Project Report: Titanic Dataset Analysis and Survival Predictions

Intern Name: Chukka Ganesh
Internship Title: Data Analytics Internship
Internship Duration: April 4 – June 4, 2025

Project Title: Titanic Dataset Analysis and Survival Predictions

---

Acknowledgment

I would like to express my sincere gratitude to my mentors and organization for providing me with the opportunity to work on this data science project. Their guidance and support were invaluable throughout the internship.

Table of Contents

---

1. Introduction

The Titanic dataset is a classic data science problem that involves predicting the survival of passengers based on various features like age sex, ticket class, etc. This project uses machine learning techniques to explore and model the survival of passengers aboard the Titanic.

---

2. Objective

To analyze the Titanic dataset and build a classification model to predict whether a passenger survived the Titanic disaster using features like age, gender, class, and fare.

---

3. Technology Stack

- **Programming Language:** Python
- **Development Tools:** Jupyter Notebook
- **Libraries Used:**
    - pandas
    - numpy
    - matplotlib
    - seaborn
    - scikit-learn
    - xgboost

---

4. Methodology

1. **Data Collection:**
   Used the provided `train.csv` and `test.csv` datasets containing passenger information.
2. **Data Preprocessing:**
    - Handled missing values
    - Encoded categorical variables using Label Encoding
    - Visualized data to understand relationships
3. **Model Building:**
   Trained a classification model using the XGBoost algorithm.
4. **Evaluation:**
   Evaluated model accuracy and plotted confusion matrix.

---

5. Implementation Details

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report

# Load datasets
train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')

# Encode categorical features
le = LabelEncoder()
train_df['feature2'] = le.fit_transform(train_df['feature2'])
test_df['feature2'] = le.transform(test_df['feature2'])
```

```
# Prepare data
X = train_df[['feature1', 'feature2']]
y = train_df['target']
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train model
model = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
model.fit(X_train, y_train)

# Evaluate
y_pred = model.predict(X_val)
print("Accuracy:", accuracy_score(y_val, y_pred))
print(classification_report(y_val, y_pred))

# Confusion Matrix
cm = confusion_matrix(y_val, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix')
plt.show()
```

---

6. Results and Evaluation

- The XGBoost classifier provided a good accuracy on the validation set.
- The confusion matrix and classification report showed balanced performance across classes.
- Visualizations helped to understand data distribution and feature impact.

---

7. Conclusion

This project successfully demonstrated the use of data analysis and machine learning to solve a classification problem using the Titanic dataset. The XGBoost model performed well in predicting passenger survival with visual support from plots and metrics.

---

8. References

- https://www.kaggle.com/competitions/titanic/data
- https://xgboost.readthedocs.io
- https://seaborn.pydata.org
- https://scikit-learn.org

---

9. Appendix

- The full implementation code is available in the Jupyter notebook: `Titanic.ipynb`
- Dataset files: `train.csv`, `test.csv`