# MACHINE LEARNING LABORATORY

## 1818128_JAYASREE T

## ☆ EXP 3 : For the Titanic dataset from kaggle guess whether the individuals from the dataset had survived or not and also calulate the gini index

## Importing the Dataset

```
[1] import pandas as pd

    df = pd.read_csv('titanic.csv', index_col='PassengerId')
```

```
[2] print(df.head())
```

```
             Survived  Pclass  ... Cabin Embarked
PassengerId                    ...
1                   0       3  ...   NaN        S
2                   1       1  ...   C85        C
3                   1       3  ...   NaN        S
4                   1       1  ...  C123        S
5                   0       3  ...   NaN        S
```

1818128_JAYASREE T_EXP 3

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

[2]
```
PassengerId                    ...
1                  0       3  ...   NaN        S
2                  1       1  ...   C85        C
3                  1       3  ...   NaN        S
4                  1       1  ...  C123        S
5                  0       3  ...   NaN        S

[5 rows x 11 columns]
```

[3]
```
df = df[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Survived']]
```

[4]
```
df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})
```

[5]
```
df = df.dropna()
```

[7]
```
X = df.drop('Survived', axis=1)
y = df['Survived']
```

## Splitting Dataset into training and testing sets

[11]
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

1818128_JAYASREE T_EXP 3

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

## Building Model

```
[8] from sklearn import tree

    model = tree.DecisionTreeClassifier()
```

```
[9] model
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

```
[12] model.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

## Accuracy of Model

1818128_JAYASREE T_EXP 3

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

# Accuracy of Model

```
[22] y_predict = model.predict(X_test)

     from sklearn.metrics import accuracy_score

     print("Accuracy: ",accuracy_score(y_test, y_predict))
```

```
Accuracy:  0.8324022346368715
```

```
[23] from sklearn.metrics import confusion_matrix

     pd.DataFrame(
         confusion_matrix(y_test, y_predict),
         columns=['Predicted Not Survival', 'Predicted Survival'],
         index=['True Not Survival', 'True Survival']
     )
```

| | Predicted Not Survival | Predicted Survival |
|---|---|---|
| **True Not Survival** | 98 | 14 |
| **True Survival** | 16 | 51 |

```
[15] from subprocess import call
```

1818128_JAYASREE T_EXP 3

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
[15] from subprocess import call

     call(['dot', '-T', 'png', 'tree.dot', '-o', 'tree.png'])
```

```
2
```

## Gini Index

```
[16] def get_gini_impurity(survived_count, total_count):
         survival_prob = survived_count/total_count
         not_survival_prob = (1 - survival_prob)
         random_observation_survived_prob = survival_prob
         random_observation_not_survived_prob = (1 - random_observation_survived_prob)
         mislabelling_survided_prob = not_survival_prob * random_observation_survived_prob
         mislabelling_not_survided_prob = survival_prob * random_observation_not_survived_prob
         gini_impurity = mislabelling_survided_prob + mislabelling_not_survided_prob
         return gini_impurity
```

```
[17] gini_impurity_starting_node = get_gini_impurity(342, 891)
     gini_impurity_starting_node
```

```
0.47301295786144265
```

```
[18] gini_impurity_men = get_gini_impurity(109, 577)
     gini_impurity_men
```

```
[18] gini_impurity_men = get_gini_impurity(109, 577)
     gini_impurity_men
```

```
0.3064437162277843
```

```
[19] gini_impurity_women = get_gini_impurity(233, 314)
     gini_impurity_women
```

```
0.3828350034484158
```

## Visualization

```
[21] from IPython.display import Image as Image
     from sklearn.externals.six import StringIO
     from sklearn.tree import export_graphviz
     import pydotplus
     dot_data=StringIO()
     export_graphviz(model,out_file=dot_data,filled=True,rounded=True,special_characters=True,
                     class_names=['0','1'])
     graph=pydotplus.graph_from_dot_data(dot_data.getvalue())
     graph.write_png('tree.png')
     Image(graph.create_png())
```

1818128_JAYASREE T_EXP 3 - Co

colab.research.google.com/drive/1YshaOuCXzRavUQIwg1njXNT-FriNMofl#scrollTo=7VFGnZOZ5ur8

1818128_JAYASREE T_EXP 3

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
[21] graph=pydotplus.graph_from_dot_data(dot_data.getvalue())
     graph.write_png('tree.png')
     Image(graph.create_png())
```