



# **MSCS 634 Residency Weekend Project**

Advanced Data Mining for Data-Driven Insights and  
Predictive Modeling

## **Project Deliverable 4 Final Presentation**

### **Team Members:**

Fahreen Dhanani  
Hajeera Hajeera  
Kajol Makhijani  
Sana Magherbi  
Gregory Renteria

Dr. Satish Penmatsa

February 15, 2026

# Agenda

---

- Dataset Overview
- Data Preparation
- Classification Approach
- Classification Results & Insights
- Clustering Analysis

- Association Rule Mining
- Challenges Faced
- Key Takeaways
- Future Improvements



# Dataset Overview

---



- Titanic dataset: passenger demographics and survival outcome
- Key features: Age, Gender, Pclass, Fare, FamilySize, IsAlone
- Mix of numeric and categorical data suitable for data mining

# Data Preparation

```
# Load dataset
df = pd.read_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv')

# Handle missing values
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])

# Drop Cabin due to excessive missing values
df = df.drop(columns=['Cabin'])

# Remove duplicates
df = df.drop_duplicates()

# Quick overview
print(df.shape)
df.head()
```

(891, 11)

|   | PassengerId | Survived | Pclass | Name  | Sex    | Age  | SibSp | Parch | Ticket           | Fare    | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|----------|
| 0 | 1           | 0        | 3      | Braund, Mr. Owen Harris                           | male   | 22.0 | 1     | 0     | A/5 21171        | 7.2500  | S        |
| 1 | 2           | 1        | 1      | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1     | 0     | PC 17599         | 71.2833 | C        |
| 2 | 3           | 1        | 3      | Heikkinen, Miss. Laina                            | female | 26.0 | 0     | 0     | STON/O2. 3101282 | 7.9250  | S        |
| 3 | 4           | 1        | 1      | Futrelle, Mrs. Jacques Heath (Lily May Peel)      | female | 35.0 | 1     | 0     | 113803           | 53.1000 | S        |
| 4 | 5           | 0        | 3      | Allen, Mr. William Henry                          | male   | 35.0 | 0     | 0     | 373450           | 8.0500  | S        |

- Filled missing Age with median, Embarked with mode
- Dropped Cabin column due to excessive missing values.
- One-hot encoded categorical variables
- Engineered features: FamilySize and IsAlone

# Classification Approach

---

- Models used: Decision Tree, K-Nearest Neighbors (KNN)
- Train-test split for model evaluation
- Feature scaling applied for distance-based models
- Hyperparameter tuning for Decision Tree (GridSearchCV)

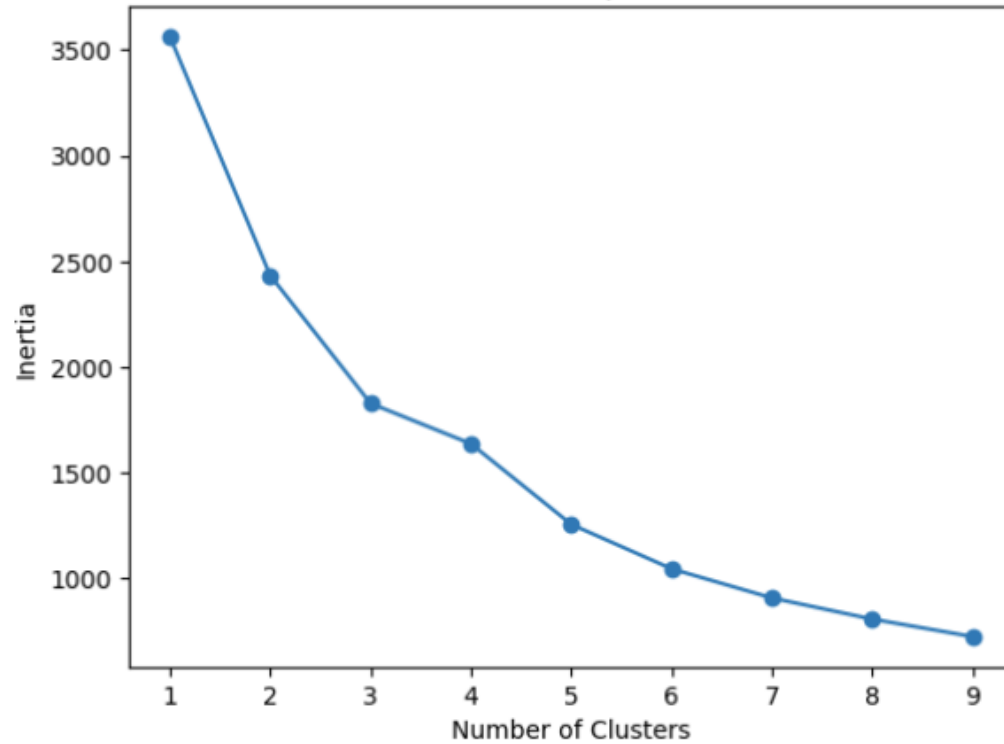
# Classification Results & Insights



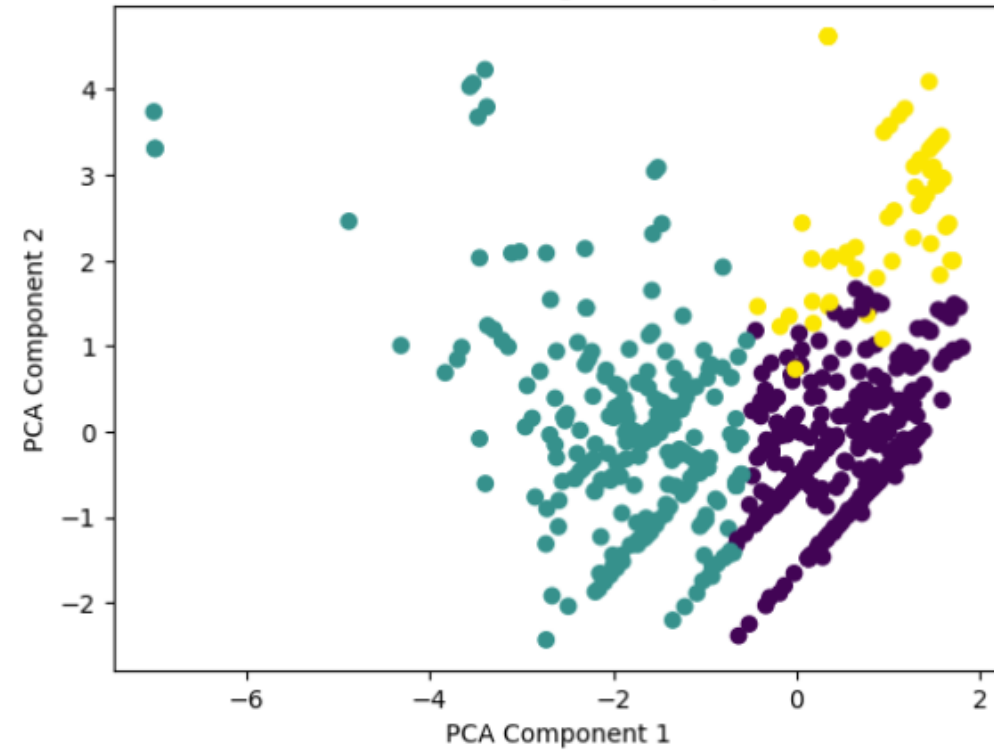
|   | Model               | Accuracy | F1 Score |
|---|---------------------|----------|----------|
| 0 | Decision Tree       | 0.726257 | 0.666667 |
| 1 | KNN                 | 0.821229 | 0.774648 |
| 2 | Tuned Decision Tree | 0.798883 | 0.739130 |

- KNN achieved the highest accuracy and F1 score
- Decision Tree tuned for interpretability
- Feature importance: Gender, Passenger Class, Fare most predictive
- Results align with historical survival patterns

Elbow Method for Optimal Clusters



K-Means Clustering (PCA Projection)



# Clustering Analysis

# Association Rule Mining

---

- Apriori algorithm applied to categorical features
- Rules highlighted relationships between Survival, Gender, Class, and Embarkation
- Male, third-class passengers = lower survival
- Female, first-class passengers = higher survival

# Challenges Faced

---

- Handling missing values without biasing the data
- Feature engineering for better model performance
- Scaling issues for KNN
- Iterative hyperparameter tuning for Decision Tree

# Key Takeaways

---

- KNN: best predictive performance
- Decision Tree: interpretable feature importance
- Clustering: meaningful segmentation
- Association rules: validated known survival patterns



# Future Improvements



---

- Explore advanced models: Random Forests, Gradient Boosting
- Additional feature engineering for complex interactions
- Techniques to detect and mitigate bias
- Expand dataset for richer predictive insights

# Conclusion

---

- Combined use of classification, clustering, and association rules
- KNN: high predictive accuracy; Decision Tree: interpretable
- Clustering and rules provided meaningful descriptive insights
- Ethical and responsible modeling ensures reliable, actionable results

**Thank you**

