Project Overview

The objective of this project is to develop a robust machine learning model for predicting the presence of heart disease in patients. The model will be trained and evaluated on a dataset comprising 14 features, encompassing demographic details and medical attributes. The project will follow a systematic approach, including exploratory data analysis (EDA), feature engineering, model training, evaluation, and refinement. The ultimate goal is to create a model that accurately predicts heart disease, aiding in early diagnosis and intervention.

Dataset Exploration

The dataset used for this project contains 270 entries, each representing a patient, and 14 columns detailing their attributes. The features within the dataset are a mix of numerical and categorical types, and notably, there are no missing values, ensuring data completeness for subsequent analysis and modeling.

- **Age:** This feature represents the age of the patients in years, ranging from 29 to 77, with an average age of 54.43 years. The distribution of age is roughly normal, indicating a balanced representation of different age groups in the dataset.
- **Sex:** This is a binary feature where 0 denotes female and 1 denotes male. The dataset has a higher proportion of males (68%) compared to females (32%).
- **Chest Pain Type:** This categorical feature ranges from 1 to 4, representing different types of chest pain experienced by the patients. The most common types are 3 and 4.
- **BP (Blood Pressure):** Measured in mm Hg (or millimeters of mercury), this feature ranges from 94 to 200, with a mean of 131.34 mm Hg. The distribution is approximately normal, with most values falling between 120 and 150 mm Hg.
- **Cholesterol:** This feature represents cholesterol levels in Milligeams per Decaliter, ranging from 126 to 564, with a mean of 249.66 mg/dL. The distribution is right-skewed, indicating the presence of some individuals with very high cholesterol levels.
- **FBS over 120 (Fasting Blood Sugar):** This binary feature indicates whether the fasting blood sugar level is over 120 mg/dL (1) or not (0). The majority of patients (85%) have FBS levels below 120 mg/dL.
- **EKG Results:** This categorical feature, ranging from 0 to 2, represents the results of an electrocardiogram test. The dataset has a relatively balanced distribution across the three categories.
- **Max HR (Maximum Heart Rate):** This feature represents the maximum heart rate achieved during exercise, ranging from 71 to 202, with a mean of 149.68. The distribution is roughly normal.
- **Exercise Angina:** This binary feature indicates whether the patient experiences exercise-induced angina (1) or not (0). Most patients (67%) do not experience exercise-induced. Exercise-induced angina is a condition that causes chest pain during exercise. It's a common symptom of coronary artery disease, and a good indicator of a heart problem.

- **ST Depression:** This feature measures ST segment depression which is a reading typically found on an ecg results. ranging from 0 to 6.2, with a mean of 1.05. The distribution is right-skewed, with most values between 0 and 2.
- **Slope of ST:** This categorical feature, ranging from 1 to 3, describes the slope of the ST segment. The dataset has a balanced distribution across the three categories.
- **Number of Vessels Fluro:** This feature represents the number of major vessels colored by fluoroscopy, this a medical imaging technique that uses a continuous X-ray beam to create a real-time video of a body part's movement on a monitor. This has a range from 0 to 3, with a mean of 0.67. Most patients have 0 or 1 major vessel colored.
- **Thallium:** This categorical feature, ranging from 3 to 7, represents the results of a thallium stress test. The distribution is concentrated around the values 3, 6, and 7.
- **Heart Disease:** This is the target variable, a binary feature indicating the presence (1) or absence (0) of heart disease.

Visualizations

The visualizations, including histograms, box plots, and correlation matrices, provide valuable insights into the data:

- **Histograms:** These plots reveal the distribution of numerical features, such as age, blood pressure, and cholesterol. The roughly normal distributions of age and blood pressure suggest a typical patient population. The right-skewed distribution of cholesterol indicates the presence of some individuals with exceptionally high levels.
- **Box Plots:** These plots display the distribution of numerical features and highlight potential outliers. Outliers are observed in blood pressure, cholesterol, and ST depression, suggesting the presence of extreme values.
- **Correlation Matrix:** This matrix quantifies the relationships between features and the target variable. Strong positive correlations are observed between "Thallium," "Number of vessels fluro," "Chest pain type," "Exercise angina," and "Heart Disease," indicating that these features are highly associated with the presence of heart disease. Conversely, a negative correlation is found between "Max HR" and "Heart Disease," suggesting that lower maximum heart rates might be linked to a higher risk of heart disease.

Feature Engineering

1. **Outlier Capping**: Outliers in BP, cholesterol, and ST depression were capped at the 95th percentile to improve model stability.
2. **Normalization**: Numerical features were standardized using `StandardScaler`.
3. **Creation Of**: Interaction terms and polynomial features to capture non-linear relationships.
4.

Model Architecture

The script explores various model architectures to predict heart disease:

1. **Baseline Models:** Random Forest and XGBoost classifiers are initially trained and evaluated on the dataset. These models serve as a baseline for comparison with subsequent refined models.
2. **SMOTE-Enhanced Models:** To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data. The Random Forest and XGBoost models are retrained on the balanced dataset, aiming to improve their performance on the minority class (heart disease).
3. **Ensemble Models:** Ensemble learning techniques are employed to combine the predictions of multiple models. Simple averaging of probabilities and stacking with logistic regression as the meta-model are explored. These approaches leverage the strengths of different models to potentially enhance overall predictive accuracy.
4. **Enhanced Models with Feature Engineering:** Polynomial features and scaling are applied to the features to capture non-linear relationships and standardize the data. The Random Forest and XGBoost models are trained on the engineered features.
5. **Stacking with Neural Network Meta-Model:** Stacking is implemented with a neural network as the meta-model. The neural network combines the predictions of the base models (Random Forest and XGBoost) to make the final prediction. This approach aims to leverage the flexibility of neural networks to capture complex patterns in the data.
6. **Stacking with Regularized Neural Network:** L2 regularization is added to the neural network meta-model to prevent overfitting and improve generalization to unseen data. The regularized stacking model is trained and evaluated.

Model Output Analysis

Model Building

- The first baseline moels:
- The Random Forest model: when evaluated on the validation set, achieved an accuracy of 70.37%. The confusion matrix reveals that the model correctly predicted 25 out of 29 instances of non-heart disease (class 0) and 13 out of 25 instances of heart disease (class 1). The classification report indicates a precision of 0.68 for class 0 and 0.76 for class 1, with recall values of 0.86 for class 0 and 0.52 for class 1. This results in f1-scores of 0.76 for class 0 and 0.62 for class 1. Overall, the model demonstrates moderate performance with a tendency to miss a significant number of true positive heart disease cases.
- XGBoost classifier: This model achieved an accuracy of 72.22% on the validation set, as indicated by the confusion matrix, which shows it correctly predicted 26 out of 29 instances of non-heart disease (class 0) and 13 out of 25 instances of heart disease (class 1). The classification report further highlights the model's strengths and weaknesses, with a precision of 0.68 for class 0 and 0.81 for class 1, and a recall of 0.90 for class 0 and again 0.52 for class 1. This resulted in an overall f1-score of 0.78 for class 0 and 0.63 for class 1.

Hyperparameter Tuning

- **Random Forest Best Parameters**: `{'bootstrap': True, 'max_depth': 40, 'min_samples_leaf': 4, 'min_samples_split': 20, 'n_estimators'` (refers to the number of individual models (estimators) that are combined to form the ensemble)`: 200}`
- **XGBoost Best Parameters**: `{'colsample_bytree'` (specific hyperparameter determines the fraction of columns (features) to be randomly sampled for each tree. In this case, 0.6 means that 60% of the total features will be randomly selected for each tree that's built during the boosting process.)`: 0.6, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100, 'subsample': 0.6}`

Enhanced Models

- In comparison, the SMOTE-augmented Random Forest model showed a slightly lower accuracy of 66.67% on the validation set. The confusion matrix shows 23 correct predictions out of 29 for non-heart disease (class 0) and 13 correct predictions out of 25 for heart disease (class 1). The classification report highlights a precision of 0.66 for class 0 and 0.68 for class 1, with recall values of 0.79 for class 0 and 0.52 for class 1. The use of SMOTE aimed to address class imbalance but resulted in a slight decrease in overall accuracy and f1-scores, indicating that while SMOTE helped in balancing the dataset, it also introduced some noise, leading to a reduction in performance.

- The SMOTE-enhanced XGBoost classifier achieved a slightly lower accuracy of 70.37% on the validation set. The confusion matrix shows it correctly predicted 25 out of 29 instances of non-heart disease (class 0) and 13 out of 25 instances of heart disease (class 1). The precision for class 0 was 0.68 and 0.76 for class 1, with a recall of 0.86 for class 0 and 0.52 for class 1. The f1-score was 0.76 for class 0 and 0.62 for class 1.
- The stacking model with a neural network meta-model performs exceptionally well on the test set but can't generalize to the validation set likely overfitting.

Ensemble Learning

- Meta-Model: Linear Regression
- Test Set Performance: The model performs robustly on the test set, achieving an accuracy of 85.19%.
- Both precision and recall are high, with precision at 0.86 for class 0 and 0.83 for class 1, and recall at 0.91 for class 0 and 0.75 for class 1. This indicates that the model can correctly classify both positive and negative cases with high accuracy.
- Validation Set Performance: The model's performance on the validation set is significantly lower, with an accuracy of 68.52%.
- There is a noticeable drop in recall for class 1 (Heart Disease), which is at 0.52, indicating that the model is missing many true positive cases. This suggests that while the model performs well on the test set, it is having a hard time with unseen data.

Stacking with Neural Network

- **Meta-Model**: Neural Network

- Test Set Performance:
- The model performs exceptionally well on the test set, with an accuracy of 90.74%.
- Both precision and recall are high. So, the model can correctly classify both positive and negative cases with high accuracy.
- Validation Set Performance:
- The model's performance on the validation set is significantly lower, with an accuracy of 68.52%.
- Recall for class 1 is still at 52% indicating that the model is still missing many true positive cases.

Conclusion

A consistent challenge observed across all models is the discrepancy between validation and test set performance.

However, this project does successfully develop and evaluated various machine learning models for predicting heart disease. Despite the slight decrease in overall accuracy, the SMOTE-enhanced model provides a more balanced approach by addressing class imbalance issues, leading to improved precision and f1-score for the heart disease class. The trade-off between precision and recall highlights the importance of balancing the dataset to ensure better generalization and robustness in predictions, particularly for the minority class.

Continuing, I would continue

improving the enhanced XGBoost model while fine-tuning the neural net stacking model with different approaches for boosting accuracy on unseen data. Potential strategies for improvement include further hyperparameter tuning, exploring different base models, applying regularization techniques, adjusting the class rates, threshold tuning, augmenting the training data and the biggest factor: enhancing the dataset with more outside data.