Feature Engineering Report – Titanic Dataset

- **Dataset Chosen**
  The dataset used is the *Titanic Survival Dataset* from Kaggle's "Machine Learning from Disaster" competition.
  It contains 891 passenger records with features such as passenger class, age, gender, ticket fare, and embarkation port.
  Goal: Predict passenger survival (0 = No, 1 = Yes).
  Type: Classification problem.

- **Feature Engineering Techniques Applied**

Step	Technique	Purpose	Reasoning
1	Missing Value Handling	Filled missing Age with median, Embarked with mode, and dropped Cabin	Prevents model errors due to NaN values while preserving data consistency
2	Encoding	Label Encoding (Sex) and One-Hot Encoding (Embarked)	Converts categorical data into numerical format suitable for ML algorithms
3	Feature Scaling	Standardization using StandardScaler	Centers numeric features (Age, Fare)

Step	Technique	Purpose	Reasoning
			to zero mean and unit variance
4	Dimensionality Reduction	PCA with 2 components	Reduces correlated numeric features and helps visualization
5	Feature Selection	SelectKBest (Chi-Squared test)	Selects most influential predictors for survival and reduces noise

- **Reasoning Behind Choices**

1. Median and mode imputation preserve dataset structure without introducing extreme bias.
2. Label and one-hot encoding ensure categorical variables are machine-readable.
3. StandardScaler helps models like Logistic Regression and SVM train efficiently.
4. PCA reduces redundant variance, simplifying the model.
5. SelectKBest filters top predictors (e.g., Pclass, Sex, Fare, Age, Embarked_S).

- **Observations on Transformed Features**

1. `Sex` was the most significant factor: females had higher survival probability.
2. `Pclass` and `Fare` also strongly influenced survival chances.

3. PCA showed two main components explaining most dataset variance.
4. Final dataset became fully numeric and clean, ready for ML modeling.
5. Dataset size after feature engineering: 891 × 8 (approx).

- **Ethical Concerns & Bias Mitigation**

If this preprocessing were used in a real-world predictive model (e.g., employee attrition or loan approval):

1. **Gender or Marital Status** could introduce *algorithmic bias*.
2. Models might unfairly favor or penalize individuals based on these features.

Mitigation Steps:

1. Exclude sensitive attributes (gender, marital status) unless ethically justified.
2. Use **fairness metrics** (demographic parity, equal opportunity) to measure bias.
3. Apply **bias mitigation tools** like IBM AI Fairness 360 or Fairlearn.
4. Maintain transparency — clearly document how features were used.

In the Titanic dataset, gender is historically relevant (social norm "women and children first"),
but in employment or credit models, such features should be excluded.

GROUP MEMBERS:

GUFRAN SHAH 202401110011

PARTH KITCHLOO 202401110014

RAGAV SHARMA 202401110012

NIKHIL SHIRSATHE 202401110003

SANDEEP KOTWAL 202401110052