Tech_Mav - AI Cure Project: PARSEC 4.0

Team Members

Vishwajeet Singh Solanki

Paramananda Bhaskar

Methods and models used during experiment :-

- **Random Forest Feature Importance**
- **Correlation with Heart Rate**
- **LASSO regression and PCA**
- **RandomForestRegressor**
- **Compute Mutual Information**
- **Support Vector Regression (SVR) with RBF kernel**
- **Feature elimination using RFE with Random Forest Regressor**

BEST MODEL : -

- **Feature selection using RFE with Random Forest Regressor**

Random Forest Feature Importance

- A RandomForestRegressor was used to determine the importance of features.
- The most significant features include MEAN_RR, MEDIAN_RR, SDRR, RMSSD, and others.

Correlation with Heart Rate

- Strong negative correlations were found with MEAN_RR (-0.943725) and MEDIAN_RR (-0.927397).
- Features like HF_NU, LF_NU, and HF_LF showed strong positive correlations with HR.
- Correlations were calculated for all features, and their absolute values were sorted in descending order for better interpretation.

Predictive Modeling

- A RandomForestRegressor with 50 estimators was trained on the dataset.
- The model's performance can be evaluated using metrics like mean absolute error, mean squared error, and R-squared on a validation set.

Recommendations

- **Feature Selection**: Based on the feature importance and correlation analysis, select features with the highest significance for model training to improve accuracy and reduce overfitting.
- **Model Optimization**: Experiment with different model parameters and other algorithms (like Lasso, SVR) for better performance.

1. LASSO Regression Analysis:

- **Objective:** To identify significant features affecting heart rate (HR) using LASSO regression.
    - **Methodology:**
    - Standardized features using StandardScaler.
    - Applied LASSO regression with alpha = 0.001.
    - **Results:**
    - The Mean Squared Error (MSE) of the model was 2.2374.
    - The most impactful features based on the absolute value of coefficients were MEAN_RR, HF_LF, LF_NU, SD1, and SDRR.
    - **Interpretation:**
    - Features with higher coefficients (both positive and negative) have a more significant impact on HR.
    - Features like MEAN_RR and SD1 have a strong negative relationship with HR, whereas HF_LF and LF_NU show a positive relationship.
    - **Insight:**
    - LASSO regression helped in feature selection by penalizing less significant features towards zero, providing a clearer view of influential variables.

2. PCA with RandomForest Regression Analysis:

- **Objective:** To predict HR using principal component analysis (PCA) and RandomForestRegressor.
    - **Methodology:**
    - Standardized features and reduced dimensions using PCA (10 components).
    - Trained a RandomForestRegressor on transformed features.
    - **Results:**
    - The Mean Absolute Error (MAE) was 0.5551.
    - The R-squared value was 0.9869, indicating a very high model fit.
    - The first two principal components explained approximately 52.61% of the variance, while all 10 components together accounted for about 95.67%.
    - **Interpretation:**
    - The RandomForest model performed exceptionally well in predicting HR with reduced feature space.
    - The high R-squared value suggests the model explains a large proportion of the variance in HR.
    - **Insight:**
    - PCA effectively reduced the dimensionality of the data without losing significant information, leading to an efficient and accurate model.

Overall Conclusion:
The LASSO regression provided valuable insights into the most significant features affecting HR. Following this, the application of PCA and RandomForestRegressor demonstrated the feasibility of making accurate predictions with reduced feature space. The combination of these methods offers a robust approach to understanding and predicting HR, balancing feature selection, dimensionality reduction, and predictive performance.

Support Vector Regression (SVR) with RBF Kernel

    - **Initial Performance**:
    - Accuracy = 84.22%
    - Root Mean Squared Error (RMSE) = 3.1645
- **Hyperparameter Tuning**: Grid Search identified optimal parameters as C=5.0, epsilon=0.5.
    - **Tuned Performance**:
    - RMSE = 1.4809
    - Accuracy = 96.54%

Mutual Information Analysis

- **Top Features**: 'MEAN_RR', 'MEDIAN_RR', 'SDRR', 'SD2', and 'HF_PCT' showed the highest mutual information scores, suggesting their significant relevance in predicting heart rate.

Conclusions

- **Random Forest with PCA** showed exceptional performance, achieving the highest R-squared value.

- **LASSO Regression** provided valuable insights into feature importance, aiding in understanding the relationship between variables and heart rate.
- **SVR with RBF Kernel**, especially after hyperparameter tuning, demonstrated a significant improvement in accuracy and reduction in RMSE, making it a strong candidate for precise heart rate prediction.

BEST MODELS AND BEST RESULTS

Random Forest Regression with RFE Elimination

- **Selected Features**: 'MEAN_RR', 'MEDIAN_RR', 'SDRR', 'SDRR_RMSSD', 'KURT', 'SDRR_REL_RR', 'RMSSD_REL_RR', 'SDSD_REL_RR', 'SDRR_RMSSD_REL_RR', 'VLF', 'LF_NU', 'HF', 'HF_PCT', 'SD2', 'higuci', 'sampen'
  - **Performance**:
  - ***Mean Squared Error (MSE): 0.0442***
  - ***R-squared: 99.93%***

Random Forest Regression with RFE Selection

- **Selected Features**: 'MEAN_RR', 'MEDIAN_RR', 'SDRR', 'SDRR_RMSSD', 'SDRR_REL_RR', 'RMSSD_REL_RR', 'SDSD_REL_RR', 'SDRR_RMSSD_REL_RR', 'KURT_REL_RR', 'VLF', 'HF', 'HF_PCT', 'HF_LF', 'SD2', 'higuci', 'sampen'
  - **Performance**:
  - ***Mean Squared Error (MSE): 0.0273***
  - ***R-squared: 99.96%***

Analysis of Selected Features

The features selected by RFE in both iterations are predominantly related to heart rate variability (HRV) metrics like 'MEAN_RR', 'SDRR', and 'RMSSD'. These metrics are known to be strong indicators of autonomic nervous system activity, which directly influences heart rate. The inclusion of features like 'HF', 'LF', and their ratios ('HF_LF', 'HF_PCT') further underscores the importance of frequency domain measures in HRV analysis.

Conclusions

- **High Performance**: Both iterations of the Random Forest model with RFE demonstrated exceptional performance with very high R-squared values and low MSE, indicating their robustness in predicting heart rate accurately.
- **Feature Importance**: RFE effectively identified key features that contribute the most to heart rate variability, aligning well with physiological insights.

Recommendations

- **Further Research**: Investigate the physiological significance of the selected features in more depth to understand their impact on heart rate.
- **Model Deployment**: These models, given their high accuracy, are suitable for implementation in real-world applications such as health monitoring systems.
- **Cross-Validation**: Implement cross-validation techniques to further validate the stability and generalizability of the models.

Future Work

- **Model Interpretability**: Explore model interpretability tools to gain deeper insights into how the selected features influence heart rate predictions.
- **Ensemble Techniques**: Consider ensemble methods that combine predictions from multiple models to potentially improve accuracy and robustness.
- **Real-Time Application**: Adapt the models for real-time heart rate prediction, considering their high accuracy and reliability.