Name: Venkata Sai Kusuma Sindhoora, Vankayala Siva

SJSU ID: 014540916

Titanic: Machine Learning from Disaster

Data Exploration and Visualization:

1. Load the train and test data into data frames using pandas.
2. Test set does not have survived column and rest all columns are same as training data.
3. There are 891 records in training data and 418 records are found in testing data.
4. Missing values are found in features Age are 177, Cabin is 687 and embarked is 2.
5. Among 891 passengers of training data, 38% are survived and rest 61% are dead.
6. Passengers from first class has more survival rate i.e. 62% compared to other class passengers.
7. The columns such as 'Survived','Pclass','Age','SibSp','Parch', 'Fare' are continuous features and 'Sex', 'Ticket', 'Cabin', 'Embarked' are categorical features.
8. The histogram of continuous feature Age tends to have normal distribution with bimodal peak whereas column Fare has exponential distribution.
9. The correlation matrix shows features Pclass and Fare has strong correlation with target variable Survived.
10. Though the population of male passengers are more than female passengers, the survival rate from bar plot shows female passengers have good survival rate when compare to male passengers.
11. The frequency plots of tickets and cabins are clumsy and not clear whereas Embarked shows highest for ‘S’ place.

Feature Engineering:

1. Merged both train and test data to perform feature engineering.
2. Added both SibSp and Parch to calculate the family size of passengers. Labelled the single passengers as ‘0’, small family with range 2 to 4 as ‘1’ and rest all passengers i.e. ‘Large Family’ as ‘2’.
3. A calculated column called ‘Is Alone’ is added to the data frame where family size== ‘0’.
4. Extracted all the titles from Name column and mapped it to Mr, Miss, Mrs and Master.
5. Found 284 ages in first class, 261 ages in second class and 501 ages in third class.
6. Impute the missing value of Age by taking the group median of pclass, sex and title and updated all the missing records of data.
7. Mapped cabin to ‘0’ if it is float or else ‘1’.
8. Impute the missing values of ‘Embarked’ with most frequent value in training set i.e ‘S’.

9. Performed one hot encoding to create the dummy columns and dropped the actual column.
10. Impute the missing values with mean of train data for Fare and converted it to categorical feature using pandas qcut with bins=4.
11. A calculated column called numeric ticket is populated with ‘1’ if it is numeric else ‘0’.
12. The categorical column Sex is converted to numerical feature using Label encoding.
13. Performed one hot encoding on Title feature and added it to data frame.
14. Drop all unnecessary columns 'PassengerId', 'Name', 'Title', 'Fare', 'Ticket', 'SibSp', 'Parch', 'Age', 'Family_size'.
15. Performed min-max normalization on pclass, Fare_cat columns.
16. Slice the train and test data based on index 891 and drop the source column.
17. Concatenated the Survived column to training data for modelling.

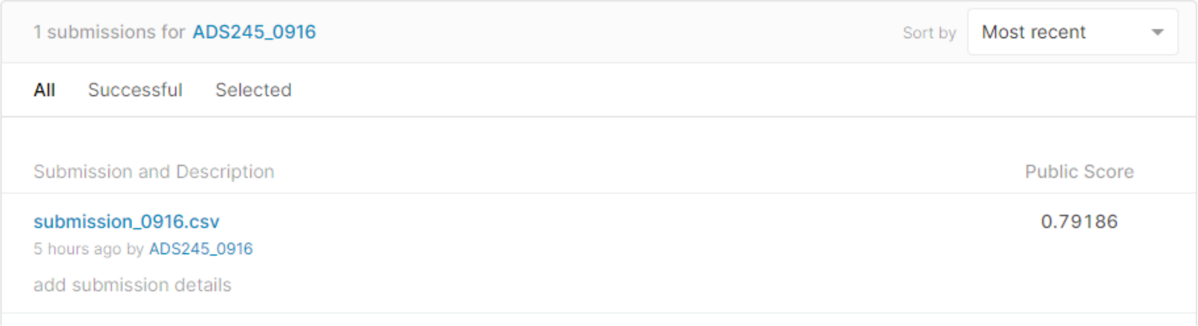
Data Modelling and Evaluation:

Implemented KNN algorithm (Supervised Learning):

Pseudo code:

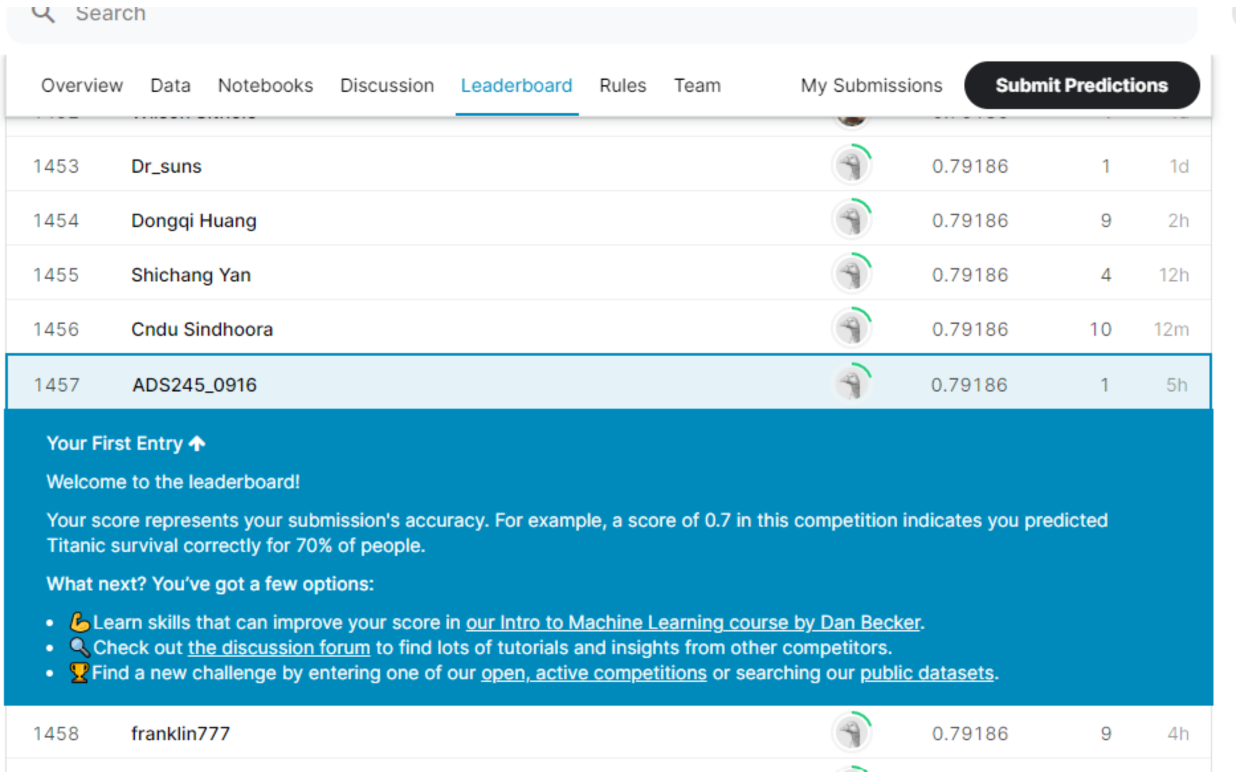
1. Load the csv file of training data. Append each record of training data into a list using the csv reader.
2. Convert all the string column of data to float.
3. Initialize the parameter ‘num_neighbors’ to select the nearest neighbors.
4. Call the evaluate algorithm with parameters as train data, model, num_neighbors and k_folds are set to 5 by default.
5. Perform cross validation split to generate (n-1) training set and one remaining test set using folds size and data split.
6. For each fold of test set, set the target column as None before calling the knn model.
7. The k_nearest neighbors function takes input as train data, test data and num_neighbors as arguments.
8. For each record in test, it will compute the Euclidean distance with entire training set. Sort the distances using numpy argsort function and indices from smallest to largest (ascending order) by distances. Pick the first num_neighbors entries from sorted entries.
9. Pick the labels of selected neighbors and return the mode of predictions.
10. The accuracy percentage is calculated for actual and prediction results.
11. Once the model is trained, pass the training set to k_nearest neighbors function to predict the survived classification and add that result to final dataframe.
12. Converted final dataframe to submission.csv keeping columns of Passenger Id and Survived for Kaggle submission.
13. The result classifies 281 passengers as ‘0’ class and 137 passengers as ‘1’ class.

Kaggle Submission:



1 submissions for ADS245_0916

Submission and Description	Public Score
submission_0916.csv 5 hours ago by ADS245_0916 add submission details	0.79186



1453	Dr_suns	0.79186	1	1d
1454	Dongqi Huang	0.79186	9	2h
1455	Shichang Yan	0.79186	4	12h
1456	Cndu Sindhoora	0.79186	10	12m
1457	ADS245_0916	0.79186	1	5h

Your First Entry ↑

Welcome to the leaderboard!

Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.

What next? You've got a few options:

- 💪Learn skills that can improve your score in our Intro to Machine Learning course by Dan Becker.
- 🔍Check out the discussion forum to find lots of tutorials and insights from other competitors.
- 🏆Find a new challenge by entering one of our open, active competitions or searching our public datasets.

1458	franklin777	0.79186	9	4h