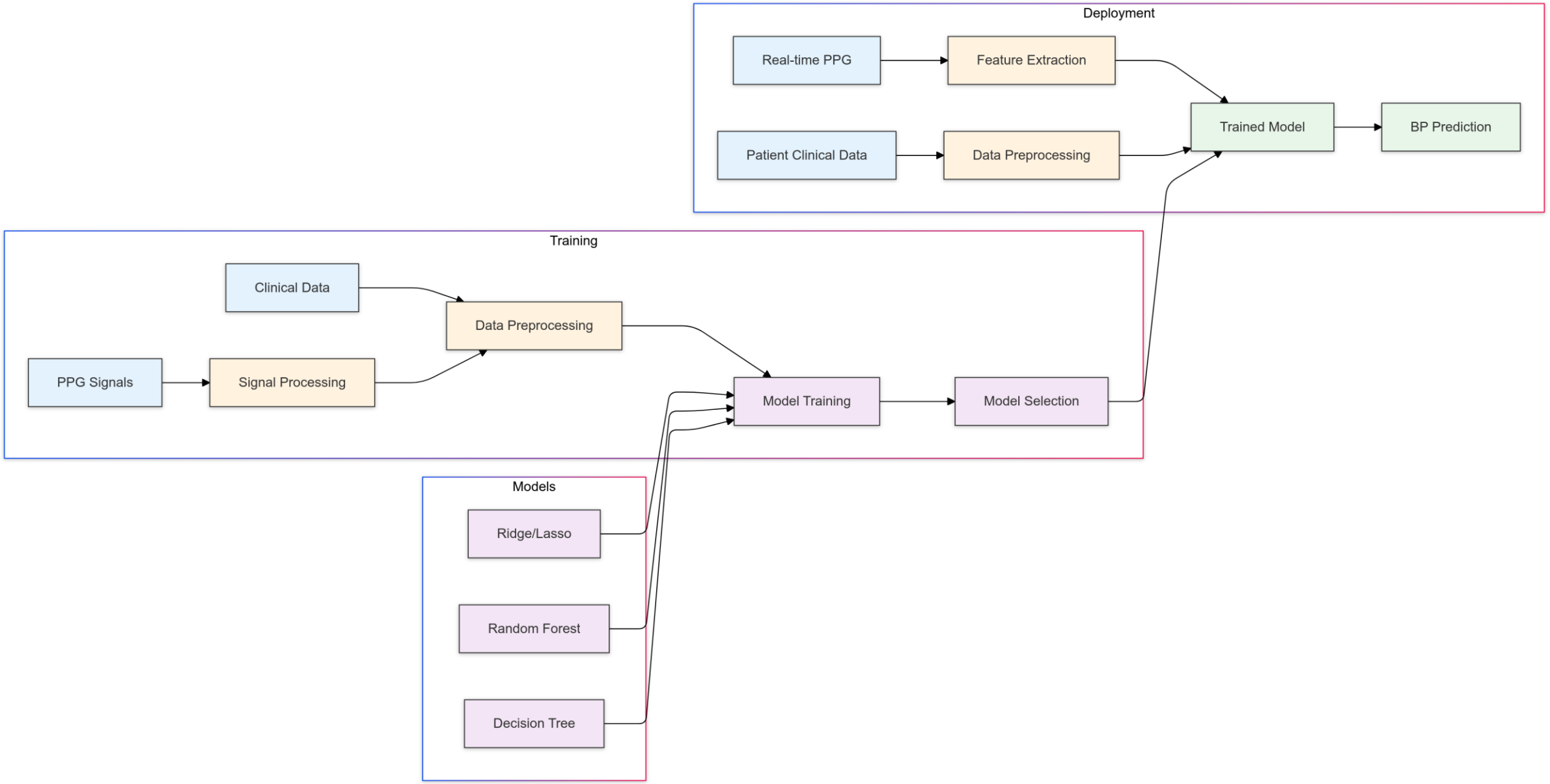
PS2

Team Name : Prompt

Github repo : https://github.com/bhavyagala279/PPG_BP_Signal-Prediction

Architecture Diagram



Feature Extraction from PPG Signals

The first step involved processing raw PPG signals and clinical metadata to create a unified dataset. This dataset (`fin_dataset.csv`) combines signal-derived features with clinical information. It serves as the foundation for accurate blood pressure predictions using machine learning.

- **PPG Signal Preprocessing**:
- Applied Butterworth filter (0.5–5 Hz) to remove noise.
- Extracted statistical features (mean, std, min, max, range) for each signal segment, which describes various features such as Peak-to-Peak Interval (P2P), Peak Amplitude, Variability Features (e.g., std)
- Created a DataFrame grouped by subject_ID.
- **Clinical Metadata Preprocessing**:
- Cleaned missing values in the metadata.
- Standardized subject_ID for consistency.
- **Combining Features**:
- Merged PPG features and clinical metadata using subject_ID.
- Created a final dataset (fin_dataset.csv) for model training.

Data Exploration and Preprocessing

- **Data Preprocessing**: Dropped unnecessary columns (`Num`, `subject_ID`) to focus on relevant features.
- **Binary Categorical Encoding**: Converted binary categories such as Diabetes, cerebral infarction and Sex into 0 and 1 values for model compatibility.
- **One-Hot Encoding**: Applied one-hot encoding to transform other categorical variables into binary columns.
- **Correlation Matrix for Numerical Variables**: Visualized the correlation matrix for numerical features to identify key relationships and avoid multicollinearity.
- **Correlation Matrix for Categorical Variables**: Analyzed the correlation between categorical features to understand their interdependencies.

Cleaned and transformed data into a ready-to-use format for machine learning, ensuring feature consistency and relevance.

Model Training

- **Target and Feature Variables**: Defined y for systolic blood pressure and z for diastolic, while X was created by removing these target columns.
- **Train-Test Split**: Split the data into training and testing sets for both systolic and diastolic blood pressure, using an 80:20 ratio for model validation.
- **Model Selection**: Trained multiple regression models, including Linear, Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, and SVR, for blood pressure prediction.
- **Training and Prediction**: Each model was trained on the training set and tested on the test set to predict systolic and diastolic blood pressure values.
- **Optimization:** The models, including Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, and SVR, were optimized using GridSearchCV with 3-fold cross-validation to find the best hyperparameters. Data scaling was applied specifically to Ridge, Lasso, and SVR models to improve performance. Hyperparameters such as regularization strength and tree depth were tuned to enhance R^2, MAE, and RMSE scores, ensuring the most accurate predictions for both systolic and diastolic blood pressure.
- **Feature Selection and Training :** performed feature selection using ANOVA to identify the most relevant features, but it resulted in decreased accuracy.

Results

Inference for y (Systolic Blood Pressure):

- **R2 Score**: Lasso Regression achieved the highest R2 score (0.8768), indicating that it explains the most variance in the data for systolic blood pressure, followed by Random Forest (0.8765) and Ridge Regression (0.8721).
- **MAE and RMSE**: Decision Tree had the lowest MAE (5.3270), while Lasso Regression and Random Forest performed similarly, both offering competitive MAE and RMSE values.

Top 3 models for y (based on RMSE and MAE):

- **Best by MAE**:
  - 1st: Decision Tree (MAE = 5.3270)
  - 2nd: Ridge Regression (MAE = 5.4442)
  - 3rd: Lasso Regression (MAE = 5.4860)
- **Best by RMSE**:
  - 1st: Lasso Regression (RMSE = 6.5199)
  - 2nd: Random Forest (RMSE = 6.5291)
  - 3rd: Ridge Regression (RMSE = 6.6426)

Inference for z (Diastolic Blood Pressure):

- **R2 Score**: Decision Tree performed the best with an R2 score of 0.5985, followed by Ridge Regression (0.5522) and Lasso Regression (0.5470).
- **MAE and RMSE**: Decision Tree also had the lowest MAE (6.2165), with Ridge Regression and Lasso Regression performing similarly with MAEs just above 6.

Top 3 models for z (based on RMSE and MAE):

- **Best by MAE**:
  - 1st: Decision Tree (MAE = 6.2165)
  - 2nd: Ridge Regression (MAE = 6.3747)
  - 3rd: Lasso Regression (MAE = 6.4334)
- **Best by RMSE**:
  - 1st: Decision Tree (RMSE = 7.5869)
  - 2nd: Ridge Regression (RMSE = 8.0118)
  - 3rd: Random Forest (RMSE = 8.0447)