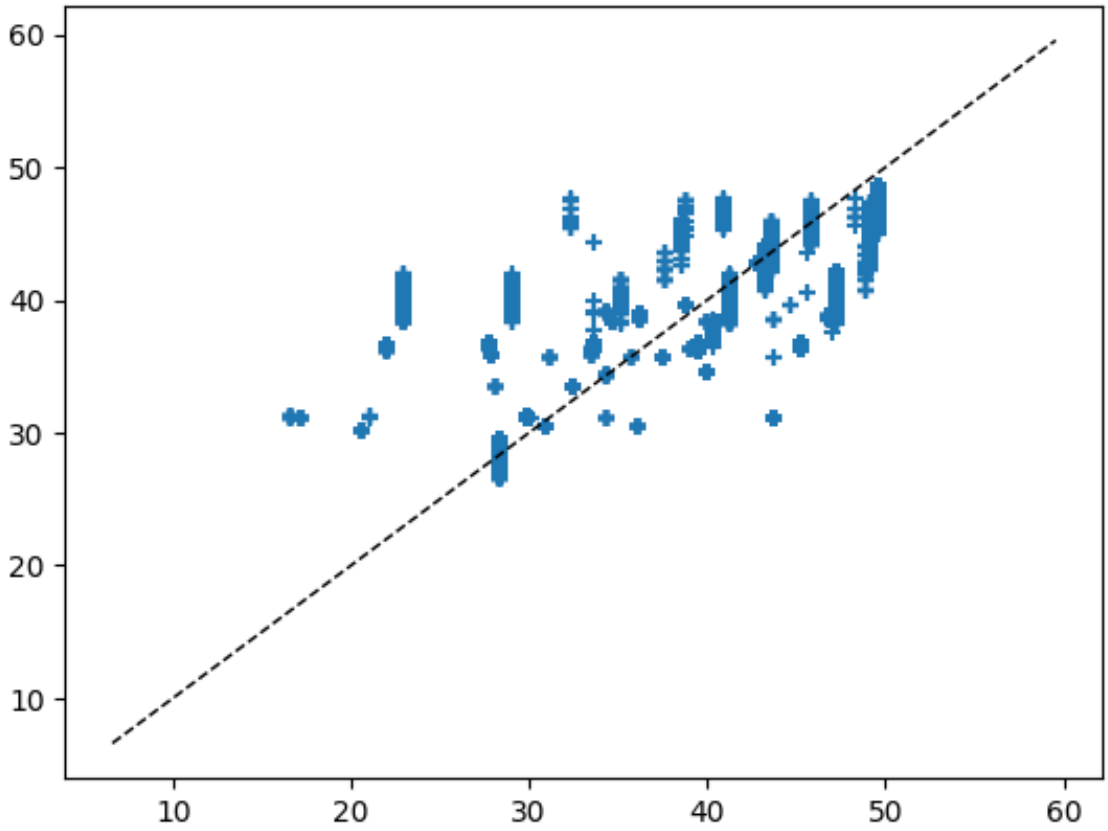

FIGURE 8 – Dispersion des prédictions de la régression linéaire

2.2 Arbre de Décision (CART)

Pour prédire les prix des billets de train dans la catégorie "Moyen", un modèle d'Arbre de Décision a été construit et évalué.

2.2.1 Résultats

L'arbre ainsi généré a une profondeur limitée à deux niveaux pour une visualisation aisée des décisions prises par le modèle.

Les résultats obtenus pour ce modèle montrent :

— Coefficient de détermination (R^2) : 0.974

— Erreur quadratique moyenne (MSE) : 3.892

— Erreur quadratique moyenne racine (RMSE) : 1.972

Ces métriques évaluent la performance du modèle d'Arbre de Décision pour la prédiction des prix des billets de train dans la catégorie "Moyen". Ces résultats constituent une étape initiale pour évaluer l'adéquation du modèle à ces données spécifiques.

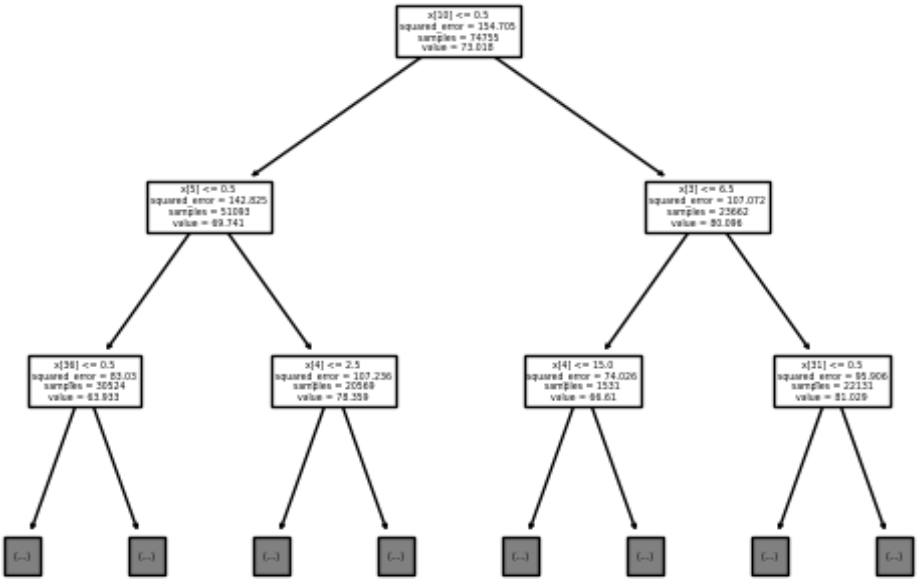

FIGURE 9 – Arbre de CART

2.3 Random Forest

Pour prédire les prix des billets de train dans la catégorie "Élevé", un modèle de Random Forest a été utilisé. Les Random Forests sont des modèles d'ensemble basés sur des arbres de décision multiples.

2.3.1 Évaluation du Modèle

Le modèle de random forest a été évalué sur l'ensemble de test pour estimer sa performance. Les métriques suivantes ont été calculées :

— Coefficient de détermination (R^2) : 0.999

— Erreur quadratique moyenne (MSE) : 0.093

— Erreur quadratique moyenne racine (RMSE) : 0.306

2.3.2 Visualisation des Résultats

Pour illustrer les performances du modèle, la figure suivant :

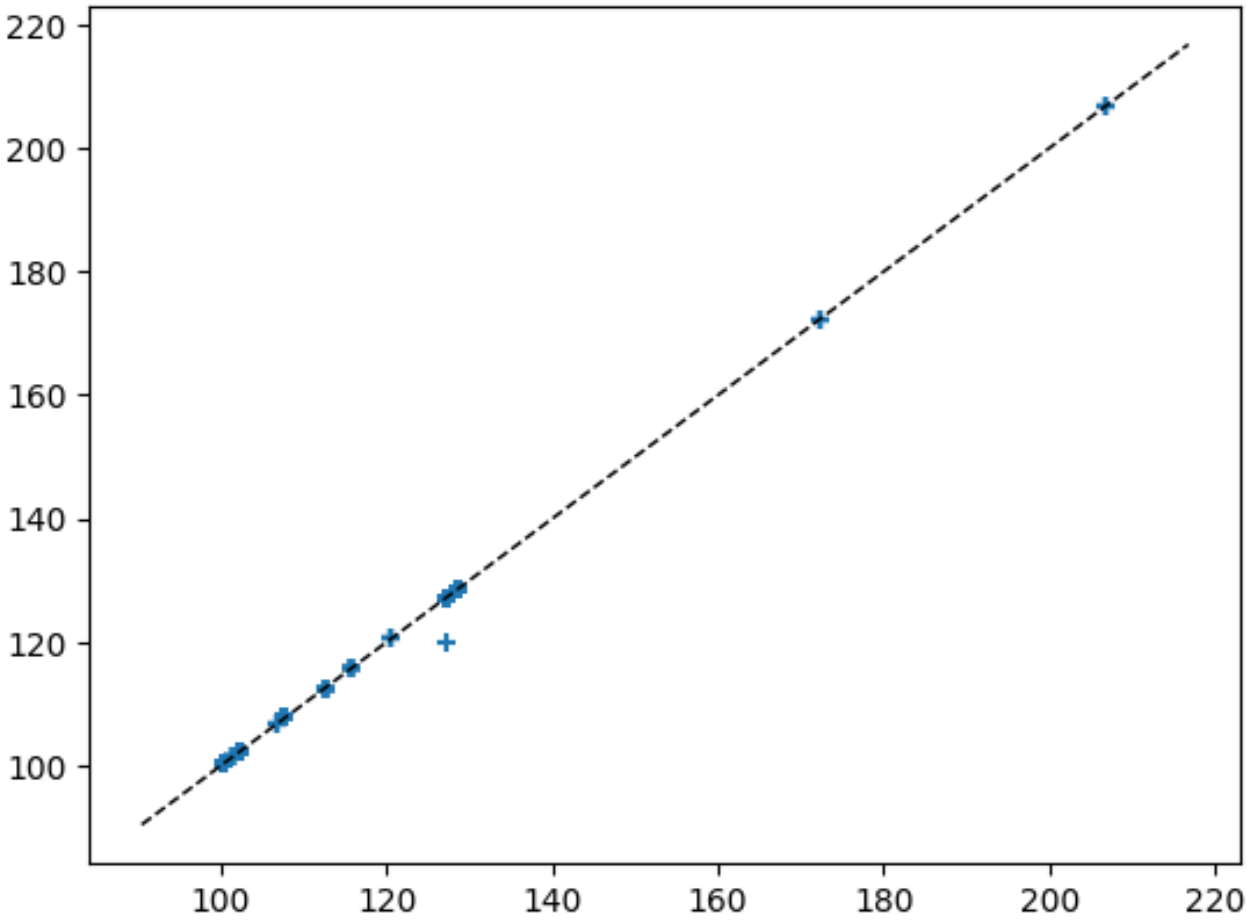

FIGURE 10 – Résultats du modèle Random Forest sur les billets (catégorie prix : élevé)

Résultats et Conclusions

1 Résultats des Modèles

Les performances des différents modèles ont varié en termes de précision, d'erreur quadratique moyenne (RMSE), de coefficient de détermination (R^2), et d'autres métriques d'évaluation.

Modèle	R^2	MSE	RMSE
Regression Linéaire	0.663	19.399	4.404
Arbre de Décision	0.974	3.892	1.972
Random Forest	0.999	0.093	0.306

TABLE 1 – Performances des Modèles

2 Interprétation des Résultats

— **Régression Linéaire :** Son coefficient de détermination (R^2) de 0.663 indique que ce modèle explique environ 66% de la variance des données. Cela signifie qu'il parvient à prédire raisonnablement bien les prix des billets de train, mais il reste de la variance non expliquée. L'erreur quadratique moyenne (MSE) de 19.399 et la racine de l'erreur quadratique moyenne (RMSE) de 4.415 indiquent que les prédictions ont une dispersion plus importante par rapport aux valeurs réelles.

— **Arbre de Décision :** Avec un R^2 de 0.972, cet arbre de décision capture une grande partie de la variance des données. Il parvient à expliquer près de 97% de la variance, ce qui suggère une très bonne adéquation du modèle aux données. Les valeurs MSE de 3.892 et RMSE de 1.972 montrent que les prédictions sont généralement proches des valeurs réelles, avec une moindre dispersion que la régression linéaire.

— **Random Forest :** Le coefficient de détermination (R^2) proche de 1 (0.999) suggère que ce modèle explique quasiment toute la variance des données, indiquant une adéquation exceptionnelle du modèle aux données d'entraînement. Les valeurs très basses de MSE et de RMSE confirment la grande précision des prédictions, avec une dispersion très faible par rapport aux valeurs réelles.

Ces comparaisons des modèles fournissent des indications clés :

— **Complexité et performance :** Les modèles plus complexes (l'arbre de décision et Random Forest) ont montré une meilleure adéquation aux données et une précision plus élevée dans leurs prédictions.

— **Adaptabilité :** L'arbre de décision a montré une très bonne performance, capturant une grande partie de la variance des données. Cependant, le Random Forest, grâce à sa nature d'ensemble d'arbres, a surpassé l'arbre de décision en offrant une précision encore plus élevée, réduisant considérablement l'erreur dans les prédictions.

— **Importance de la complexité :** Les modèles plus complexes, tels que l'arbre de décision et le Random Forest, semblent mieux convenir à cette tâche de prédiction de prix de billets de train en raison de leur capacité à capturer les relations non linéaires entre les caractéristiques et les prix des billets.

En conclusion, pour cette tâche de prédiction de prix de billets de train, le Random Forest semble être le choix le plus performant en termes de précision.

3 Limitations

Les performances élevées de l'Arbre de Décision et du Random Forest pourraient suggérer un surajustement (overfitting) du modèle aux données d'entraînement, nécessitant une évaluation plus poussée sur des données inconnues.

Pour la régression linéaire, l'erreur relativement élevée pourrait être due à la simplicité du modèle, qui peut ne pas capturer de manière optimale les relations non linéaires entre les caractéristiques et les prix des billets de train.

Les résultats exceptionnellement bons du Random Forest pourraient indiquer une forte sensibilité aux données d'entraînement spécifiques, nécessitant une évaluation approfondie sur différents ensembles de données pour vérifier sa généralisation.

4 Conclusion

En résumé, cette étude a démontré la faisabilité de prédire les prix des billets de train en Espagne à partir de différentes caractéristiques des trajets. Les modèles ont montré des performances variables, avec des résultats prometteurs pour certains, mais avec des limitations à considérer.

Application de Prédiction de prix de billets de train

L'interface graphique développée permet aux utilisateurs de saisir diverses informations telles que la date et l'heure de départ, l'origine et la destination du trajet, le type de train, la classe et le tarif. Ces variables sont utilisées pour prédire le prix du billet de train. L'interface offre une manière conviviale et intuitive de fournir ces données et fournit en retour une estimation du prix du billet en fonction des entrées spécifiées.

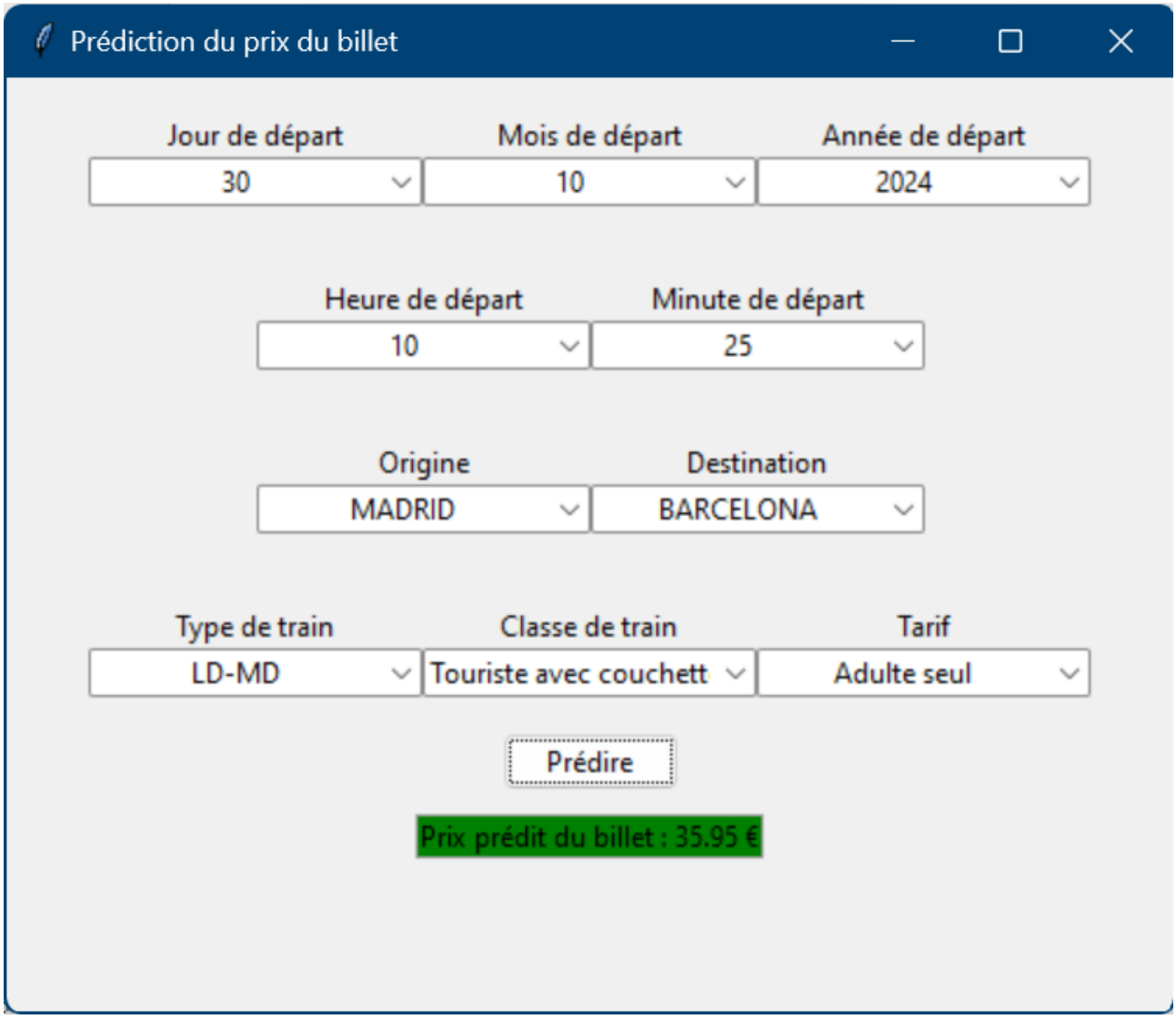

FIGURE 11 – Interface graphique : Prédiction de prix de billets de train

Perspectives Futures

1 Améliorations Possibles

Pour améliorer la précision des prédictions, des ajustements pourraient être apportés aux modèles existants en incluant de nouvelles variables pertinentes ou en utilisant des techniques avancées de prétraitement des données. Une optimisation des hyperparamètres des modèles pourrait également être entreprise pour obtenir de meilleures performances.

2 Extensions du Projet

Pour élargir ce projet, l'ajout de données supplémentaires telles que les conditions météorologiques, les événements spéciaux ou les niveaux de demande pendant certaines périodes pourraient enrichir les modèles et améliorer leur capacité à prédire les prix des billets de train avec précision. L'exploration d'autres méthodes de prédiction comme les réseaux de neurones ou les méthodes de traitement de langage naturel (NLP) pourrait également être envisagée pour une analyse plus approfondie et des résultats plus précis.

Bibliographie

[1] Kaggle. Spanish rail tickets pricing - renfe. Consulté le 9 décembre 2023.