Titanic

MACHINE LEARNING FROM DISASTER

Gene Olafsen

Tragedy

- The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.
- One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Challenge

- Kaggle asks you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Goal

- It is your job to predict if a passenger survived the sinking of the Titanic or not.
- For each in the test set, you must predict a 0 or 1 value for the variable.

Benchmark

- Your score is the percentage of passengers you correctly predict. This is known simply as "accuracy”.

Accuracy

- Accuracy is also used as a statistical measure of how well a "binary classification" test correctly identifies or excludes a condition.
- In the Titanic sense... the binary classification is *survived* or *died*.
- The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.
- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- where: TP = True positive; FP = False positive; TN = True negative; FN = False negative

Features

- A feature is a measurable property of the object you're trying to analyze.
- When you look at your dataset, features appear as columns.

Training Data

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Test Data

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

- pclass: A proxy for socio-economic status (SES)
  1st = Upper
  2nd = Middle
  3rd = Lower

  age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

  sibsp: The dataset defines family relations in this way...
  Sibling = brother, sister, stepbrother, stepsister
  Spouse = husband, wife (mistresses and fiancés were ignored)

  parch: The dataset defines family relations in this way...
  Parent = mother, father
  Child = daughter, son, stepdaughter, stepson
  Some children travelled only with a nanny, therefore parch=0 for them.

Test.csv Sample Record

- PassengerId,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
- 892,3,"Kelly, Mr. James",male,34.5,0,0,330911,7.8292,,Q

Prediction Record

- Kaggle also includes gender_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

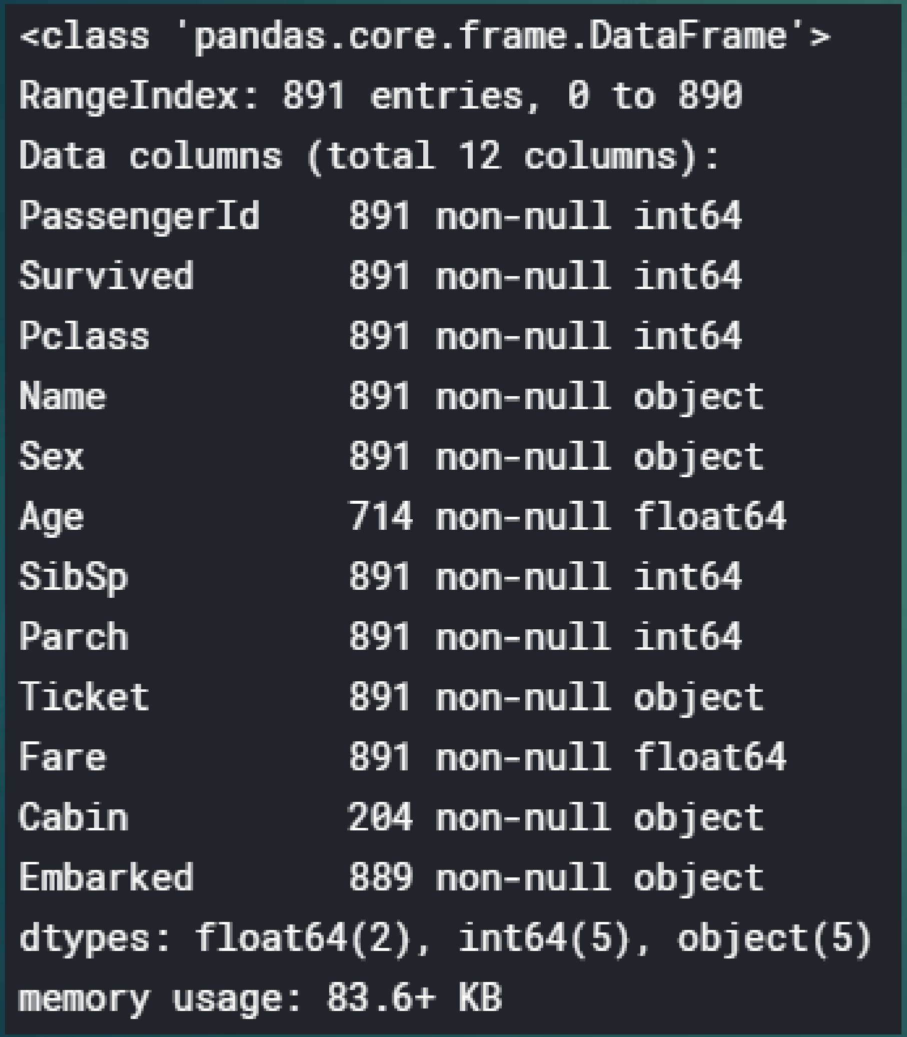
- PassengerId,Survived
- 892,0

Feature Engineering

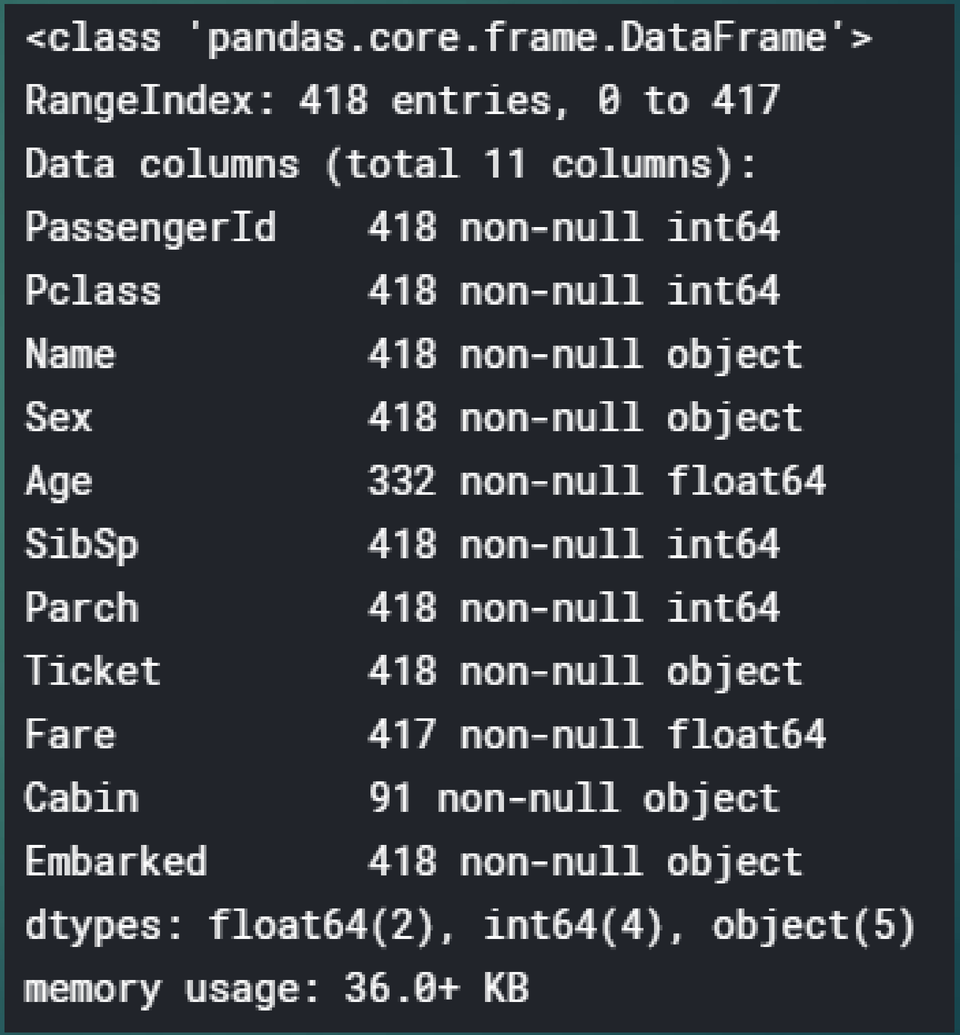
- Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.
- Feature engineering is fundamental to the application of machine learning, and can be both difficult and expensive.

Loading Data

train_df.info()
test_df.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
PassengerId    418 non-null int64
Pclass         418 non-null int64
Name           418 non-null object
Sex            418 non-null object
Age            332 non-null float64
SibSp          418 non-null int64
Parch          418 non-null int64
Ticket         418 non-null object
Fare           417 non-null float64
Cabin          91 non-null object
Embarked       418 non-null object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

Columns having missing values are

1.Train data - Age,Embarked, Cabin

2.Test data - Age, Cabin

First Look

- Look for the obvious. (relationships)
- Look for missing or incomplete data.
- Start to consider what data to keep and what to drop.

What is Important

- Observation
  - Out of 891 survived, 577 are male(~65%).
  - Most of the survived(644) embarked from port S(~72%).
- Analysis
  - Age and Embarked play a important role in suvival rate.
- Action
  - We need to fill the missing values.

What is not Important

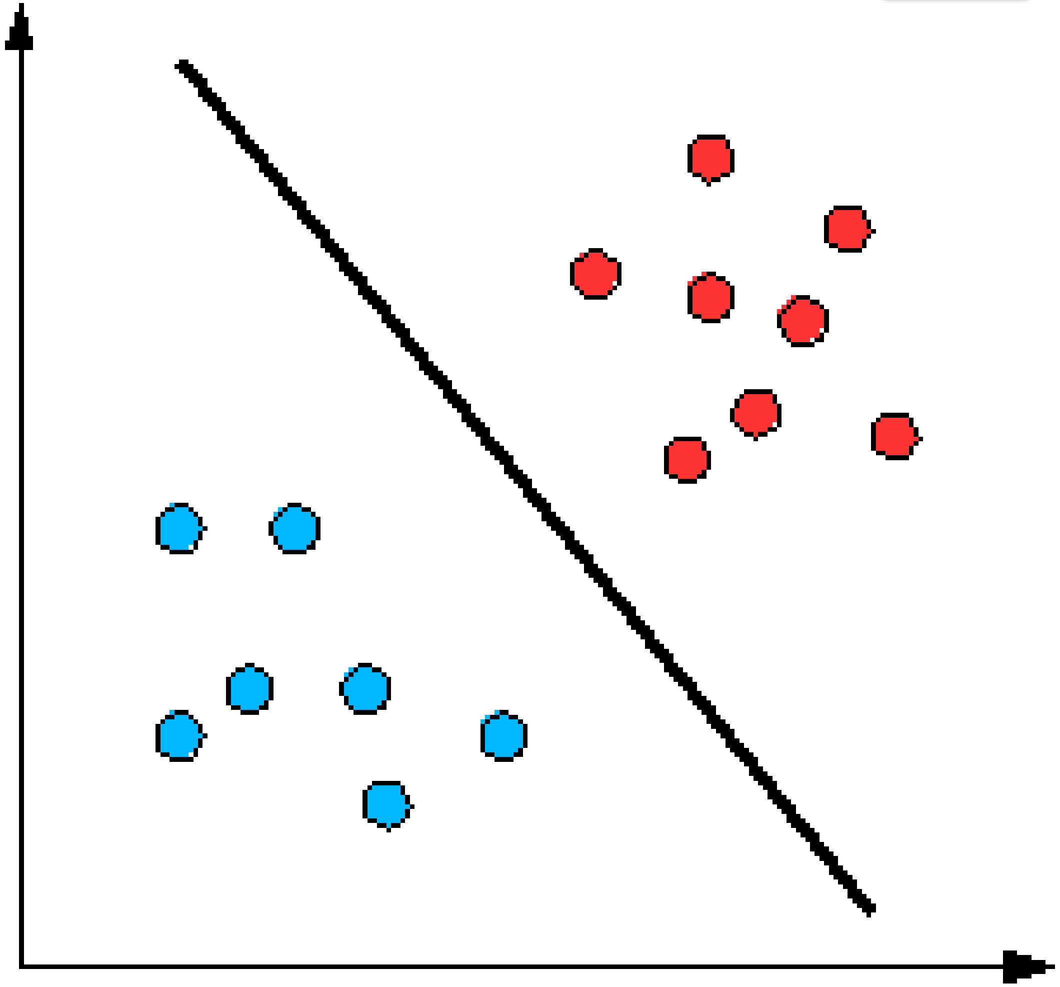
- We also need to drop certain features...
- Observation
  - Cabin is highly incomplete- drop this feature
  - Passenger ID, Name and Ticket may also be dropped as they are irrelevant in determining the survival.
    - Could you determine the 'class'/'status' of the person from their name?

Kernel, Kernel Trick, Kernel Function

- Has nothing to do with KFC.
- In machine learning, a "kernel" is usually used to refer to the *kernel trick*, a method of using a linear classifier to solve a non-linear problem.
- The kernel trick entails transforming linearly inseparable data- to linearly separable ones.
- The kernel function is what is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable.

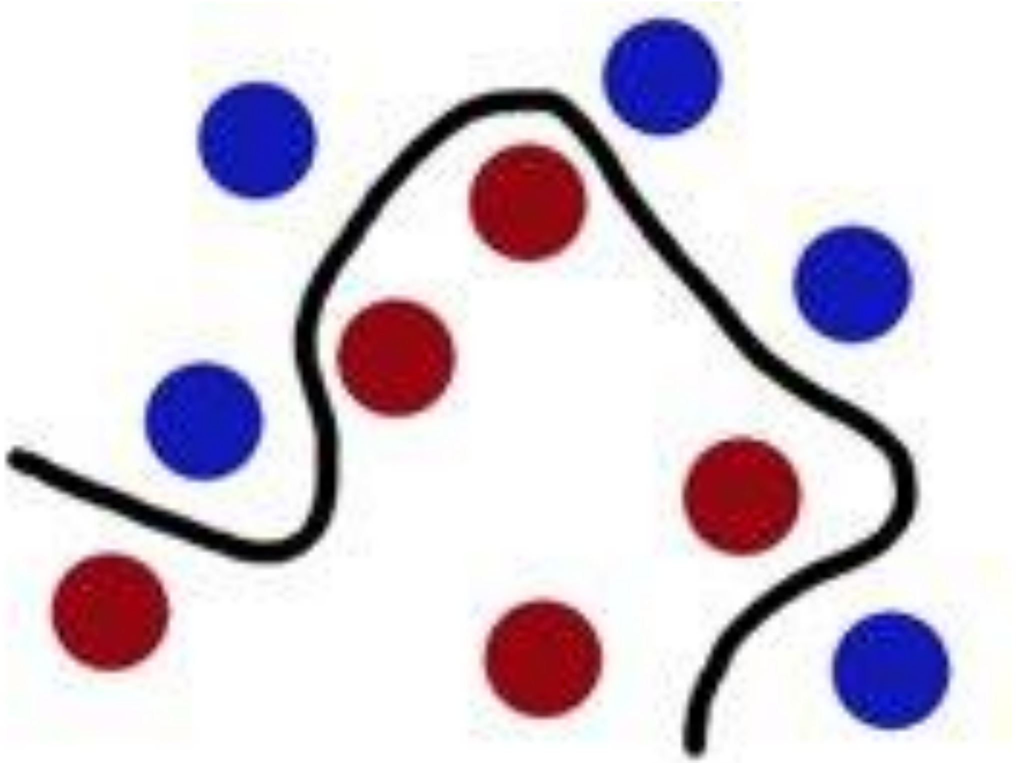
Hyperplane

- A Hyperplane in two dimensional space.

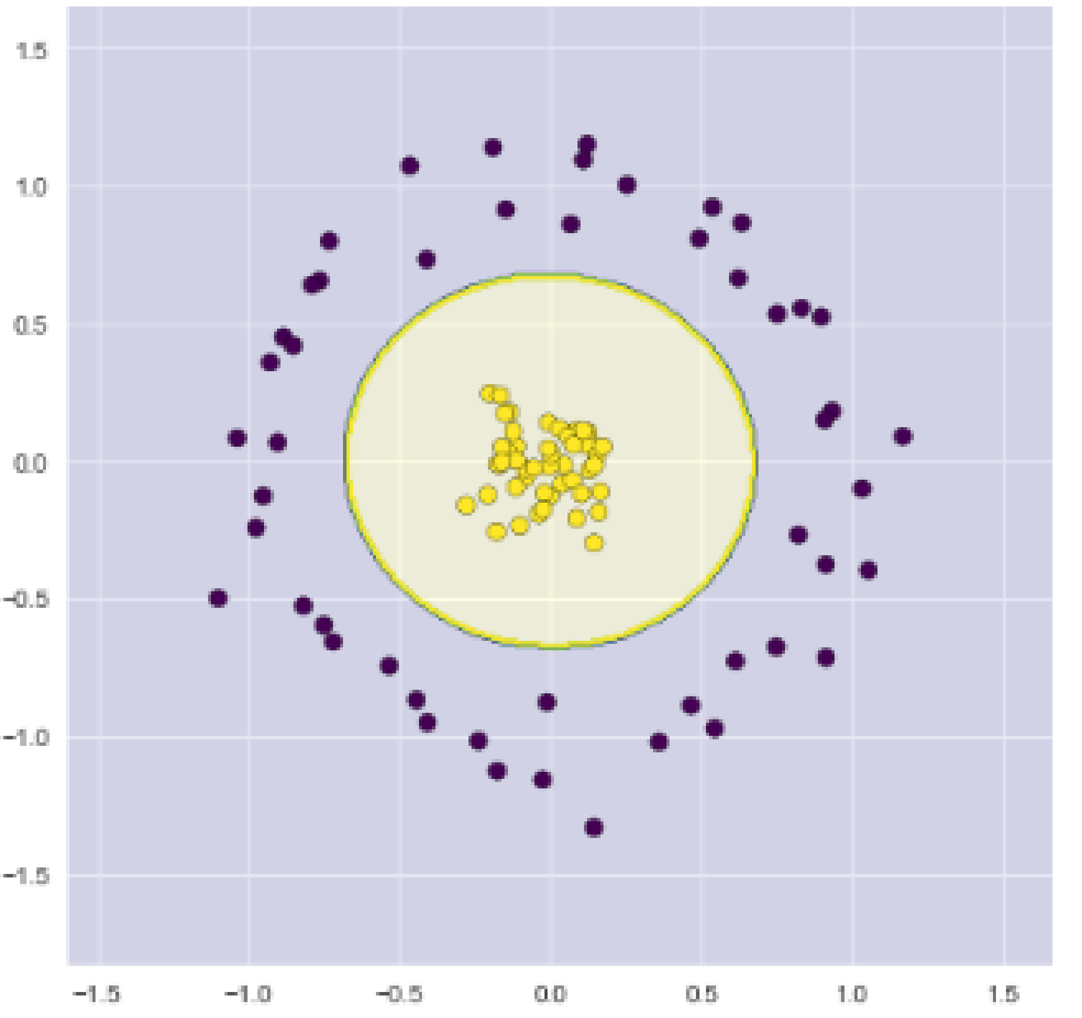


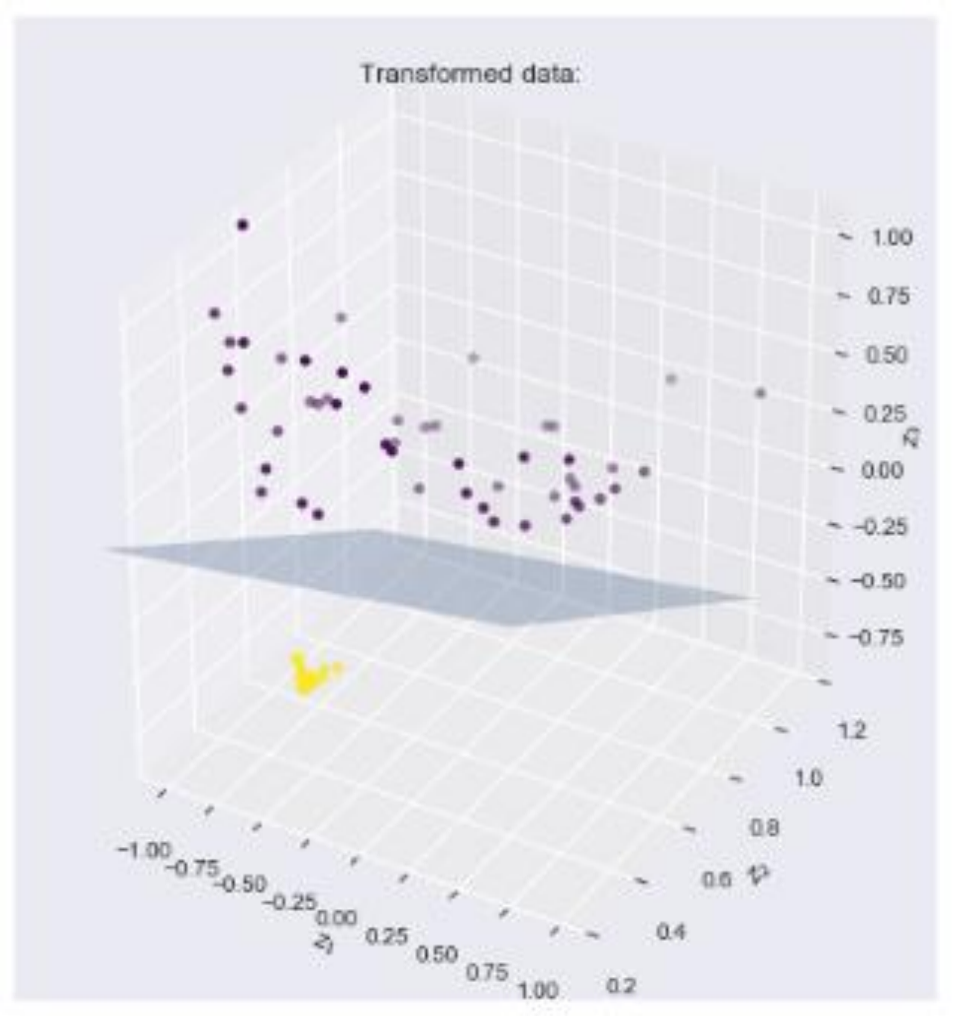
Real-world Distribution

- The red and blue dots cannot be separated by a straight line as they are "randomly distributed" and this, in reality, is how most real life problem data are -randomly distributed



Visualize Hyperplane





Important Links

- https://www.kaggle.com/c/titanic/overview
- https://www.kaggle.com/sriloksagar/titanic-survival-prediction-98-accuracy