SAVITRIBAI PHULE PUNE UNIVERSITY

The Mini Project Based On

Build a machine learning model

Submitted By:

Siddhi Sachin Diwadkar

Seat No: A-29

Under Guidance of:

Prof. A.P Bhalke

In partial fulfillment of

Laboratory Practice-III (310258)

DEPARTMENT OF COMPUTER ENGINEERING)

SAVITRIBAI PHULE PUNE UNIVERSITY 2024-25

CERTIFICATE

This is to certify that the Mini Project based on,

Build a machine learning model

has been successfully completed by,

Name: Siddhi Sachin Diwadkar

Exam seat number: A-29

Towards the partial fulfilment of the Final Year of Computer Engineering as awarded by the Savitribai Phule Pune University, at PDEA's College of Engineering, Manjari Bk," Hadapsar, Pune 412307, during the academic year 2024-25.

Prof. A.P Bhalke	**Dr. M. P. Borawake**
Guide Name	**H.O.D**

Acknowledgement

My first and for most acknowledgment is to my guide Prof. A.P Bhalke During the long journey of this study, she supported me in every aspect. She wasthe one who helped and motivated me to proposer search in this field and inspired me with her enthusiasm on research, her experience, and her lively character.

I express true sense of gratitude to my guide Prof. A.P Bhalke for her perfect valuable guidance, all the time support and encouragement that he gave me.

I would also like to thanks our head of department Dr. M. P. Borawake and Principal Dr. R. V. Patil and management inspiring me and providing all lab and other facilities, which made this mini project very convenient.

I thankful to all those who rendered their valuable help for successful completion on Internship presentation.

Name: Siddhi Sachin Diwadkar

Index

Abstract

This project focuses on the development of a machine learning model to predict the survival of passengers aboard the Titanic, utilizing the well-known Titanic dataset. By employing various data preprocessing techniques, including handling missing values and encoding categorical variables, we prepare the dataset for analysis. Multiple classification algorithms, such as Logistic Regression, Decision Trees, and Random Forest, are trained and evaluated to determine their predictive performance. Feature selection methods are also applied to identify the most significant attributes influencing survival outcomes. The model's effectiveness is assessed using metrics such as accuracy, precision, recall, and F1-score, along with cross- validation techniques to ensure robustness. Ultimately, this project aims to provide insights into the factors affecting survival on the Titanic, demonstrating the application of machine learning in historical data analysis.

Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most tragic maritime disasters in history, resulting in the loss of over 1,500 lives. The event has captivated public interest for over a century, leading to extensive analysis of the factors that contributed to survival. This project aims to leverage machine learning techniques to predict which passengers survived the disaster based on a variety of attributes, including demographic information, ticket class, and embarkation details.

The Titanic dataset, widely used in data science education and practice, offers a rich source of information for exploratory data analysis and predictive modeling. By applying statistical and machine learning methods, we can uncover insights into the characteristics that influenced survival rates. This not only serves as an engaging application of machine learning but also provides a historical perspective on the societal dynamics of the early 20th century.

In this study, we will implement a series of preprocessing steps to clean and prepare the data, followed by the application of multiple classification algorithms. The performance of these models will be rigorously evaluated, allowing us to identify the most effective approach for predicting survival. Through this endeavor, we aim to enhance our understanding of the Titanic disaster while demonstrating the practical applications of machine learning in historical context.

Objectives

1. Data Preprocessing:

- Clean the Titanic dataset by handling missing values, outliers, and irrelevant features.
- Encode categorical variables (e.g., Sex, Embarked) to make them suitable for machine learning algorithms.

2. Exploratory Data Analysis (EDA):

- Analyze the dataset to uncover trends and relationships between passenger attributes and survival rates.
- Visualize the data to identify significant features influencing survival (e.g., age, gender, class).

3. Model Development:

- Implement multiple machine learning algorithms, including: Logistic Regression, Decision Trees, Random Forest
- Optimize model parameters using techniques such as Grid Search and Cross-Validation.

4. Feature Selection:

- Apply feature selection methods (e.g., Recursive Feature Elimination, Feature Importance) to identify the most relevant attributes impacting survival.
- Enhance model performance and interpretability by reducing irrelevant features.

5. Model Evaluation:

- Evaluate the performance of the developed models using various metrics, including: Accuracy, Precision, Recall, F1-Score
- Compare the effectiveness of different algorithms to determine the best-performing model for predicting survival.

6. Insights and Reporting:

- Provide insights into the factors that contributed to survival on the Titanic (e.g., class, age, gender).
- Prepare a comprehensive report summarizing findings, methodologies, and implications for understanding historical data through machine learning.

7. Educational Value:

- Demonstrate the application of machine learning techniques in a real-world context.
- This project serves as a learning resource for data science practitioners and enthusiasts, highlighting the importance of data preprocessing, feature selection, and model optimization.

System Specification

Software Requirement:

- Python Libraries
- Jupyter Notebook
- Google Collab

Hardware Requirement:

- Processor - 11th Gen Intel(R) Core(TM) i5-1155G7 @ 2.50GHz 2.50 GHz
- System Type - 64-bit operating system, x64-based processor
- RAM Size - 16 GB

Methodology

1. Data Collection:

- Obtain the Titanic dataset from a reliable source, such as Kaggle or the Titanic database hosted by the University of California, Irvine (UCI). This dataset contains information on passengers, including their demographics, ticket details, and whether they survived.

2. Data Exploration:

- Load the dataset and perform an initial exploration to understand its structure. This includes identifying the number of rows and columns, as well as the types of data present.
- Identify and document the features available in the dataset, such as:
  - PassengerId
  - Name
  - Age
  - Sex
  - Ticket
  - Fare
  - Embarked
  - Survived

3. Data Preprocessing:

- Handling Missing Values:
  - Identify columns with missing data (e.g., Age, Cabin, Embarked) and decide on appropriate strategies to handle them. For instance, impute missing values for the `Age` column using the median or mean, and drop unnecessary columns if needed.

- **Data Encoding:**
  - Convert categorical variables (e.g., `Sex`, `Embarked`) into numerical formats using techniques like One-Hot Encoding or Label Encoding to make them suitable for machine learning algorithms.

- **Feature Scaling:**

  - Normalize or standardize continuous features such as `Age` and `Fare` to improve model performance, especially for distance-based algorithms like Support Vector Machines (SVM).

4. Exploratory Data Analysis (EDA):

- Use visualizations such as histograms, box plots, and correlation matrices to explore the relationships between different features and survival rates.
- Analyze survival rates based on different groups, such as by gender, passenger class, and age, to identify patterns and insights that could influence model development.

5. Feature Selection:

- Apply techniques such as Recursive Feature Elimination (RFE) or feature importance from tree-based models to select the most relevant features for modeling. This step helps enhance model performance by focusing on the most critical attributes that influence survival.

6. Model Development:

- **Dataset Splitting:**

  - Split the dataset into training and testing sets, using a standard split ratio such as 80/20, where 80% of the data is used for training and 20% for testing.

- **Model Training:**

  - Train multiple classification algorithms to predict survival, including:

    - Logistic Regression

    - Decision Trees

    - Random Forest

    - Support Vector Machines (SVM)

- **Model Optimization:**

  - Use techniques such as cross-validation to ensure that the models are robust and generalizable. Cross-validation helps to avoid overfitting by evaluating model performance on different subsets of the training data.

7. Model Evaluation:

- Evaluate each model using various performance metrics, including:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - ROC-AUC (Receiver Operating Characteristic - Area Under Curve)
- Compare the results of these metrics to identify the best-performing model for predicting survival on the Titanic.

8. Interpretation of Results:

- Analyze the results of the models to understand which features had the most significant impact on survival rates. For example, factors such as gender, passenger class, and age are likely to be influential.
- Visualize model performance using metrics like confusion matrices, ROC curves, and feature importance plots for better interpretability.

9. Reporting Findings:

- Document the methodologies, results, and insights gained from the analysis in a comprehensive report.
- Prepare visualizations to clearly present findings, highlighting key factors that influenced survival, and use these visualizations to explain the most relevant features in an understandable manner.

10. Future Work and Improvements:

- Suggest potential improvements or further analyses, such as testing additional algorithms, applying ensemble methods (e.g., Gradient Boosting, XGBoost), or exploring deeper feature engineering to improve model accuracy.
- Additionally, investigating more complex relationships between features, such as interaction terms or polynomial features, could lead to further insights.

By following this structured methodology, the project aims to effectively predict survival on the Titanic and provide valuable insights into the underlying factors that influenced this historic tragedy.

Sample Code

```
# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt import seaborn as sns

from sklearn.model_selection import train_test_split, GridSearchCV from sklearn.preprocessing import OneHotEncoder, StandardScaler from sklearn.impute import SimpleImputer

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score


# Load the dataset

data = pd.read_csv('titanic.csv')

# Explore the dataset

print(data.head())

print(data.info())

print(data.describe())

# Data Preprocessing

# Handle missing values

data['Age'].fillna(data['Age'].median(), inplace=True)

data.drop(columns=['Cabin', 'Ticket'], inplace=True) # Drop unnecessary columns data.dropna(subset=['Embarked'], inplace=True) # Drop rows with missing Embarked

# Encode categorical variables

data = pd.get_dummies(data, columns=['Sex', 'Embarked'], drop_first=True)
```

```
# Feature Selection

features = ['Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'Sex_male', 'Embarked_Q', 'Embarked_S'] X = data[features]

y = data['Survived']

# Split the dataset into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Feature Scaling (optional, depending on the model) scaler = StandardScaler()

X_train = scaler.fit_transform(X_train) X_test = scaler.transform(X_test)

# Model Development: Random Forest Classifier model = RandomForestClassifier(random_state=42) model.fit(X_train, y_train)

# Predictions

y_pred = model.predict(X_test)

# Model Evaluation

print("Accuracy:", accuracy_score(y_test, y_pred)) print("\nClassification Report:\n", classification_report(y_test, y_pred)) print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))

# Visualize Feature Importance

feature_importances = model.feature_importances_

features_df = pd
```

Output

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
```

Future Scope

1. Advanced Modeling Techniques:

- Explore more sophisticated machine learning algorithms, such as Gradient Boosting Machines (GBM) or Neural Networks, to improve predictive accuracy.

2. Hyperparameter Tuning:

- Implement advanced hyperparameter tuning techniques (e.g., Bayesian Optimization) to enhance model performance further.

3. Feature Engineering:

- Investigate additional feature engineering methods, such as creating interaction terms or using domain-specific knowledge to generate new features.

4. Integration of External Data:

- Incorporate external datasets (e.g., weather conditions, ship's route) to enrich the analysis and potentially improve predictions.

5. Real-time Prediction Models:

- Develop a web application or API for real-time predictions of survival based on passenger data, making the model accessible for educational or analytical purposes.

6. Longitudinal Studies:

- Conduct longitudinal studies to compare findings from the Titanic dataset with other historical datasets or modern scenarios to assess changes in survival factors over time.

7. Educational Outreach:

- Create workshops or tutorials based on the project to educate others about the applications of machine learning in historical data analysis.

Conclusion

This project successfully demonstrates the application of machine learning techniques to analyze and predict the survival of passengers aboard the Titanic. Through a systematic methodology encompassing data preprocessing, exploratory analysis, model development, and evaluation, we identified key factors influencing survival rates. The insights gained from this analysis not only deepen our understanding of the Titanic disaster but also showcase the potential of machine learning in uncovering historical patterns.

The findings reveal significant disparities in survival based on attributes such as gender, class, and age, highlighting the social dynamics of the time. By utilizing various classification algorithms, we assessed their predictive performance, with the best models providing valuable insights into which characteristics were most impactful.

Moving forward, the scope for further exploration and enhancement of this project is vast. With advancements in modeling techniques and the potential for integrating external data, the analysis can be refined and expanded, offering even deeper insights. This project serves as a foundation for understanding the interplay between data science and historical analysis, illustrating the relevance and power of machine learning in interpreting past events.

Reference

- https://www.github.com
- https://chat.openai.com/
- https://www.python.org/
- https://www/w3schools.com