ISG Bizerte : Institut Supérieur de Gestion de Bizerte

Rapport de Projet : Analyse du Dataset Titanic avec Machine Learning

Superviseur : [Olfa Chebbi]

Hanin Hedhli & Hadir Dridi & Ella Soussi & Jihen Souissi
02/05/2025

Table des Matières

1. Introduction

Le dataset Titanic est un ensemble de données bien connu dans le domaine de la science des données, souvent utilisé pour pratiquer les techniques de machine learning. Ce projet, réalisé à l'Institut Supérieur de Gestion de Bizerte, vise à appliquer des algorithmes de machine learning pour prédire la survie des passagers en fonction de caractéristiques telles que l'âge, le sexe, la classe, et plus encore. Ce dataset a été choisi pour sa documentation complète, ses caractéristiques variées et son problème clair de classification binaire (survie ou non-survie).

Pourquoi le Dataset Titanic ?

Le dataset Titanic a été sélectionné pour les raisons suivantes :

- **Bien documenté :** De nombreuses informations sont disponibles sur Kaggle et d'autres plateformes.
- **Caractéristiques variées :** Inclut des attributs comme l'âge, le sexe, la classe des passagers et le port d'embarquement, permettant une analyse riche.
- **Classification binaire simple :** Le résultat de survie (0 = décédé, 1 = survivant) est facile à comprendre, idéal pour apprendre les concepts de machine learning.

2. Méthodologie

Le projet suit un flux de travail structuré de machine learning, incluant la configuration de l'environnement, la préparation des données, l'entraînement des modèles, leur évaluation et la visualisation.

2.1. Configuration de l'Environnement

L'environnement de travail a été configuré comme suit :

1. **Python et Jupyter Notebook :** Python a été installé avec Jupyter Notebook pour un codage interactif.
2. **Bibliothèques installées :**
    - **pandas :** Pour la manipulation des données et la gestion des fichiers CSV.
    - **scikit-learn :** Pour les algorithmes de machine learning (arbres de décision, SVM, KNN).
    - **matplotlib et seaborn :** Pour la visualisation des données, y compris les matrices de confusion.

- **jupyter :** Pour créer des notebooks interactifs (optionnel mais recommandé). Commande utilisée dans le terminal :

```
PS C:\Users\dridi> python -m pip install pandas scikit-learn matplotlib seaborn jupyter notebook
```

2.2. Préparation des Données

Un dossier de projet, Projet_Titanic, a été créé pour organiser le dataset et les notebooks. Les datasets utilisés incluent :

- **train.csv** : Contient 891 enregistrements de passagers avec leurs caractéristiques (PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked) et leur statut de survie.
- **test.csv** : Contient 418 enregistrements de passagers avec les mêmes caractéristiques sauf Survived, utilisé pour tester les modèles.
- **gender_submission.csv** : Un fichier de soumission d'exemple prédisant la survie selon une règle simple (les femmes survivent, les hommes non).

Étapes de prétraitement des données :

- **Valeurs manquantes :** Les valeurs manquantes dans **Age** et **Fare** ont été remplies avec leurs médianes respectives, et **Embarked** avec le mode.
- **Encodage :** Les variables catégoriques **Sex** et **Embarked** ont été converties en valeurs numériques à l'aide de **LabelEncoder.**

2.3. Entraînement des Modèles

Trois algorithmes de classification ont été implémentés avec scikit-learn :

- **Arbre de Décision :** DecisionTreeClassifier(random_state=42).
- **Machine à Vecteurs de Support (SVM) :** SVC(kernel="linear", random_state=42).
- **K-Nearest Neighbors (KNN) :** KNeighborsClassifier(n_neighbors=3).

Le dataset a été divisé en 80 % pour l'entraînement et 20 % pour le test à l'aide de train_test_split avec test_size=0.2 et random_state=42. Chaque modèle a été entraîné sur l'ensemble d'entraînement.

2.4. Évaluation des Modèles

Les performances des modèles ont été évaluées à l'aide de :

- **Précision :** Calculée avec accuracy_score.
- **Matrice de Confusion :** Visualisée avec des heatmaps seaborn pour montrer les vrais positifs, faux positifs, vrais négatifs et faux négatifs.

3. Résultats

Les performances des modèles sont résumées ci-dessous.

3.1. Arbre de Décision

- **Matrice de Confusion :**
  - Vrais Positifs (Décédé) : 83
  - Faux Positifs : 18
  - Faux Négatifs : 56
  - Vrais Négatifs (Survivant) : 22
- **Précision :** Décédé : 82.2 %, Survivant : 28.2 %
- **Rappel :** Décédé : 59.7 %, Survivant : 55 %
- **Précision Globale :** 78 %

3.2. SVM

- **Matrice de Confusion :**
  - Vrais Positifs (Décédé) : 99
  - Faux Positifs : 16
  - Faux Négatifs : 23
  - Vrais Négatifs (Survivant) : 41
- **Précision Globale :** Environ 78 % (estimée à partir de la matrice de confusion).

3.3. KNN

- **Matrice de Confusion :** Non détaillée complètement, mais précision fournie.
- **Précision Globale :** 72.63 %
- **Précision et Rappel :**
  - Décédé : Précision 70 %, Rappel 92 %
  - Survivant : Précision 80 %, Rappel 45 %.

3.4. Comparaison

Un graphique en barres a été généré pour comparer les précisions :

- Arbre de Décision : 78 %
- SVM : ~78 % (estimée)
- KNN : 72.63 %

Les modèles Arbre de Décision et SVM ont surpassé KNN, l'Arbre de Décision offrant un équilibre entre précision et rappel pour la classe des décédés.

4. Discussion

Ce projet a permis de démontrer avec succès l'application du machine learning au dataset Titanic. Les principaux enseignements incluent :

- **Expérience Pratique :** Manipulation pratique des données, entraînement et évaluation des modèles.
- **Comparaison des Algorithmes :** Les Arbres de Décision et SVM ont montré une meilleure précision que KNN, probablement en raison d'une meilleure gestion de l'espace des caractéristiques du dataset.
- **Importance du Flux de Travail :** Le processus structuré de chargement, prétraitement, division, entraînement et évaluation est crucial pour des résultats fiables.
- **Visualisation :** Les matrices de confusion ont offert des insights clairs sur les performances des modèles, mettant en évidence les erreurs de classification.

5. Conclusion

Ce projet a offert une introduction complète au machine learning avec le dataset Titanic. En implémentant les Arbres de Décision, SVM et KNN, nous avons atteint des précisions allant de 72.63 % à 78 %. Cette expérience a souligné l'importance du prétraitement des données et de l'évaluation des modèles pour obtenir des prédictions significatives. Les travaux futurs pourraient explorer l'ingénierie des caractéristiques ou des méthodes d'ensemble pour améliorer les performances.

6. Références

- **Dataset Titanic de Kaggle :** https://www.kaggle.com/c/titanic
- **Documentation Scikit-learn :** https://scikit-learn.org/stable/
- **Documentation Pandas : https:** //pandas.pydata.org/