Titanic Data Analysis Project

CS379

Nickolas Marquez

2/27/2025

Contents

[Section 1: Unsupervised Clustering Using K-Means 3](#_Toc191587992)

[Goal: 3](#_Toc191587993)

[Steps Taken: 3](#_Toc191587994)

[Diagram: 3](#_Toc191587995)

[Section 2: Supervised Classification using RandomForestClassifier 4](#_Toc191587996)

[Goal: 4](#_Toc191587997)

[Steps Taken: 4](#_Toc191587998)

[Diagram: 5](#_Toc191587999)

[References 6](#_Toc191588000)

# Section 1: Unsupervised Clustering Using K-Means

## Code:

## 

## Goal:

Group passengers into clusters based on similarities in their features (passenger class, gender, age, and fare) without using any labeled outcome. This helps to uncover hidden patterns or groupings within the data.

## Steps Taken:

1. Load and Review Data:

The dataset is read from CSV file and inspected for its shape and missing values.

1. Clean and Prepare Data:

Only the necessary columns are Missing values in ‘pclass’, ‘age’, ‘fare’ are filled with the median. The ‘sex’ column in converted from text to numbers.

1. Feature Scaling:

All features are scaled so that they contribute equally to the clustering process.

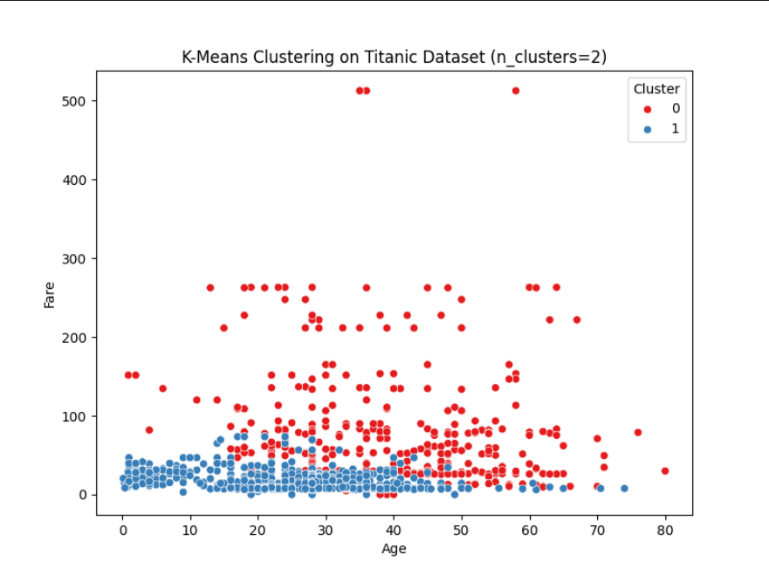
1. Clustering with K-means:

K-Means clustering is applied with 2 clusters (n\_clusters=2), and the resulting cluster labels are added to the data.

1. Visualization:

A scatter plot is generated (with age on the x-axis and fare on the y-axis) to visually show how the clusters are separated.

## Diagram:



# Section 2: Supervised Classification using RandomForestClassifier

## Code:



## Goal:

Predict whether a passenger survived the titanic disaster using labeled data. In this case, the target variable is ‘survived’.

## Steps Taken:

1. Load and Inspect Data:

The dataset is loaded and checked for missing values and overall structure.

1. Data Cleaning and Preparations:

A focused subset of the data is selected (including ‘pclass’, ‘survived’, ‘sex’, ‘age’, and ‘fare’). Missing values in ‘pclass’, ‘age’, and ‘fare’ are filled with median values. The ‘sex’ column is converted to numeric form, and any rows missing the survival outcome are removed.

1. Feature and Target Setup:

The cleaned data is split into features 9X) and the target (y – survived).

1. Optional Feature Scaling:

The features are standardized, even though scaling isn’t strictly needed for RandomForest, to keep preprocessing consistent.

1. Data Splitting:

The data is split into training (80%) and testing (20%) sets.

1. Model Training:

A RandomForesterClassifier is trained on the training data.

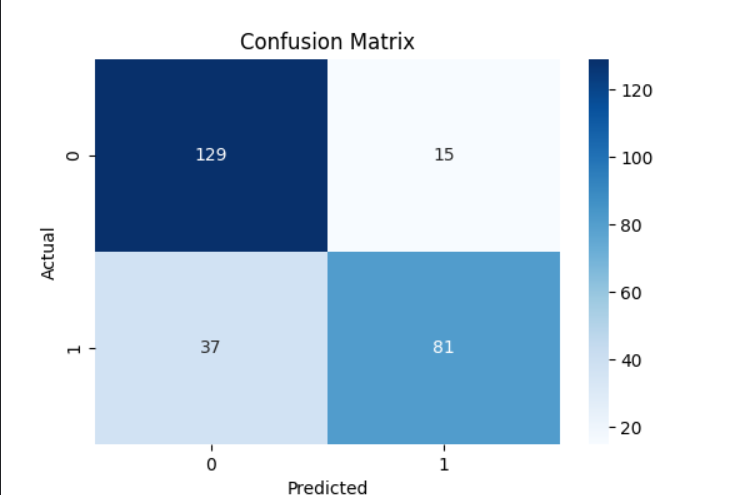
1. Model Evaluation:

The trained model is used to predict on the test set. The performance is measured using accuracy, a confusion matrix, and a detailed classification report (precision, recall, and f1-score).

1. Visualization:

A heatmap is created to visually represent the confusion matrix.

## Diagram:



# References

Mosh. (n.d.). *Python Machine Learning Tutorial (Data Science)*. YouTube. https://www.youtube.com/watch?v=7eh4d6sabA0

Ryan, M. (n.d.). *Beginner Data Science Portfolio Project Walkthrough (Kaggle Titanic)*. YouTube. https://www.youtube.com/watch?v=6IGx7ZZdS74

X, C. (2021, July 22). *Using KMEANS clustering to predict survivors of the Titanic*. Medium. https://tracyrenee61.medium.com/using-kmeans-clustering-to-predict-survivors-of-the-titanic-ae3d3e959eb8

Comment, info, M., GeeksforGeeks, & GeeksforGeeks. (2024, September 23). *Supervised and unsupervised learning*. GeeksforGeeks. https://www.geeksforgeeks.org/supervised-unsupervised-learning/