Insights Report – Sub Task 1: Feature Engineering and Data Transformation

1. Dataset Overview

The Titanic dataset from seaborn was used for this task. It contains demographic and travel details of passengers, with the goal of predicting survival (0 = No, 1 = Yes). Selected relevant features included survival status, passenger class (pclass), sex, age, number of siblings/spouses (sibsp), number of parents/children (parch), fare, and port of embarkation (embarked).

2. Missing Value Imputation

Missing values were primarily found in the 'age' and 'embarked' columns. Two imputation strategies were applied:
- Mean Imputation: Replaced missing values in 'age' with the column's mean.
- KNN Imputation: Encoded categorical values, then used KNN with 3 neighbors to impute missing values.

3. Feature Encoding

Categorical variables ('sex' and 'embarked') were encoded using one-hot encoding, with the first category dropped to avoid multicollinearity. This enabled compatibility with machine learning algorithms.

4. Feature Selection

Three different statistical techniques were used to assess feature importance:
- Chi-Square Test: Applied on binned numerical features. Showed that 'sex_male' and 'pclass' had the highest dependency with survival.
- ANOVA F-Test: Revealed 'fare' and 'sex_male' as highly significant predictors.
- Mutual Information: Confirmed 'sex_male', 'fare', and 'pclass' as top contributors. Mutual information captured non-linear relationships between features and the target.

5. Dimensionality Reduction

Two techniques were used to reduce dimensionality for visualization:
- PCA (Principal Component Analysis): Projected features into two components, showing reasonable separation between survived vs. not.
- t-SNE (t-distributed Stochastic Neighbor Embedding): Produced a more non-linear separation of classes, better capturing clusters of survivors.

6. Conclusion

The data was successfully preprocessed and explored. Statistical tests and visualizations consistently identified 'sex', 'fare', and 'pclass' as the most informative features. Dimensionality reduction techniques provided visual insights into the separation of survivors, indicating that the selected features have strong predictive potential for downstream modeling.