Task 1 – Titanic Dataset: Model Choice & Evaluation Summary

Objective

The goal of this task was to understand the fundamentals of supervised machine learning using classification algorithms on a real-world dataset. The Titanic dataset was chosen for its balanced class distribution and rich mix of numerical and categorical features.

Model Choice & Justification

We used two supervised classification algorithms:

1. **Logistic Regression**
    - A linear model well-suited for binary classification tasks such as survival prediction.
    - Offers interpretable coefficients and is computationally efficient.
2. **K-Nearest Neighbors (KNN)**
    - A non-parametric model that classifies a test instance based on the majority class among its nearest neighbors.
    - Useful for capturing non-linear relationships in the data.

These models were selected to contrast a **linear classifier** with a **distance-based, non-linear one**, providing a well-rounded understanding of different ML approaches.

Data Preprocessing Steps

- Dropped irrelevant or redundant columns like 'deck', 'who', and 'class'.
- Imputed missing values in 'age' (with median) and 'embarked' (with mode).
- Encoded categorical variables like 'sex' and 'embarked' using one-hot encoding.
- Scaled the features using **StandardScaler** to normalize ranges for both models, especially important for KNN.

Evaluation Metrics Used

Each model was evaluated using:

- **Accuracy Score** – overall prediction correctness.
- **Confusion Matrix** – breakdown of true/false positives/negatives.
- **Classification Report** – including precision, recall, and F1-score.

Results Summary

Metric	Logistic Regression	KNN (k=5)
Accuracy	79.9%	**80.4%**
Precision (class 1)	0.77	0.77
Recall (class 1)	0.73	**0.74**
F1-Score (class 1)	0.75	**0.76**

- Both models performed similarly well.
- **KNN slightly outperformed Logistic Regression** in overall accuracy and F1-score.
- Confusion matrices and prediction distribution plots helped visualize performance gaps.

Insights

- The Titanic dataset is linearly separable to some extent, allowing Logistic Regression to perform well.
- KNN handled the data’s non-linearities slightly better, benefiting from scaled features.
- Data preprocessing, especially scaling and encoding, was critical to both models' success.

Conclusion

This task provided a practical introduction to supervised machine learning. We learned how different models behave on the same dataset, how to evaluate their performance effectively, and how preprocessing significantly influences results.