Tests d’Hypothèse

Les femmes et les enfants d’abord

Les enfants ont été privilégiés lors du naufrage

Question : Les enfants ont-ils été privilégiés lors du naufrage.

Hypothèse H0 : les enfants n'ont pas été privilégiés.

Hypothèse H1 : les enfants ont été privilégiés.

Test utilisé : comparaison de moyennes de 2 échantillons (Test de Student).

Conditions : les données suivent une loi Normale, le nombre d'échantillons > 30 : ok

Etapes :

- Sélection des données à comparer:

- Ages des enfants.

- Ages de l’ensemble des passagers.

- Nettoyage des données.

- Application du Test de Student

Conclusion :

La p-value est égale à 3.907198e-56, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et conclure, que les moyennes des enfants ayant été sauvés et de toutes les personnes sauvées ne sont pas identiques.

Les femmes ont été privilégiées lors du naufrage

Question : Les femmes ont-elles été privilégiées lors du naufrage.

Hypothèse H0 : les femmes n'ont pas été privilégiées.

Hypothèse H1 : les femmes ont été privilégiées.

Test utilisé : Test de comparaison à une proportion (Test de Pearson).

Etapes :

- Sélection des données à comparer:

- Nombre de femme sauvés.

- Total des passagers.

- Nettoyage des données.

- Application du Test de Pearson.

Conclusion :

La p-value est égale à 1.005984e-11, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et conclure, que les femmes ont été privilégiées lors du naufrage.

Influence du prix du billet

Le prix du billet a une influence sur la survie d’un passager

Question : Le prix du billet a une influence sur la survie d'un passager.

Hypothèse H0 : Le prix du billet a une influence.

Hypothèse H1 : Le prix du billet a une influence.

Test utilisé : comparaison de moyennes de 2 échantillons (Test de Student).

Conditions : les données suivent une loi Normale, le nombre d'échantillons > 30 : ok

Etapes :

- Sélection des données à comparer:

- Prix du billet des survivants.

- Prix du billets de l’ensemble des passagers.

- Nettoyage des données.

- Application du Test de Student

Conclusion :

La p-value est égale à 3.632247e-08, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et on peut conclure que, le prix du billet a une influence sur la survie d'un passager.

Régression linéaire

Régression Logistique

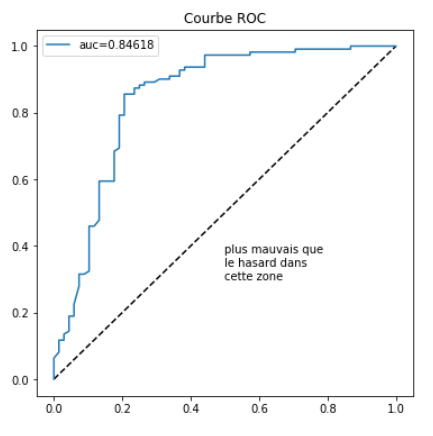
L’objectif est de trouver un modèle qui nous permettra de prédire si un passager est mort ou vivant en se basant sur les données du dataset.

Etapes :

* Importation des données : read.csv
* Analyse des données :
  + On constate qu’il y a beaucoup de valeurs manquantes dans la colonne Age et Cabin.
  + 5 colonnes sur les 12 sont qualitatives.
* Choix des colonnes :
  + Choix de la Target y et des data X :
    - Y : l’objectif de l’exercice est de prédire, si un passager est mort ou vivant. Nous avons sélectionné la Colonne Survived.
    - X : Le reste des features.
  + Suppression des faetures de X non pertinentes pour notre analyse :
    - Cabin : Trop de valeurs manquantes et valeurs uniques.
    - PassengerId : Valeurs uniques.
    - Name : Valeurs uniques.
    - Ticket : Valeurs uniques.
* Nettoyage des données :
  + Colonne Age : En se basant sur la moyenne et l’écart type des valeurs de la colonne (15,30,45). Nous avons remplacé les valeurs manquantes, par tranches.
  + Colonnes Sex et Embarked : Transformation des valeurs qualitatives en quantitatives, via LabelEncoder de Sklearn.
  + Utilisation de RFECV de Sklearn pour terminer la sélection.
* Entrainement du modèle : LogisticRegression (algorithme imposé).
* Evaluation du modèle :
  + Score : 82% résultat correct, qui pourrait être optimisé.
  + Matrice de confusion :

|  |  |  |
| --- | --- | --- |
|  | Predis Vrai | Predis Faux |
| Actuels Vrai | 99 | 12 |
| Actuels Faux | 19 | 49 |

* + Jutesse (Accuracy score)  : 0.84



* Application du modèle au données test fournit et exports du résultat en CSV.