Titanic Survival Prediction



Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most tragic maritime disasters in history. This iconic event claimed the lives of over 1,500 passengers and crew members, highlighting the critical importance of safety measures and emergency preparedness at sea.

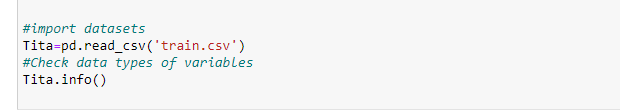
Project Overview

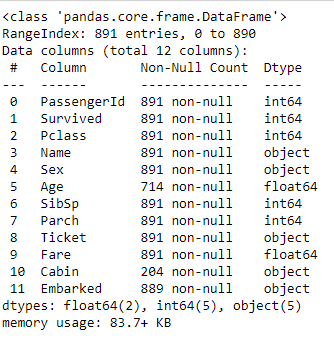
The Titanic Survival Prediction Project is a data-driven endeavor aimed at developing a predictive model that can determine the likelihood of a passenger surviving the Titanic disaster based on various attributes and features. By analyzing the available dataset, which includes information about passengers' age, gender, class, cabin, ticket fare, and more, we aim to uncover patterns and insights that can shed light on the dynamics of survival during this tragedy.

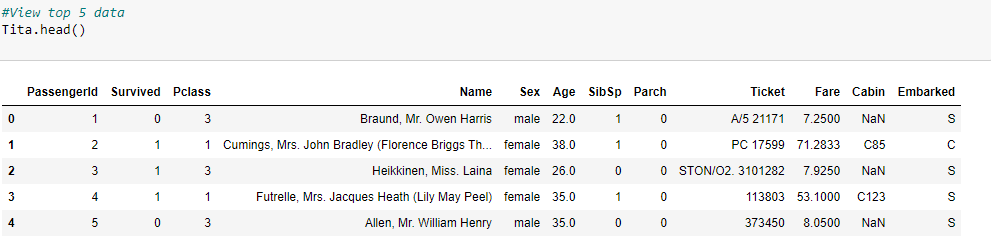
Objectives

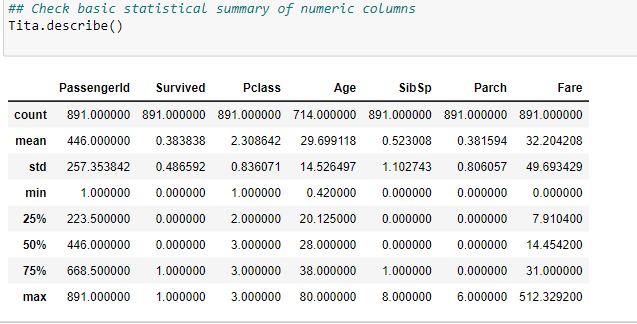
1. Predictive Modeling: The primary objective of this project is to build an accurate predictive model.
2. Feature Analysis: We will conduct a thorough analysis of the dataset to identify which features or characteristics played a significant role in determining a passenger's likelihood of survival.
3. Data Visualization: Visualization techniques will be employed to create informative graphs and plots, enabling us to present the data in an easily digestible manner

1.Importing Required Libraries

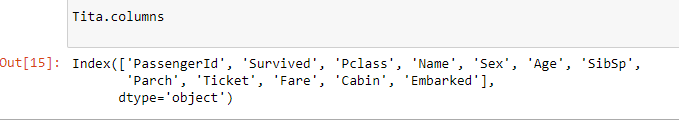


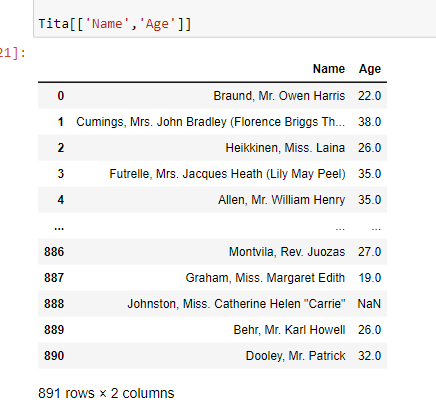




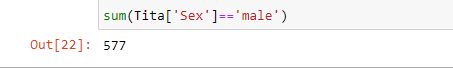


2.Data Filtering





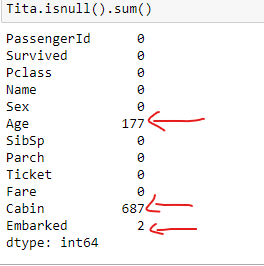
* We have 577 male was traveled on board of Titanic

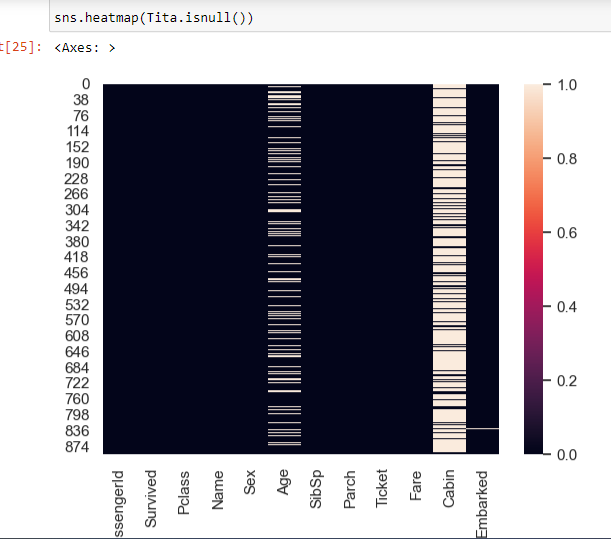


* 342 person were survived on board of Titanic



Check Null values in the Dataset

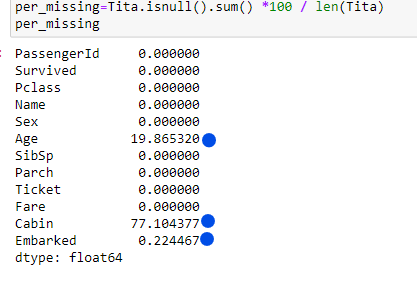




By this we can see that Cabin is having more number of null values than

Age.

3.Handling Missing Data



As seen from above cell, some columns have missing values.

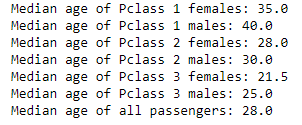
dataset have missing values in Age , Cabin and Embarked columns.

3.1 Age

Missing values in Age are filled with median age, but using median age of the whole data set is not a good choice. Median age of Pclass groups is the best choice because of its high correlation with Age (0.408106) and Survived (0.338481). It is also more logical to group ages by passenger classes instead of other features.

Une image contenant texte, Police, ligne, capture d’écran

Description générée automatiquement



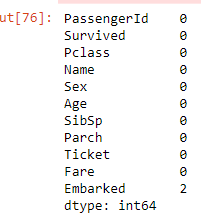
In order to be more accurate, Sex feature is used as the second level of groupby while filling the missing Age values. As seen from above cell, Pclass and Sex groups have distinct median Age values. When passenger class increases, the median age for both males and females also increases. However, females tend to have slightly lower median Age than males. The median ages below are used for filling the missing values in Age feature.

3.2 Cabin

Cabin feature is little bit tricky and it needs further exploration.

Une image contenant texte, Police, capture d’écran, blanc

Description générée automatiquement

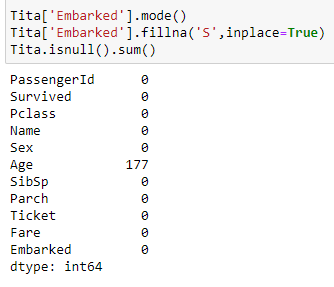


3.2 Embarked

Embarked is a categorical feature and there are only 2 missing values in whole data set. Both of those passengers are female, upper class and they have the same ticket number.

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement



Categorical Data Encoding

Une image contenant texte, ligne, Police, nombre

Description générée automatiquement

As we show that in our dataset there are many categorical columns for ML algorithm. So, we have to change categorical columns into numeric columns. For this I used sklearn LabelEncoder to Encode columns.

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

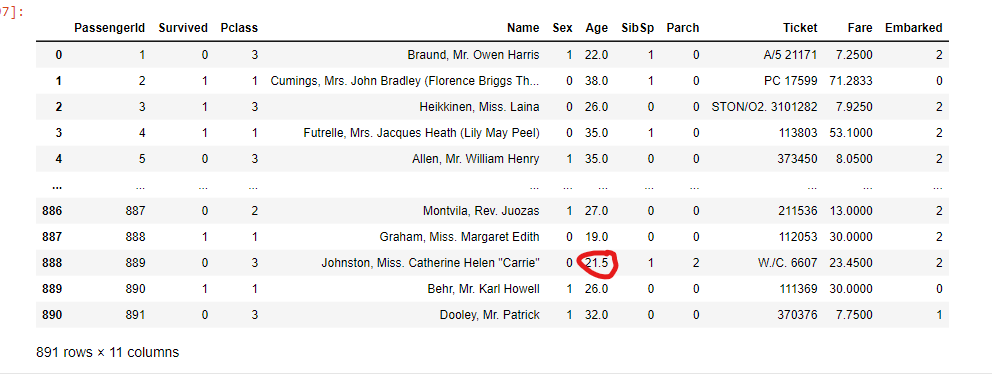
Une image contenant texte, capture d’écran, nombre, logiciel

Description générée automatiquement

So by success I change the type of columns but I realize that there exists a NaN value in line 888.( data cleaning

Une image contenant texte, Police, capture d’écran, ligne

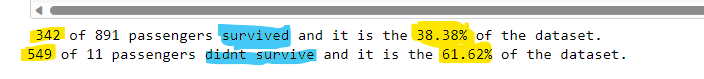
Description générée automatiquement

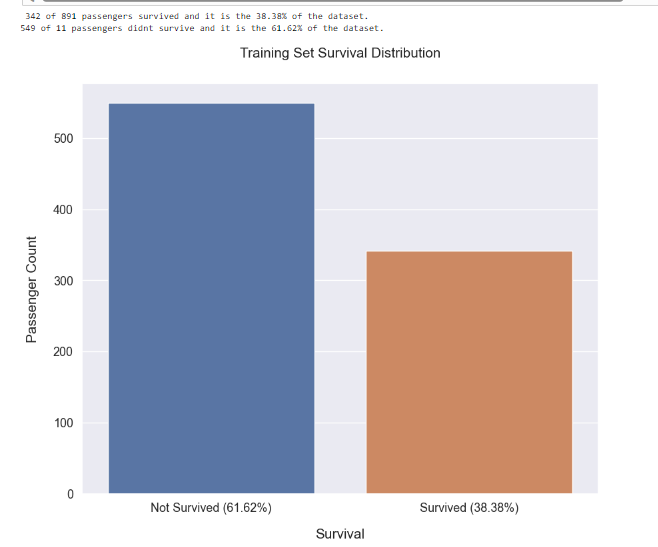


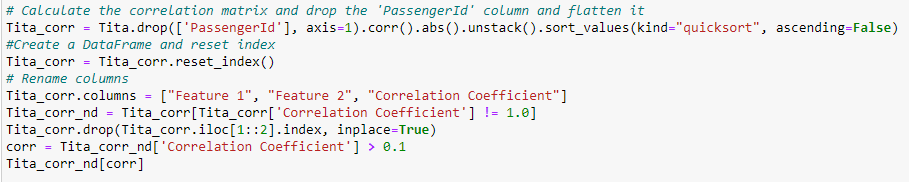
4.Target Distribution

Une image contenant texte, capture d’écran, Police

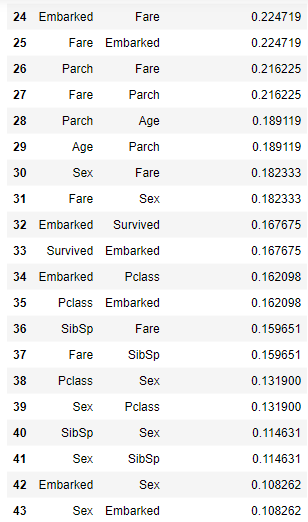
Description générée automatiquement





5.Correlations

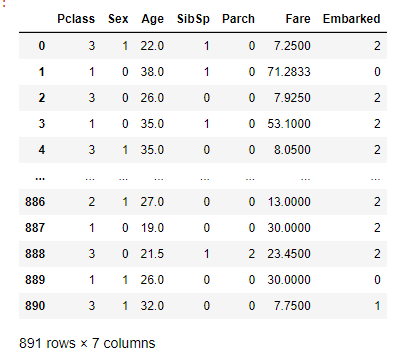
Une image contenant texte, capture d’écran, nombre, menu

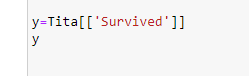
Description générée automatiquement

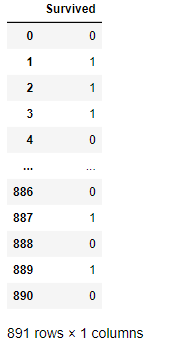
By sorting the correlation values in descending order, as we see the highest correlation between features is 0.549500 in data set between Fare and Pclass and Survived and Sex . The other features are also highly correlated. There are 27 correlations in our data set that are higher than 0.1.

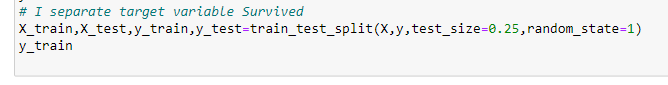
6.Train/Test split

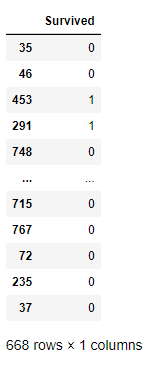


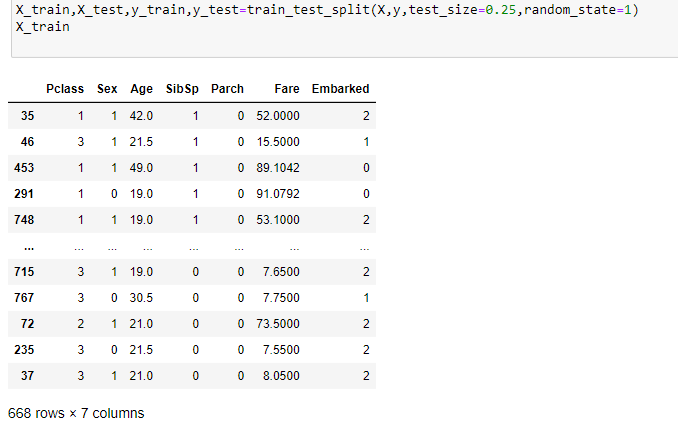








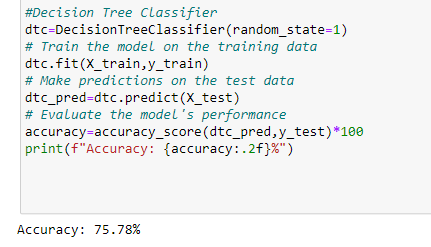




The training set is used to train the models, and the testing set is used to evaluate their performance.

Take test\_size ratio of 75:25 it means we give 75% data to training and 25% for testing random\_state = 1 (it means that it takes observations randomly.)

7.Prediction



* In above cell i fit a model DecisionTreeClassifier on Training setMake prediction on Test set
* Got accuracy is 75.34% which is approximately good.
* the Decision Tree Classifier is a powerful and interpretable algorithm that can be effective for various classification tasks, including predicting the survival of passengers in the Titanic dataset.However, there are always opportunities to improve the model's accuracy by exploring different strategies and techniques to optimize its performance.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

In above cell i fit a model Support Vector Machine(SVM) and Takes kernel- "linear" on Training set

* Make prediction on Test set
* Got accuracy is 79.37% which is better in comparision of Decision Tree Classifier .
* An accuracy of 79.37% means that the SVM model correctly predicted the survival outcome for approximately 79.37% of the

passengers in the dataset.

* Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks.

The linear kernel in SVM means that the algorithm uses a linear decision boundary to separate the data points belonging to different

classes. It works well when the data is linearly separable.

* The reasons for the SVM performing better in this specific case might be:

1. Handling Non-Linearity: The SVM with a linear kernel can still handle some non-linear relationships in the data to some extent, and

in this particular dataset, it might be able to capture the underlying patterns better than the Decision Tree.

2. Less Prone to Overfitting: Decision Trees can easily overfit the training data, especially when not pruned properly. SVM with a linear

kernel is generally less prone to overfitting, which might have contributed to its better generalization performance on the test data.

3. Simpler Model: SVM with a linear kernel is a simpler model compared to Decision Trees. When the dataset has noise or irrelevant

features, a simpler model can often perform better by focusing on the most informative features.

4. Handling Outliers and Irrelevant Features: SVM with a linear kernel is less affected by outliers, and if the dataset contains outliers or

irrelevant features, the SVM might have been able to handle them more effectively than the Decision Tree.

5. Feature Engineering and Selection: The success of SVM with a linear kernel can be influenced by the quality and relevance of the

features used. If the features were well-engineered and selected appropriately, the SVM could have taken advantage of the most

informative ones.

6. Data Size: SVMs often perform well with a relatively small amount of data, and if the dataset size is relatively small, the SVM might

have been able to generalize better.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Got accuracy is 76.68% .

An accuracy of 76.68% means that the GaussianNB model correctly predicted the survival outcome for approximately 76.89% of the

passengers in the dataset.

Gaussian Naive Bayes (GaussianNB):

1.GaussianNB is a probabilistic classifier based on the Bayes theorem, assuming that the features follow a Gaussian distribution.

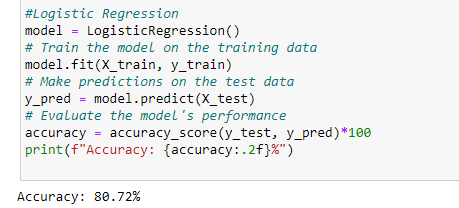
2.It is particularly useful when dealing with continuous or numerical features that can be assumed to be normally distributed.

3.The algorithm calculates the likelihood of each feature value for each class and combines them with the prior probability of each class to make predictions.

GaussianNB is a simple and efficient classification algorithm that is often used for datasets with continuous features that can be

assumed to be normally distributed.

Given that the Titanic dataset contains both categorical and numerical features, using GaussianNB might not be the most appropriate choice.



* Got accuracy is 80.72% .
* An accuracy of 80.72% means that the Logistic Regression model correctly predicted the survival outcome for approximately 80.72%

of the passengers in the dataset.

Logistic Regression:

1. Logistic Regression is a popular and widely used classification algorithm in machine learning.

2. It is used for binary classification tasks where the target variable has two classes (in this case, Survived or Not Survived).

3. Logistic Regression models the relationship between the features and the target variable by using the logistic function (sigmoid

function) to produce probabilities of belonging to each class.

4. The model then predicts the class with the highest probability as the final outcome.

8.Understanding Accuracy Score

Based on the accuracy scores provided, the Logistic Regression classifier gives the best fit model for the Titanic dataset. The accuracy score for the Logistic Regression classifier is 80.72% , indicating that it approximately predicted all instances in the test set.

On the other hand, the Support Vector Machine classifier achieved an accuracy score of 79.37% , and the Decision

Tree Classifier achieved an accuracy score of 75.34% and the GaussianNB achieved an accuracy score 76.68% . While

three classifiers have good accuracy scores, the Logistic Regression classifier achieved a perfect score, indicating that it provided the best fit model for the Titanic dataset in terms of accuracy.

It's important to note that accuracy alone may not be the only metric to consider when evaluating classifier performance. Depending on the specific problem and requirements, other metrics such as precision, recall, or F1 score may provide a more comprehensive evaluation of the classifiers' performance.

Additionally, it's recommended to perform further evaluation and validation techniques, such as cross-validation or assessing the models on unseen data, to ensure the chosen classifier's robustness and generalization ability.

Ultimately, the choice of the best algorithm depends on various factors, including the dataset size, the complexity of the relationships between features and the target, and the specific problem requirements.

It's essential to experiment with multiple algorithme and perform proper evaluation and validation to select the most appropriate one for your specific problem.

9.Why Logistic Regression is best fit

The reason why Logistic Regression is giving the best accuracy in comparison to Decision Tree Classifier, SVM, and GaussianNB for the Titanic dataset could be attributed to various factors, which we'll explore below:

1. Suitable for Linear Separability: The Titanic dataset might have a linearly separable relationship between the features and the target variable (Survived or Not Survived). Logistic Regression is well-suited for such cases as it models the relationship between the features and the target using a linear function (log-odds) with the sigmoid activation function.

2. Robust to Noise: If the dataset contains noise or irrelevant features, Logistic Regression can still perform relatively well. It is less sensitive to noisy data compared to Decision Trees, which are prone to overfitting with noisy data.

3. No Kernel Selection: Logistic Regression does not involve the selection of a kernel (like SVM). Choosing the right kernel and tuning its parameters can be challenging, especially when the dataset is not clearly separable.

4. Fewer Hyperparameters to Tune: Logistic Regression has fewer hyperparameters to tune compared to SVM or Decision Trees. This simplicity might make it easier to find a good set of hyperparameters, leading to better accuracy.

5. Handling Class Imbalance: If the dataset has an imbalanced class distribution (survived vs. not survived), Logistic Regression can handle it relatively well. It naturally balances the class distribution using the logistic function.

6. Data Size and Linearity: Logistic Regression often performs well on small to medium-sized datasets and when the relationship between the features and target is approximately linear. For some datasets, these conditions might be met, leading to better accuracy.

7. Feature Scaling: Logistic Regression is less sensitive to feature scaling compared to some other algorithms like SVM, which can be affected by the scale of the feature.

9.Conclusion

The Titanic Survival Prediction project is a machine learning project that uses historical data about the Titanic disaster to predict whether a passenger would survive. The data includes information about the passenger's sex, age, class, fare, and whether they were traveling with family. The project uses a variety of machine learning algorithms to train a model that can predict survival. The most common algorithms used for this project are logistic regression, decision trees, and Naïve Bayes.

The accuracy of the model depends on the quality of the data and the complexity of the model. In general, more complex models can achieve higher accuracy, but they are also more likely to overfit the data.