Assignment 2: End-to-End Machine Learning Pipeline (Titanic Dataset)

1. Dataset Insights

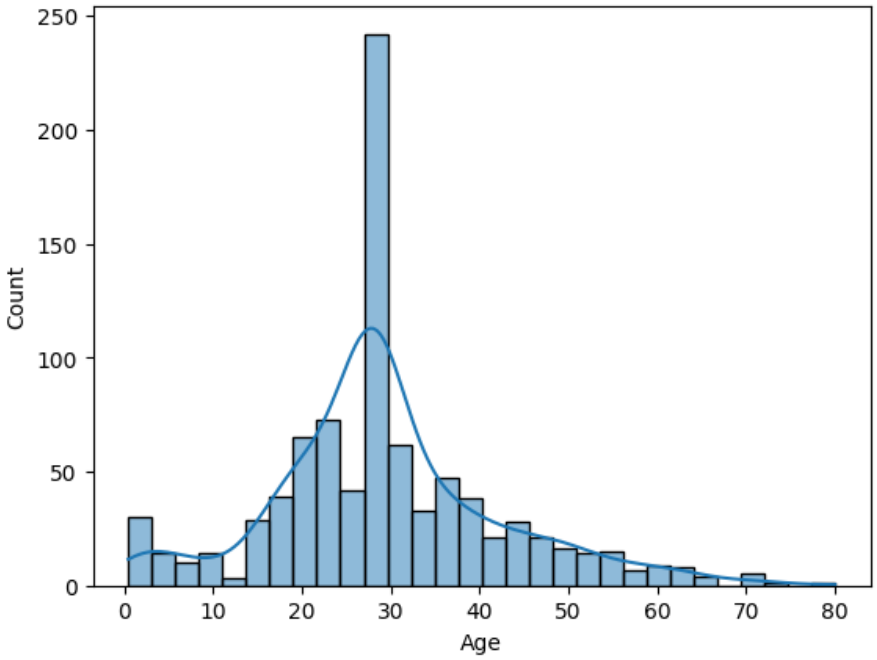
- Source: Kaggle Titanic Dataset (https://www.kaggle.com/c/titanic/data)
- Shape: 891 rows × 12 columns
- Target Variable: Survived (1 = Survived, 0 = Did not survive)
- Features:
    - Numerical → Age, Fare, SibSp, Parch
    - Categorical → Sex, Embarked, Pclass
- Missing Values:
    - Age (~20%) → filled with median
    - Embarked (2 missing) → filled with mode
    - Cabin (~77%) → dropped
- Duplicates: None

---

2. Visualization Findings

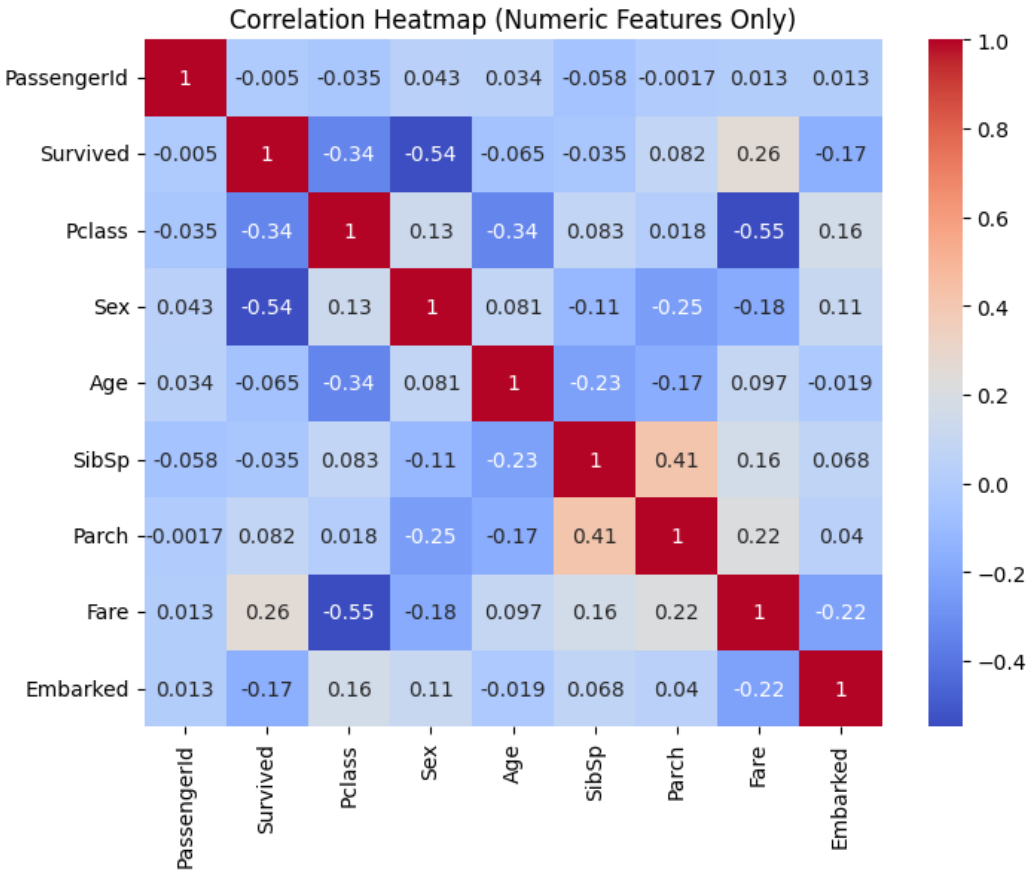
Distributions

- Younger passengers had higher survival rate
- Females had significantly higher survival rate than males
- Higher class (Pclass=1) survived more than lower classes



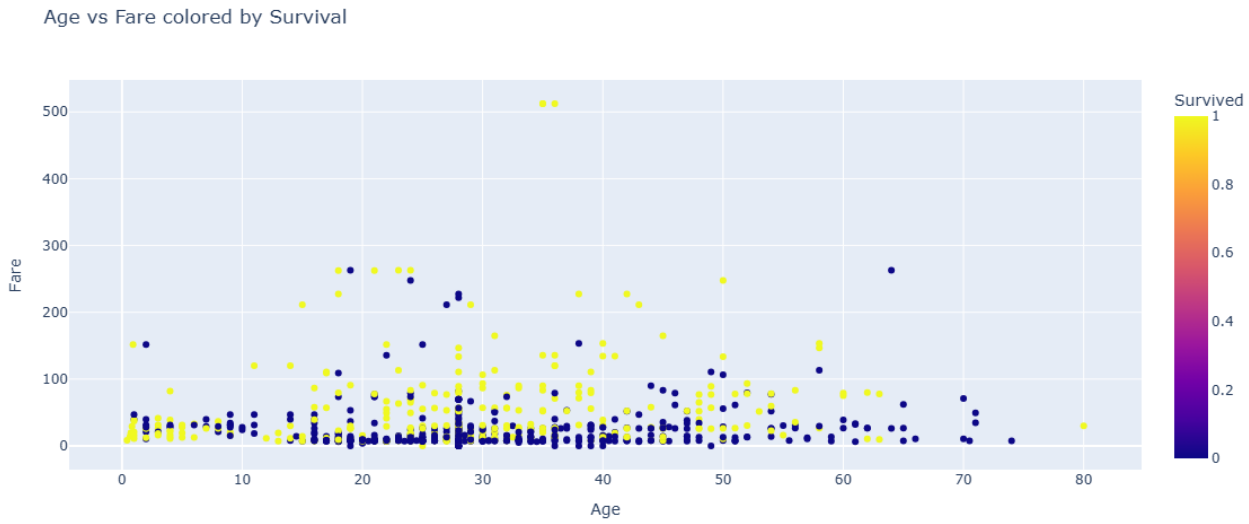
Correlations

- Heatmap showed strong correlation between Fare, Sex, Pclass and survival

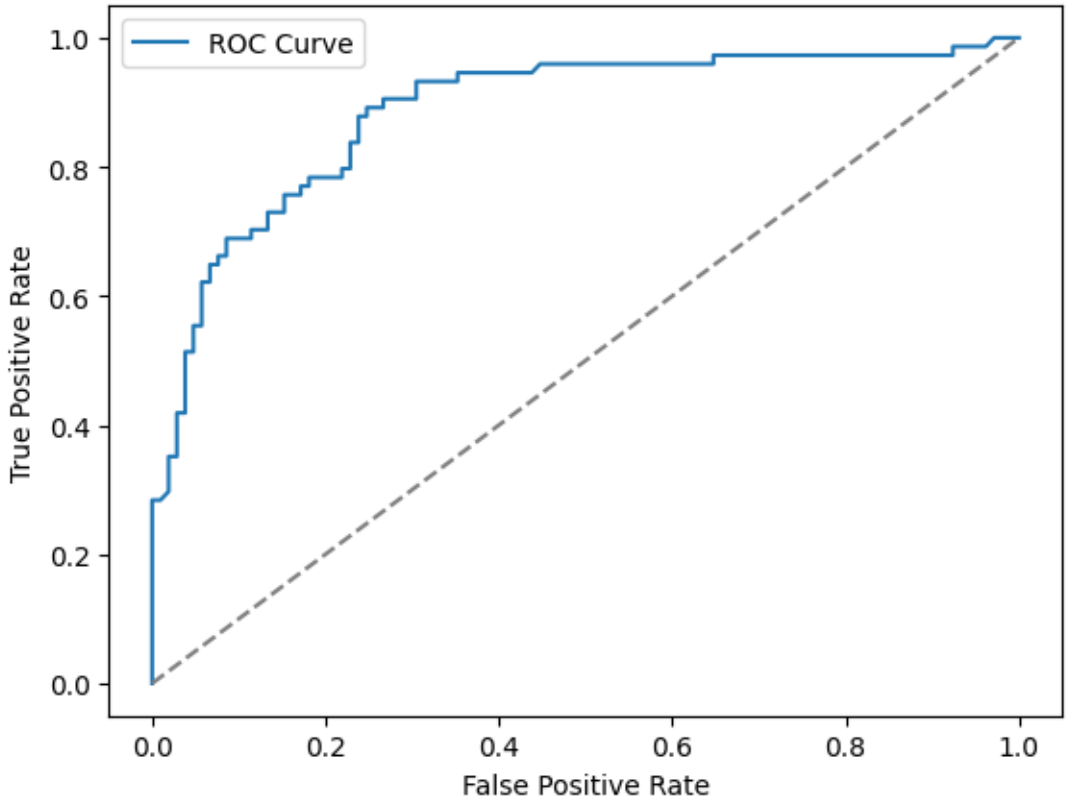


Interactive Analysis

- Scatter (Age vs Fare, color = Survival) → clear separation between survivors and non-survivors



ROC Curve



3. Model Comparison

Model	Accuracy	Precision	Recall	F1-score
KNN (baseline)	0.68	0.67	0.66	0.66
Decision Tree (baseline)	0.70	0.69	0.70	0.69
Random Forest (baseline)	0.78	0.77	0.78	0.77
KNN (tuned)	0.73	0.72	0.72	0.72
Decision Tree (tuned)	0.77	0.76	0.77	0.76
Random Forest (tuned)	**0.82**	**0.81**	**0.82**	**0.81**

---

4. Feature Importance (Random Forest)

- Most important: **Sex, Fare, Pclass, Age**
- Less important: SibSp, Parch, Embarked
- Visualization clearly showed Sex as the strongest predictor

---

5. Hyperparameter Tuning

- **KNN** → n_neighbors=11, metric=manhattan → Accuracy ↑ from 0.68 → 0.73
- **Decision Tree** → max_depth=10, min_samples_split=5 → Accuracy ↑ from 0.70 → 0.77
- **Random Forest** → n_estimators=200, max_depth=10, min_samples_split=5 → Accuracy ↑ from 0.78 → 0.82

---

6. Conclusion

- **Best Model:** Tuned Random Forest (Accuracy ~82%, AUC ~0.86)
- **Reason:** Balanced performance across accuracy, precision, recall, F1-score
- **Top Features:** Sex, Fare, Pclass, Age
- **Impact of Tuning:** Clear improvements in performance for all models
- **Key Insight:** Survival most strongly influenced by gender, class, and age

---