AISSMS
INSTITUTE OF INFORMATION TECHNOLOGY
(IOIT)
ADDING VALUE TO ENGINEERING
An Autonomous Institute Affiliated to Savitribai Phule Pune University
Approved by AICTE, New Delhi and Recognised by Govt. of Maharashtra
Accredited by NAAC with "A+" Grade | NBA - 5 UG Programmes

2022-2023

सत्याला मरण नाही

Department of Computer Engineering
MINI PROJECT REPORT ON

"Implement merge sort and multithreaded merge sort. Compare time required by both the algorithms. Also analyse the performance of each algorithm for the best case and the worst case."

Submitted By

76	Atharva Mohan Tirkhunde	72145918J

Guided by
Ms. Shilpa Pimpalkar

Project Aim

ML model to predict who survived the Titanic shipwreck using Random Forest Classifier.

Problem Statement

Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data (i.e. name, age, gender, socio-economic class, etc.). Dataset Link: https://www.kaggle.com/competitions/titanic/data

Project Objective

- To build a model for classification.
- To analyze its performance on Titanic Dataset.
- To use different ML and Feature Selection concepts
- To optimize the model's performance.

Project Scope

- While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.
- Predicting the survival rate of the people in a disastrous accident.

System Requirements

Operating System: 64 bit Linux
or its derivatives / Windows.
Python Programming
Language >= 3.6
Jupyter Notebook >= 4.1.11
Pip >= 3.0.2
Numpy >= 1.18.2
Pandas >= 1.0.3
Matplotlib >= 1.0.1
Seaborn >= 1.8.5
Scikit Learn >= 1.1.3

Theory

Binary Classification

Binary classification is a supervised learning algorithm that categorizes new observations into one of two classes

Standard Scaling

Standardization is a scaling technique where the value are manipulated such that it becomes centered around the mean with a unit standard deviation.

01

02

03 Confusion Matrix

Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems.

Confusion matrices represent counts from predicted and actual values. The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.

Theory

The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

1.TP (True Positive): TP represents the number of patients who have been properly classified to have malignant nodes, meaning they have the disease.

2.TN (True Negative): TN represents the number of correctly classified patients who are healthy.

3.FP (False Positive): FP represents the number of misclassified patients with the disease but actually they are healthy. FP is also known as a Type I error.

4.FN (False Negative): FN represents the number of patients misclassified as healthy but actually they are suffering from the disease. FN is also known as a Type II error.

Theory

04 RandomForestClassifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

In random forests (see RandomForestClassifier and RandomForestRegressor classes), each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set.

Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features. (See the parameter tuning guidelines for more details).

Theory

Theory

Modules

Numpy For Line Algebra and Maths in the Notebook.

Pandas For Data Processing in the Notebook.

Matplotlib For Plotting charts and graphs for better visualization.

Seaborn For lightweight, powerful visualization of data.

Scikit Learn For model evaluation, preprocessing, Data Splitting etc.

Project Outcome

・ The model was implemented using the Random Forest Classifier.

・ The Model was used to predict the survival of an individual using various parameters with a state of the art accuracy.

Algorithm

1

Import

Import necessary libraries and datasets.

2

Statistics

Take a statistical look at the dataset.

3

Missing Values

Handle missing values in the dataset.

4

EDA

Plot various graphs for gaining insights from the Exploratory Data Analysis.

5

ML Model

Split dataset into training and validation set and build Machine Learning model

6

Predictions
Get predictions on
Test dataset and
display model
evaluation results.

Results
The model performed well with an accuracy of 91.99% on training dataset and 82.68% on validation dataset.

Conclusion
Hence, a machine learning model
using Random Forest Classifier has
been build, that predicts the type of
people who survived the Titanic
shipwreck using the passenger data.