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

"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.)."

Submitted By

76	Atharva Mohan Tirkhunde	72145918J

Guided by
Ms. Shilpa Pimpalkar

Project Aim

To 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.).

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

Data Pre-processing:

- Data Preprocessing is a critical step in the machine learning pipeline that involves cleaning and transforming the raw data into a format suitable for model training and evaluation. It includes the following key aspects:

- Data Cleaning: Data cleaning is the process of identifying and addressing missing, inconsistent, or erroneous data. Techniques such as imputation, removal of outliers, and correction of data anomalies are applied.

- Data Scaling and Normalisation: Scaling and normalisation are used to standardise the numerical features. Scaling ensures that all features have the same scale, preventing certain features from dominating others during model training.

- Handling Categorical Data: Categorical data, such as gender or class, needs to be converted into numerical format for machine learning models. This is typically done through techniques like one-hot encoding or label encoding.

- Feature Engineering: Feature engineering involves creating new features or transforming existing ones to capture more meaningful information

- Data Splitting: The dataset is typically split into training and testing sets. The training set is used to train the machine learning model, while the testing set is reserved for evaluating its performance.

- Data Visualization: Data visualisation is used to explore and understand the dataset. Visualisation techniques include histograms, scatter plots, and correlation matrices, aiding in feature selection and model understanding.

K-Nearest Neighbors (KNN)

suitable for both classification and regression tasks. KNN is simple to implement but can be sensitive to the choice of k and is computationally expensive for large datasets. The k-nearest neighbour classifier fundamentally relies on a distance metric. The better that metric reflects label similarity, the better the classification will be. The most common choice is the Minkowski distance.

Decision Trees

It is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. The decisions or the test are performed on the basis of features of the given dataset. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. Below diagram explains the general structure of a decision tree:

Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent given the class and models data using Gaussian distributions. It is simple, computationally efficient, and works well with high-dimensional data.

Support Vector Classifier (SVC)

SVC is a powerful classification algorithm that aims to find a hyperplane that best separates data points into different classes. It is effective for both linear and non-linear classification tasks when used with kernel functions. For multi-class classification problems, SVC can be extended using one vs-one or one-vs-all strategies, where multiple binary classifiers are combined to make multi-class predictions.

Extremely Randomised Trees

Extremely Randomised Trees, is an ensemble learning method that builds multiple decision trees using random feature subsets and random thresholds. It introduces additional randomness compared to traditional Random Forests, which can reduce overfitting and improve model generalisation. Extra Trees are computationally efficient and effective for high-dimensional data.

Gradient Boosting

Gradient Boosting is an ensemble learning technique that combines weak learners, typically decision trees, into a strong predictive model. It works by sequentially adding models that correct the errors of the previous ones. However, they can be computationally intensive and may require careful parameter selection.

Conclusion:

Machine learning can be a powerful tool for predicting the survival of Titanic passengers. By following the steps above, you can build a machine learning model that can accurately predict the type of people who are most likely to survive a shipwreck.