Titanic Survival Prediction Project

Key Findings and Insights

Survival on the Titanic was strongly influenced by **gender, passenger class, age, and fare**.

- **Gender:** Women and children had a much higher survival rate compared to men.
- **Class:** 1st class passengers had significantly higher chances of survival than those in 2nd or 3rd class.
- **Age:** Children survived at higher rates than adults, while older passengers had lower survival chances.
- **Fare:** Higher fares (a proxy for socioeconomic status) correlated positively with survival.

These findings align with historical records ("women and children first") and reflect the impact of socioeconomic disparities on survival outcomes.

Model Performance Summary

- **Algorithm Used:** Custom implementation of a Random Forest Classifier.
- **Performance Metrics:**
    - Accuracy: **78.77%**
    - Precision: **86.49%**
    - Recall: **49.23%**
    - F1 Score: **62.75%**
- **Confusion Matrix:** TP = 32, TN = 109, FP = 5, FN = 33

Interpretation:
The model is strong at correctly identifying non-survivors (high precision), but it misses a fair number of actual survivors (lower recall). This trade-off suggests the model is conservative in predicting survival. Feature importance analysis highlighted **Sex, Pclass, Age, and Fare** as the most influential predictors.

Challenges Faced and Solutions

1. **Handling Missing Data**
    - *Challenge:* Missing values in Age, Embarked, and Fare.

- *Solution:* Applied imputation (mean/median for continuous features, mode for categorical).

2. **Feature Engineering**
    - *Challenge:* Raw data lacked predictive clarity.
    - *Solution:* Created age groups, family size, and one-hot encoded categorical variables.
3. **Model Implementation**
    - *Challenge:* Full Random Forest coded from scratch (no pre-built library).
    - *Solution:* Implemented bootstrap sampling, random feature selection, decision trees, and ensemble voting manually.
4. **Bias–Variance Trade-off**
    - *Challenge:* Avoiding overfitting/underfitting.
    - *Solution:* Limited tree depth and validated performance on a separate set.

Future Improvement Suggestions

1. **Cross-validation:** Use k-fold cross-validation for more reliable evaluation.
2. **Advanced Models:** Compare with Logistic Regression, Gradient Boosting, XGBoost, and LightGBM.
3. **Hyperparameter Tuning:** Optimize tree depth, number of estimators, and feature subset size systematically.
4. **Enhanced Feature Engineering:** Extract additional features such as cabin deck, passenger titles, and family survival grouping.
5. **Explainability Tools:** Apply SHAP or LIME for deeper interpretability of model predictions.