Research Paper Draft Based on the Notebook:

Title: ***Random Forest Classification of Titanic Survival Data: A Machine Learning Approach***

Abstract:

This paper presents a machine learning-based classification approach to predict passenger survival in the Titanic disaster using the Titanic dataset. A Random Forest classifier was employed to model the data after performing data preprocessing, transformation, and feature scaling. The model achieved an accuracy of 72%, with 'Fare' and 'Age' emerging as the most important features.

1. Introduction:

The sinking of the RMS Titanic is a well-known historical event that led to the loss of many lives. Predicting passenger survival based on available data can help understand which factors contributed most to survival. In this paper, we employ a Random Forest classifier to build a predictive model using various passenger attributes.

2. Methodology:

2.1. Data Preprocessing: The dataset included various passenger attributes such as 'Age', 'Sex', 'Pclass', and 'Fare'. Columns irrelevant to the analysis, such as 'PassengerId', 'Name', 'Ticket', and 'Cabin', were dropped. Missing values in the 'Age' and 'Fare' columns were imputed using the mean, while the 'Embarked' column was filled with the most frequent value (mode). Categorical variables such as 'Sex' and 'Embarked' were converted to numerical formats for modeling.

2.2. Data Transformation: Feature scaling was applied to numeric features ('Age', 'SibSp', 'Parch', 'Fare') using the StandardScaler to normalize the data, ensuring uniformity across feature magnitudes.

2.3. Data Mining: A Random Forest classifier, a robust ensemble learning method, was chosen for the classification task. The dataset was split into training (80%) and testing (20%) sets. The model was trained on the training set and evaluated on the test set using accuracy, precision, recall, and F1-score metrics.

3. Results:

The Random Forest model achieved an accuracy of 72% on the test set. The classification report indicated balanced performance across the two classes (survived vs. not survived), with the model favoring precision over recall. Feature importance analysis revealed that 'Fare' and 'Age' were the most influential factors in determining survival.

Feature	Importance
Fare	41.88%
Age	38.93%
Pclass	8.15%
SibSp	6.25%
Parch	4.78%
Sex	0%
Embarked	0%

4. Discussion:

The results show that socioeconomic status (represented by 'Fare' and 'Pclass') and age were the most critical factors in survival. Surprisingly, 'Sex' and 'Embarked' did not play a significant role in this model, likely due to the imbalanced dataset or the influence of other factors.

5. Conclusion:

This study demonstrates the effectiveness of Random Forest in classifying Titanic survival data, with accuracy reaching 72%. Future work could explore tuning the model further or testing other algorithms like Gradient Boosting or Support Vector Machines.

References:

- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Kaggle Titanic dataset (https://www.kaggle.com/c/titanic).