# Classification dataset

General information about dataset

|  |  |
| --- | --- |
| Name | titanic |
| No. of classes | 11 |
| Total no. of samples | 891 |
| No. of samples in training | 712 |
| No. of samples in testing | 179 |

## About Dataset

### Context

This Data was originally taken from [Titanic: Machine Learning from Disaster](https://www.kaggle.com/c/titanic/data) .But its better refined and cleaned & some features have been self engineered typically for logistic regression . If you use this data for other models and benefit from it , I would be happy to receive your comments and improvements.

### Content

There are two files namely:-  
**train\_data.csv** :- Typically a data set of 792x16 . The **survived** column is your target variable (The output you want to predict).The **parch & sibsb** columns from the original data set has been replaced with a single column called **Family size**.

All Categorical data like **Embarked , pclass** have been re-encoded using the one hot encoding method .

Additionally, 4 more columns have been added , re-engineered from the **Name column** to **Title\_1 to Title\_4** signifying males & females depending on whether they were married or not .(Mr , Mrs ,Master,Miss). An additional analysis to see if Married or in other words people with social responsibilities had more survival instincts/or not & is the trend similar for both genders.

All missing values have been filled with a median of the column values . All real valued data columns have been normalized.

**test\_data.csv** :- A data of 100x16 , for testing your model , The arrangement of **test\_data** exactly matches the **train\_data**

I am open to feedbacks & suggesstions

## Preprocessing phase:

1. Removing Irrelevant Columns: PassengerId, Name, Ticket, and Cabin columns are dropped.
2. Handling Missing Values: Age and Embarked columns' missing values are filled.
3. One-Hot Encoding: Categorical columns are one-hot encoded.
4. Standardization: Numeric columns are standardized.
5. Splitting Data: Data is split into training and testing sets (80:20 ratio).

**Ann model:**

Model Definition:

Sequential model with three layers:

Dense layer (64 neurons, ReLU activation)

Dropout layer (dropout rate 0.5)

Dense layer (32 neurons, ReLU activation)

Dropout layer (dropout rate 0.5)

Dense layer (1 neuron, sigmoid activation)

**Hyperparameters Used:**

Initial Learning Rate: 0.001

Optimizer: Adam optimizer

Regularization: Dropout layers (with a dropout rate of 0.5)

Batch Size: 32

Number of Epochs: 300

Loss Function: Binary Crossentropy Metrics: Accuracy

**Results details:**

**train accuracy: 0.85.8%**

**test accuracy: 0.82%**

A graph with blue lines

Description automatically generated A graph with blue lines

Description automatically generated

A blue squares with white text

Description automatically generated A graph of a function

Description automatically generated with medium confidence

# DECISION TREE MODEL:

model = DecisionTreeClassifier(max\_depth=4)

model.fit(x\_train,y\_train)

**Hyperparameters Used:**

max\_depth =4

**Result Details:**

train accuracy:0.8426966292134831

test accuracy 0.8044692737430168

A blue squares with white text

Description automatically generated

A graph of a function

Description automatically generated with medium confidence

**Regression dataset**

General information about dataset

|  |  |
| --- | --- |
| Name | **Salary** |
| No. of classes | 8 |
| Total no. of samples | 6684 |
| No. of samples in training | 4678 |
| No. of samples in testing | 2006 |

**About Dataset :-**

The primary dataset used is "Salary Data", which encompasses several features:

Age: The age of the employees.

Years of Experience: The total working experience of the employees**.**

Gender: The gender of the employees.

Job Title: The designation or role of the employees.

Education Level: The highest educational qualification of the employees.

Salary: The annual salary of the employees (our target variable.

**Data Exploration and Visualization:**

Loaded the dataset into a pandas DataFrame for easy manipulation and analysis. Conducted a comprehensive overview of the dataset, understanding its structure and content. Used statistical measures to understand the distribution and central tendencies of the data. Visualized the data using libraries like matplotlib and seaborn to understand the relationships between different features and the target variable.

**Data Preprocessing:**

Handled missing values and outliers.

Utilized ColumnTransformer to apply different preprocessing to different columns: encoding categorical variables and scaling numerical ones, all in a single step, maintaining the efficiency of the workflow.

Split the dataset into training and testing sets to train and subsequently evaluate our machine learning models.

**SVM MODEL :-**

A screen shot of a computer

Description automatically generated

**Hyperparameters Used:**

* C = 2000
* Gamma = ‘auto’

**Kernel type:**

RBF

Calculate mean squared error

**Calculate mean squared error & R-squared :**

A computer screen shot of a program

Description automatically generated

**Result Details :**

  Mean Squared Error : 656630700.8715392

  Mean absolute Error : 18740.1469156983

  R-squared train : 0.7650738393715122

  R-squared test : 0.7518264485203661

A diagram of a scatter plot

Description automatically generated