Titanic Survival Prediction

📌 Project Overview

This project applies machine learning techniques to predict passenger survival on the Titanic. We explore classification, resampling, and clustering methods to determine the best approach for prediction.

📊 Dataset

- **Source:** Kaggle - https://www.kaggle.com/datasets/shuofxz/titanic-machine-learning-from-disaster
- **Description:** Passenger demographic and ticket information, along with survival outcomes.

❓ Key Questions

- What factors influenced passenger survival the most?
- How accurately can machine learning predict survival?
- How do different models compare in performance?

🚀 Methods Used

◇ Data Preprocessing

- Handled missing values (Age, Embarked)
- Feature engineering (FamilySize, categorical encoding)
- Normalized numerical features

◇ Machine Learning Models

1. **Logistic Regression** - Baseline classification model
2. **Random Forest** - Ensemble learning for better accuracy
3. **K-Means Clustering** - Unsupervised learning for pattern discovery

◇ Model Evaluation

- Metrics: Accuracy, Precision, Recall, F1-score
- Feature importance analysis
- Clustering visualization

📈 Results Summary

✅ **Random Forest performed best (81% accuracy)**
✅ **Key survival factors:** Gender, class, fare
✅ **K-Means clustering showed passenger groupings by ticket class and fare**
❗ **Limitations:** Dataset size, historical bias

🛠️ Installation & Usage

```
# Clone the repository
git clone https://github.com/

# Navigate to the project directory
cd Titanic-ML-Project

# Install dependencies
pip install -r requirements.txt

# Run Jupyter Notebook
jupyter notebook
```

- Open `Titanic_Analysis.ipynb` and execute all cells.

📂 Project Structure

```
/ Titanic-ML-Project
  ├── data/          # Dataset files
  ├── notebooks/     # Jupyter Notebooks
  ├── results/       # Graphs, reports, visualizations
  ├── README.md      # Project overview & instructions
```

🔍 Future Improvements

- Implement deep learning models for higher accuracy
- Use hyperparameter tuning for better Random Forest performance
- Incorporate external datasets for a more comprehensive analysis

👥 Contributors

Musa Malhi

📜 License