Group 3 Project Proposal

Group Members: Beomseong Kim, Will Tappa, Duane Hu, Joy Hong, Taqi Alyousuf

Chosen Dataset

- Heart Failure Prediction
- The dataset contains 918 instances and 12 features.

Variables

- **Age**: Patient's age in years | **Sex**: Gender [M/F]
- **ChestPainType**: Type of chest pain [Typical Angina, Atypical Angina, Non-Anginal Pain, Asymptomatic]
- **RestingBP**: Resting blood pressure [mm Hg] | **Cholesterol**: Serum cholesterol [mm/dl]
- **FastingBS**: Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG**: Resting electrocardiogram results [Normal, ST, LVH]
- **MaxHR**: Maximum heart rate achieved
- **ExerciseAngina**: Presence of exercise-induced angina [Y/N]
- **Oldpeak**: ST depression induced by exercise
- **ST_Slope**: Slope of the peak exercise ST segment
- **HeartDisease**: Target variable indicating presence of heart disease [1: heart disease, 0: Normal]

Codes

```python
import pandas as pd
df = pd.read_csv("heart.csv")
def resumetable(df):
    print(f"data shape: {df.shape}")
    summary = pd.DataFrame(df.dtypes, columns = ['data type'])
    summary = summary.reset_index()
    summary = summary.rename(columns = {"index": "feature"})
    summary["Num_Null"] = df.isnull().sum().values
    summary["Num_Unique"] = df.nunique().values
    summary["First_Value"] = df.loc[0].values
    summary["Second_Value"] = df.loc[1].values
    summary["Third_Value"] = df.loc[2].values

    return summary
resumetable(df)
```

```
data shape: (918, 12)
```

	feature	data type	Num_Null	Num_Unique	First_Value	Second_Value	Third_Value
0	Age	int64	0	50	40	49	37
1	Sex	object	0	2	M	F	M
2	ChestPainType	object	0	4	ATA	NAP	ATA
3	RestingBP	int64	0	67	140	160	130
4	Cholesterol	int64	0	222	289	180	283
5	FastingBS	int64	0	2	0	0	0
6	RestingECG	object	0	3	Normal	Normal	ST
7	MaxHR	int64	0	119	172	156	98
8	ExerciseAngina	object	0	2	N	N	N
9	Oldpeak	float64	0	53	0.0	1.0	0.0
10	ST_Slope	object	0	3	Up	Flat	Up
11	HeartDisease	int64	0	2	0	1	0

Question Ideas:

- Which clinical features are most predictive of heart disease?
- How does stress level impact the likelihood of developing a sleep disorder?
- Which classification algorithm yields the highest accuracy?
- How does feature engineering affect model performance for this dataset?

Methods

- Data Preprocessing
    - clean data by handling potential outliers
    - perform exploratory data analysis to understand variable distributions
    - Check for and address class imbalance of target feature
    - Encode categorical variables + Scale numerical features
- Feature Selection & Engineering
    - Apply correlation analysis to identify relationships between features
    - Use feature importance techniques to identify most predictive variables
    - Create interaction terms between important features
    - Conduct feature selection methods (SelectKBest)
- Model Development
    - Split data into training, validation, and test sets
    - Implement multiple classification algorithms
        - Logistic regression
        - Random Forest (Classifier)
        - Support Vector Machine (SVC)
        - Ensemble Model (Voting)
    - Conduct Hyperparameter Tuning using GridsearchCV + RandomSearchCV
    - Threshold optimization for each model
- Model Evaluation
    - Compare models using evaluation metrics
        - Accuracy, Precision, Recall, F1-Score, ROC Curve, and AUC