



Machine Learning

Datasets and Project Ideas

REVA Academy for
Corporate Excellence - RACE

Machine Learning Datasets for Novices in Data Science

It can be challenging to find the appropriate dataset when doing research for machine learning or data science projects. And you need a large data to create reliable models. Don't worry though; a lot of researchers, groups, and people have made their work available, and we may use their datasets in our projects. We will examine three most popular machine learning datasets in this article that you can use to develop your upcoming data science project.

Titanic Dataset

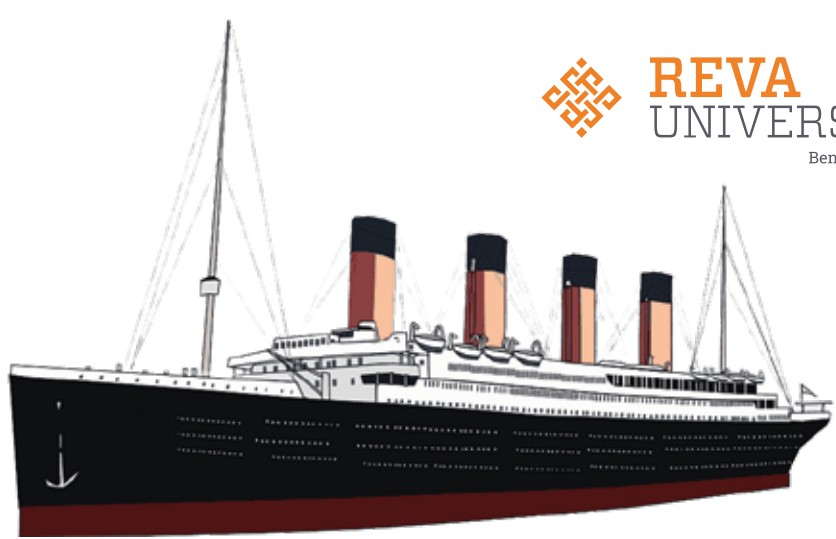
Introduction:

The largest passenger ship ever built was involved in an iceberg collision on April 15, 1912. Out of 2224 passengers and crew, 1502 died when the Titanic sank. The international society was stunned by this shocking catastrophe, which prompted improved ship safety rules. There weren't enough lifeboats for the passengers and crew, which was one of the factors contributing to the shipwreck's high death toll. Some groups of people had a higher chance of surviving the sinking than others, even though there was some element of luck involved.

Data of the actual Titanic passengers can be found in the titanic.csv file. One person is represented by each row. The columns give information about the person's age, passenger class, sex, whether they survived, whether they paid a fare, and other characteristics.

Data Link:

[Titanic Survival Data Set | Kaggle](#)



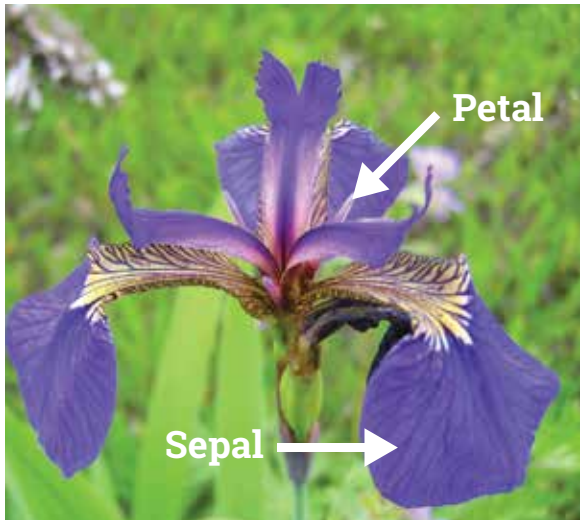
Data Description:

- **Survived:** Outcome of survival (0 = No; 1 = Yes)
- **Pclass:** Socio-economic class (1 = Upper class; 2 = Middle class; 3 = Lower class)
- **Name:** Name of passenger
- **Sex:** Sex of the passenger (Male/Female)
- **Age:** Age of the passenger (Some entries contain NaN)
- **SibSp:** Number of siblings and spouses of the passenger aboard
- **Parch:** Number of parents and children of the passenger aboard
- **Ticket:** Ticket number of the passenger
- **Fare:** Fare paid by the passenger
- **Cabin:** Cabin number of the passenger (Some entries contain NaN)
- **Embarked:** Port of embarkation of the passenger (C = Cherbourg; Q = Queenstown; S = Southampton)

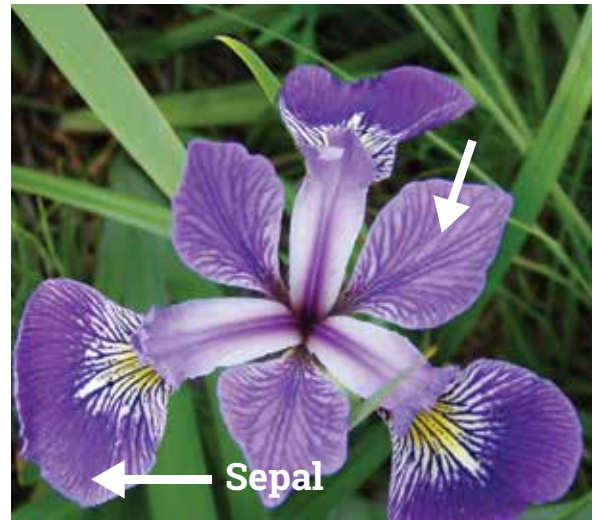
Data Science Project Idea:

You can create a model to determine whether a passenger would have survived the Titanic. For this, you can use linear regression.

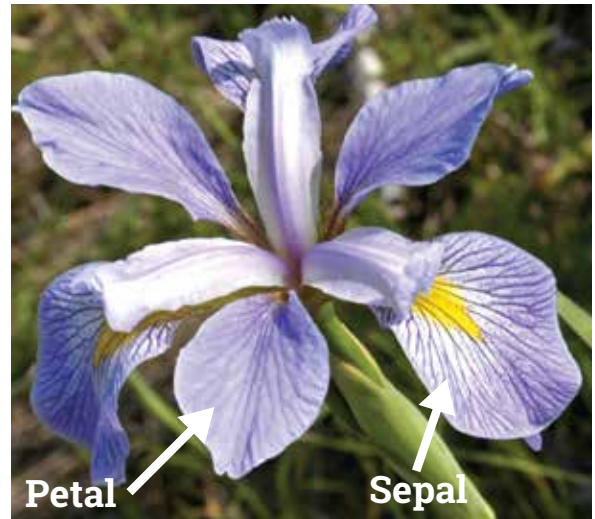
https://public.tableau.com/views/Titanic_265/Titanic?:showVizHome=no



Iris Setosa



Iris Versicolor



Iris Virginica

Iris Dataset

Introduction:

The Iris flower data set is a multi-variate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper for the use of multiple measurements in taxonomic problems. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 150 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Data Description:

- **PetalLengthCm:** Petal Length in cm
- **PetalWidthCm:** Petal Width in cm
- **SepalLengthCm:** Sepal Length in cm
- **SepalWidthCm:** Sepal width in cm
- **Class(Species):** Iris-setosa, Iris virginica, and Iris-versicolor.

Data Science Project Idea:

You can build a machine learning classification model on the dataset.

Data Link:

<https://www.kaggle.com/code/jchen2186/machine-learning-with-iris-dataset/data>



The Boston Housing Dataset

Introduction:

This is a popular dataset used in pattern recognition. It contains information about the different houses in Boston based on crime rate, tax, number of rooms, etc. It has 506 rows and 14 different variables in columns. You can use this dataset to predict house prices.

The data was originally published by

Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Data Link:

<https://www.kaggle.com/datasets/altavish/boston-housing-dataset?select=HousingData.csv>

Data Description:

- **CRIM** - per capita crime rate by town
- **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS** - proportion of non-retail business acres per town.
- **CHAS** - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- **NOX** - nitric oxides concentration (parts per 10 million)
- **RM** - average number of rooms per dwelling
- **AGE** - proportion of owner-occupied units built prior to 1940
- **DIS** - weighted distances to five Boston employment centres
- **RAD** - index of accessibility to radial highways
- **TAX** - full-value property-tax rate per \$10,000
- **PTRATIO** - pupil-teacher ratio by town
- **B** - $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
- **LSTAT** - % lower status of the population
- **MEDV** - Median value of owner-occupied homes in \$1000's

4

Data Science Project Idea:

Predict the housing prices of a new house using linear regression. Linear regression is used to predict values of unknown input when the data has some linear relationship between input and output variables.