Explication des variables :

1. Explication du jeu de données

La base de données Titanic est une base de données constituée à partir des données collectées des passagers qui sont montés au bord du célèbre paquebot transatlantique britannique Titanic. Le but dans ce jeu de données est de prédire, à partir des différentes variables, si la personne a survécu ou non.

1. Visualisation des données :

Nous remarquons en premier lieu que les variables sont de différentes types. Nous y trouvons des données numériques, qui sont discrètes (« Age », « Sex », etc.) ou continues « Fare », et des données catégorielles *(« Survived »)*. L’onglet *Summary* nous donne une idée sur les différentes mesures statistique de chaque variable : la médiane, l’écart-type, etc.

1. Analyse Unidimensionnelle

Pour l’analyse unidimensionnelle, nous avons, pour les variables à valeurs numériques, calculer le Box Plot, Le Box Plot Interactif, le Pie et l’histogramme. Pour la variable catégorielle « Survived », nous avons tracé un Barplot à la place de l’histogramme. D’après l’analyse des différents graphes, nous pouvons conclure que :

1. L’âge de la plupart des passagers entre dans la catégorie 1, suivie par la catégorie 2 ;
2. Nous avons plus d’hommes que de femmes ;
3. La majorité des passagers n’ont pas survécu ;
4. Le prix des billets tournaient autour de 14.45 Unité (médiane), avec comme prix max 512.3 Unité ;
5. La majorité des passagers ont embarqué par la porte 0 ;
6. La majorité des passagers étaient en class 3 ;
7. La majorité des passagers n’étaient pas accompagnés par un membre de leur famille.
8. Analyse Bidimensionnelle

Bien que nécessaire, l’analyse unidimensionnelle nous ne donne pas d’informations sur la relation qui existent entre les différentes variables. Pour cela, nous avons utilisé le nuage de point, entre deux variables ainsi que la table de corrélation. Nous avons pu conclure que :

1. Il existe une forte corrélation entre la variable Sex et Survived. Une femme a en effet plus de chances de survivre qu’un homme (74% vs 18%) ;
2. Les passagers dont l’âge est dans la catégorie 4 ont beaucoup moins de chances de survivre, contrairement aux passagers dont l’âge est dans la catégorie 1 ;
3. Il existe une faible corrélation entre la porte d’embarquement et la variable Survived ;
4. Il existe une forte corrélation entre la variable Survived et la classe. En effet, ceux qui sont en classe 3 ont moins chances de survivre que ceux qui sont en classe 1.
5. Il existe une forte corrélation entre la varibale Title et Survived. En effet, nous remarquons que les passagers de Title 3 ont plus de chances de survivre et que les passagers de Titre 1 ont moins de chances de survivre ;
6. Que les passagers accompagnés par au moins un membre de leur famille ont plus de risque de se noyer.
7. Clustering

En se basant sur les différentes variables, en en convertissant la variable catégorielle Survived en une variable numérique discrète, nous divisons notre base de données en deux sous-ensembles : entraînement et test. Nous utilisons 4 algorithmes de Clustering :

* 1. KNN : le meilleur résultat est obtenu en choisissant 14 voisons avec un taux de bonne prédiction de 65.73% ;
  2. Logistic Regression : un taux de bonne prédiction 79.91% ;
  3. SVM : le meilleur résultat a été obtenu avec le kernel linear avec un taux de bonne prédiction de 81.36% ;
  4. Réseaux de Neurones : le meilleur résultat obtenu avec 74 neurones dans la couche cachée avec un taux de bonne prédiction d’environ 85.00%.

Les meilleurs algorithmes sont le SVM et le Réseau de Neurones.