TITANIC SURVIVAL PREDICTION: A COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING MODELS

ABSTRACT:

This project predicted Titanic passenger survival using Logistic Regression, Decision Tree, Naïve Bayes, and Artificial Neural Network (ANN) models. ANN demonstrated superior performance among the models. Exploratory Data Analysis revealed higher survival rates for Class 1 passengers, females, aged individuals, and children. The study successfully applied both machine learning and deep learning models to predict survival outcomes. It also identified significant survival patterns, enhancing understanding of the Titanic dataset.

ABOUT DATA:

The data set is a subset of a Titanic passenger dataset. Here's a detailed explanation of each column in the dataset:

1. **Name**: This column contains the names of the passengers. It is of type object, indicating that it contains string data.
2. **survived**: This column indicates whether a passenger survived or not. It is of type object, which means it likely contains categorical string data such as "yes" or "no".
3. **sex**: This column denotes the gender of the passengers. It is of type object, containing string data such as "male" or "female".
4. **age**: This column contains the ages of the passengers. It is of type float64, indicating that it contains numerical data. There are some missing values in this column, as indicated by the count of non-null values (1046 out of 1309).

5. **passengerClass**: This column indicates the class in which the passenger was traveling (e.g., First, Second, or Third class). It is of type object, containing string data.

PROJECT OBJECTIVES

1. **Predict Titanic Passenger Survival**: Develop predictive models to estimate the likelihood of survival for passengers aboard the Titanic.
2. **Comparative Analysis of Models**: Evaluate and compare the performance of various machine learning and deep learning models including Logistic Regression, Decision Tree, Naïve Bayes, and Artificial Neural Network (ANN).
3. **Exploratory Data Analysis (EDA)**: Conduct a thorough EDA to uncover significant patterns and relationships in the Titanic dataset, particularly focusing on factors that influenced survival rates.
4. **Model Performance Evaluation**: Assess the accuracy and effectiveness of each predictive model using appropriate metrics to determine the best performing model.
5. **Identification of Key Survival Factors**: Identify and interpret significant factors influencing survival rates, such as passenger class, gender, age, and other relevant features.

PROJECT OUTCOMES

1. **Successful Survival Prediction Models**: Developed and tested multiple models for predicting survival, with the SVM model demonstrating superior performance compared to traditional machine learning models.
2. **Performance Metrics**: The SVM model outperformed others in key metrics such as accuracy, recall, and F1 score, indicating its effectiveness in predicting survival outcomes.
3. **Insights from EDA**: The EDA revealed critical insights:

    - Higher survival rates were observed among Class 1 passengers.
    - Females had a significantly higher survival rate compared to males.
    - Children and aged individuals showed higher chances of survival.
4. **Enhanced Understanding of the Dataset**: The project provided a comprehensive understanding of the factors that affected survival on the Titanic, contributing valuable insights into historical data analysis.
5. **Practical Application of ML and DL Models**: Successfully applied both machine learning and deep learning techniques to a real-world dataset, showcasing the practical applications and benefits of these methods in predictive analytics.
6. **Significant Survival Patterns**: Identified and confirmed significant survival patterns which could be useful for historical analysis and for developing similar predictive models in other domains.

By achieving these objectives and outcomes, the project not only demonstrated the practical application of predictive models but also contributed to the deeper understanding of the Titanic disaster through data analysis.

Project Description and Findings

This project aimed to predict Titanic passenger survival using a variety of machine learning and deep learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM) with linear and polynomial kernels, and Artificial Neural Network (ANN). Among these models, the SVM with a linear kernel demonstrated superior performance, achieving an accuracy score of 81%.

Key Findings and Insights

1. **Model Performance**:
    - **SVM with Linear Kernel**: Achieved the highest accuracy of 81%, outperforming other models.
    - **Random Forest**: Provided robust performance, capturing complex patterns in the data.
    - **Logistic Regression**: Offered a solid baseline with interpretable results.

- **ANN**: Leveraged deep learning techniques but did not outperform SVM in this case.

2. **Exploratory Data Analysis (EDA)**:
   - **Passenger Class**: Higher survival rates were observed among Class 1 passengers.
   - **Gender**: Females had a significantly higher survival rate compared to males.
   - **Age**: Both aged individuals and children showed higher chances of survival.
   - **Fare**: Higher fare prices correlated with higher survival rates, indicating socioeconomic factors played a role.
   - **Embarked Location**: Passengers who embarked from certain locations had different survival rates, suggesting geographic and possibly socioeconomic influences.
3. **Model Application**:
   - Successfully applied a range of machine learning and deep learning models to predict survival outcomes.
   - Demonstrated the practical application of SVM with a linear kernel as a top performer for this dataset.
4. **Significant Survival Patterns**:
   - Identified and confirmed critical factors influencing survival, providing deeper insights into historical data.
   - Highlighted the importance of passenger class, gender, and age as key determinants of survival on the Titanic.
5. **Methodological Approach**:
   - Employed rigorous cross-validation techniques to ensure the reliability and robustness of the models.
   - Used feature engineering and selection to enhance model performance, including handling missing data and encoding categorical variables.
6. **Future Work and Recommendations**:
   - Explore additional features and interactions to further improve model accuracy.
   - Investigate other advanced machine learning techniques such as ensemble methods and boosting.
   - Conduct a more in-depth analysis of the impact of socio-economic status and other demographic factors on survival rates.

Conclusion

This project aimed to predict Titanic passenger survival using Logistic Regression, Random Forest, Support Vector Machine (SVM) with linear and polynomial kernels, and Artificial Neural Network (ANN) models. The SVM with a linear kernel demonstrated superior performance with an accuracy score of 81%. Exploratory Data Analysis (EDA) revealed higher survival rates for Class 1 passengers, females, aged individuals, and children. The study successfully applied both machine learning and deep learning models to predict survival outcomes. It also identified significant survival patterns, enhancing the understanding of the Titanic dataset. Key factors influencing survival included passenger class, gender, age, fare, and embarkation location. The findings highlight the importance of EDA and the effective application of various modeling techniques to derive meaningful insights from historical data.