# ML Project Report: Tasks 8, 9, and 10

## Task 8: Comparing Validation Techniques

### 1. Validation Techniques Implemented:

- Holdout Validation: An 80-20 split was used for training and testing.   
- K-Fold Cross-Validation: K=5 was chosen to balance computational cost and performance.   
- LOOCV: Every record was used for testing once, with the rest for training.

### 2. Performance Metrics:

- Metrics: Accuracy, RMSE, F1-score were used for comparison.   
- Results showed that K-Fold CV provided more stable performance.

### 3. Bias-Variance Tradeoff:

- Holdout Validation: Higher variance due to single split.  
- K-Fold CV: Reduced variance, balanced bias.  
- LOOCV: Low bias but high variance.

### 4. Recommendation:

K-Fold Cross-Validation is recommended for this dataset due to its balance of bias and variance.

## Task 9: Stratified Sampling and Marginal Probability

### 1. Stratified Sampling Approach:

- Stratified sampling was used to ensure proportional representation of target classes.  
- Stratification was applied during train-test splits.

### 2. Marginal Probability:

- Marginal probabilities were calculated for each class and compared to the original dataset distribution.  
- The sample distribution closely matched the original distribution, proving effective representation.

## Task 10: Handling Text and Categorical Attributes

### 1. Handling Categorical Attributes:

- Missing Values: Filled with mode for categorical columns.  
- Encoding:   
 - Label Encoding for ordinal columns (e.g., education level).  
 - One-Hot Encoding for nominal columns (e.g., gender, region).

### 2. Handling Text Attributes:

- Text Preprocessing: Cleaning, lowercasing, and removing special characters.  
- Vectorization: Applied TF-IDF to convert text into numerical features.

### 3. Numerical Feature Scaling:

- Applied Standardization to numerical features for model compatibility.

### 4. New Features:

- Created a derived feature: 'income\_per\_age' by dividing income by age to capture earning capacity per year of life.

## Final Conclusion:

- K-Fold Cross-Validation provided the best balance between bias and variance.  
- Stratified sampling ensured proper representation of classes.  
- Text and categorical features were effectively processed using appropriate techniques.  
- The dataset is ready for model training with optimized features and preprocessing.