Analysis of Relationships between Biometric Variables and Blood Pressure

Lily Rademacher [1] **Bacheler Burt** [1] **Mia Cachion** [1]

Abstract Lily Rademacher

This study investigates how demographic, lifestyle, and clinical factors relate to abnormal blood pressure in a cohort of 1,529 patients from Jamalpur Medical College Hospital in Bangladesh, with the goal of improving early risk stratification for hypertension and cardiovascular disease. These conditions are deeply prevalent within the United States and globally. Over half of American adults experience high blood pressure. Cardiovascular disease, similarly, is the leading cause of mortality in the United States. Both conditions are principal causes of other conditions, such as stroke, kidney failure, and dementia (Couch et al., 2025). Thus, this study factors into the larger body of knowledge on the comorbidities between the above conditions; it underscores the importance of understanding blood pressure and heart problems as they relate to larger public health concerns. Using a cleaned and processed subset of the Jamalpur Medical College Hospital dataset, systolic and diastolic blood pressure and derived blood pressure categories serve as primary outcomes, while predictors include age, sex, body mass index, smoking status, diabetes status, physical activity, cardiovascular risk score, and key laboratory measures. Two complementary modeling strategies are employed: a random forest classifier to predict categorical blood pressure status and linear regression models to estimate continuous systolic blood pressure, with standardized continuous features and one-hot encoded categorical variables supporting robust training and evaluation on test data. The random forest classifier achieves high predictive performance, while maintaining very low rates of clinically dangerous misclassification for severe hypertension. The best random forest configuration explains a large share of variation in blood-pressure–related outcomes, indicating alignment between predicted and observed values. Linear regression models for systolic blood pressure attain an R^2 of approximately 0.80 with moderate prediction error, and coefficient analysis highlights diabetes status and cardiovascular risk score as dominant predictors, consistent with existing cardiovascular epidemiology. Across models, feature importance patterns underscore the central role of systolic and diastolic blood pressure, diabetes, and aggregate cardiovascular risk, while lifestyle and demographic variables play secondary but non-negligible roles. Taken together, the results show that flexible non-linear methods offer superior predictive accuracy, whereas linear models provide transparent parameter estimates that clarify how specific risk factors contribute to elevated blood pressure in this resource-limited clinical setting.

1. Data Bacheler Burt

The dataset used for this study includes a range of health-related variables. These include demographic details, lifestyle factors, and numerical medical measurements. Together, these data enable an analysis of health outcomes. The dataset specifically focuses on patient samples taken in Jamalpur, Bangladesh. 1,529 patient samples are included in the data set, all data was collected from January 20, 2024, to January 1, 2025 (Nirob et al., 2025). All data was gathered within ethical guidelines that ensured both patient confidentiality and informed consent.

1.1. Key Variables

The key variables used for analysis are Sex, Age, Weight, Height, Body Mass Index (BMI), Total Cholesterol, Fasting Blood Sugar, both Systolic and Diastolic Blood Pressure (measured in mmHg) and, by extension, Blood Pressure Category. Cardiovascular Disease Risk, Smoking Status, Diabetes Status and Physical Activity Level will also be used. These variables are essential for identifying patterns that will help predict health risks and inform preventive health strategies.

*Equal contribution [1]University of Virginia, Charlottesville, VA, USA. Correspondence to: Mia Cachion <vwd6uv@virginia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1.2. Data Reading and Preparation for Analysis

The data set contains more variables than we require for our analysis and answering our question. We are exploring the data with the hopes of gaining insight into potential factors that affect blood pressure. For the sake of clarity and focus, we are sacrificing analysis of a broad set of variables in favor of narrowing down several key variables to examine the possible relationships that exist between them and blood pressure. Wrangling and tidying the data will be necessary to accomplish this. This means that many variables must be omitted from our final data set on which we will perform EDA and visualization. Also present in most raw data sets are cosmetic blemishes that will complicate our EDA and visualization processes. These include missing observations, straggling commas, random spaces, inconsistent upper/lower case values, and other factors that may complicate how Python processes these observations. So, tidying will be necessary for EDA procedures to work. The tidying process will be done through the use of python packages, seaborn and matplotlib.

1.3. Data Cleaning

Several methods can be used to wrangle and tidy data that is affected by these imperfections. The packages pandas and numpy will be essential for our wrangling, tidying, and analysis to take place. At the highest level, we'll need to drop several columns and leave only the ones we want to use for our analysis. We can accomplish this by using the df.drop(columns=[...]) function in pandas. As for cleaning the individual observations, functions such as df.dropna() (to drop missing values), .str.strip() (to remove spaces), and .str.lower() (to correct lower case letters) will be needed. However, for these things to be carried out, we will first need to coerce them as strings so that they are read correctly and so we can perform the desired operations on them.

2. Methodology Bacheler Burt

2.1. Methodology Overview

The goal of this analysis is to study the effect of certain biometric data and lifestyle choices ("predictor" variables) on the presence of abnormal blood pressure. A Multiple Regression model will be used to handle the numeric, continuous target variables of interest (Systolic and Diastolic Blood Pressure). A Random Forest Classification model will be used to handle the Blood Pressure Categories target variable of interest given that this is a categorical variable. The results will then be visualized and summarized in an easily-interpretable manner. The next process in the methodology section will be an evaluation of the training procedure for the model, including subset ratios and regression feature logic. As stated, the mix of target variable data types must include different evaluation metrics on a case-by-case basis depending on metric relevance (i.e. RMSE for regression models and an F1 Score for classifier models). Fixed hyperparameters will also be denoted and explained to increase model stability and to ensure reproducibility.

2.2. Model Selection and Justification

To tackle the question at hand, this analysis will employ Multiple Regression and Random Forest analysis techniques in order to gain insight into the relative effects of these metrics on the presence of abnormal blood pressure.

There are several reasons this analysis will use these models. We'll first use Random Forest Classification to handle the "Blood Pressure Categories" target variable of interest. This is a categorical variable, meaning that it is expressed in four discrete categories that correspond to a range of potential precise blood pressure measurements. As a result, this variable will need to be encoded. Still, Random Forest's strength of performing classification analysis makes it the ideal choice in targeting this variable. The corresponding evaluation metrics will be employed (Accuracy, Precision, Recall, and F1 Score). These address a classifier's ability to predict the target variable successfully. For instance, the F1 score is a composite metric taking both precision and recall into consideration.

Multiple regression enables us to factor in multiple variables and assess their relative importance in each of their relative impacts on the outcome. Given Multiple Regressions strength in assessing relative predictor importance of continuous variables, this analysis will employ this model for this specific subtask. The corresponding evaluation metrics (namely the R^2 Value and RMSE) will be used.

Given the nature of medical analyses, interpretability is of the utmost importance as all sectors of society are implicated. It enables a wide range of people, many of whom may not necessarily possess a deep understanding of the potential risk factors affecting their physical health, to extract actionable insights from the analysis. This has major implications for the overall health of a society as it can better enable people to make healthy, informed lifestyle decisions. It can also help elected officials create better-informed health policy to better mitigate these risk factors.

2.3. Model Training Procedure

The modeling strategy evolved during the project to use not only multiple linear regression—as initially planned for continuous blood pressure values—but also random forest classification to categorize patients into clinically meaningful classes: Normal, Elevated, Hypertension Stage 1, and Hypertension Stage 2.

The dataset will be divided into two subsets: a training set

comprising 80 percent of the data (approximately 1,223 patient records) and a test set consisting of the remaining 20 percent (around 306 records), with all modeling and validation conducted exclusively on non-overlapping samples to ensure reliable out-of-sample performance estimation. This split ensures a robust model evaluation, allowing the training subset to inform the regression model while the test subset serves to assess its predictive generalization. Given that the sample size of this project is only 1,529, the authors of this paper believe that a more conservative training to test ratio is most apt. Given the mix of categorical and numeric data in the set, several preprocessing steps are necessary. Categorical variables like smoking status will be encoded numerically using one-hot encoding techniques. Numeric features will be standardized to balance feature scales and optimize algorithm performance, while categorical variables will be converted to numeric format using label encoding for compatibility across model types.

There are also some variables that can be dropped from our model as they are superfluous to the analysis or are too closely related to other variables. Some of these variables include abdominal circumference, waist to height ratio, and estimated low-density lipoprotein levels.

Our regression model regresses the following (the equation will be regressed twice as "Blood Pressure" here represents both Systolic and Diastolic Blood Pressure).

$$\begin{aligned}\text{Blood Pressure} &= \beta_0 + \beta_1(\text{Sex}) + \beta_2(\text{Age}) + \beta_3(\text{Age}^2) \\ &+\beta_4(\text{Weight}) + \beta_5(\text{Height}) + \beta_6(\text{BMI}) \\ &+\beta_7(\log(\text{Total\_Cholesterol})) + \beta_8(\log(\text{Fasting\_Blood\_Sugar})) \\ &+\beta_{11}(\text{BP\_Category}) + \beta_{12}(\text{Smoking\_Status}) \\ &+\beta_{13}(\text{Diabetes\_Status}) + \beta_{14}(\text{Physical\_Activity\_Level}) \\ &+\beta_{15}(\text{Family\_History\_CVD}) + \beta_{16}(\text{CVD\_Risk\_Level}) + \varepsilon\end{aligned}$$

The rationale for having both age and age squared is that the risk of cardiovascular disease grows increasingly with age, not at a constant rate (Zhu et al., 2010). have shown the relationship between age and maximum heart rate to be one where estimated MHR has a quadratic relationship to age; maximum heart rate, which is a conceptual proxy for heart health, is shown to decrease slower before age 40 and faster after age 40. Therefore, using a quadratic form of age should lead to better model fit. We have chosen to express cholesterol as a log variable because cholesterol has been shown to have a diminishing margin effect at higher concentrations, and transforming the variable into a log will correct skewness and approximate a more accurate relationship with cardiovascular outcomes (Wang et al., 2021).

2.4. Model Validation Plan

This section of the analysis will validate the model, assessing whether it has overfit or underfit the values by comparing training and test set performance. This section will also evaluate model performance generally, in order to confirm that the model structure explains effectively the relationship between the independent and dependent variables.

In the Random Forest Classification, accuracy, precision, recall, and F1-score will be measured only on the test data to prevent information leakage and overfitting. Confusion matrices and feature importance rankings will be incorporated to provide visual and quantitative assessment of quality. Visualization tools such as confusion matrices (via seaborn and matplotlib) will be used for diagnostic checks, reinforcing the model's internal validity and interpretability within biomedical contexts. Additionally, accuracy, precision, recall, and F1 score will be employed to evaluate model performance. Precision measures the number of predicted positive cases that are truly positive. Recall measures the amount of positive cases the model actually catches. Finally, accuracy measures the proportion of all predictions that were actually correct. Together, they account for subtle differences in classifier evaluation.

For our regression models, the relevant performance metrics will include the coefficient of determination, otherwise known as R^2, and root mean squared error (RMSE). R^2 quantifies how much of the variance in blood pressure can be explained by the explanatory variables, such as BMI, smoking, and systolic blood pressure. RMSE complements R^2 by expressing average prediction errors, allowing a direct interpretation of error magnitude.

This thorough, future-oriented approach will enable nuanced analysis of both blood pressure prediction and multi-class risk category classification, establishing a rigorous technical foundation for downstream clinical and epidemiological interpretation. It is important that this model is both statistically sound as well as biomedically sound, and that its implications stand up to logical reasoning based on the research we have seen surrounding cardiovascular disease risk.

2.5. Implementation Details and Reproducibility

As described earlier, the data set was randomly split into a common-place 80/20 train/test group to promote model robustness. To do this, scikit-learn's train_test_split function was used. We elected to keep a random_state seed value of 42 for consistency and reproducibility across all models. For the Random Forest Classification, the analysis used scikit-learn's RandomForestClassifier with the same seed value (42). Additionally, the "n_estimators" parameter was set to 100 in order to create a fitting and replicable number of decision trees. The same seed value was selected for our regression models.

The analysis will be conducted using Python within a Google Colab environment. Key libraries will include pandas and numpy for data wrangling, matplotlib and seaborn

for data visualization, and scikit-learn for model building, preprocessing, and evaluation. To ensure reproducibility, random seeds will be fixed using np.random.seed(). Per the project instructions, all scripts will be documented through GitHub commits for transparency and traceability. Code organization will follow a modular approach, separating data cleaning, preprocessing, model training, evaluation, and visualization and interpretation of results into distinct Python scripts or notebook sections. Each section will include detailed comments and markdown annotations explaining procedural logic.

Model	Hyperparameter		
	Seed Value	Estimators	Positive (Allows only positive feature importances)
Random Forest	42	100	False
Multiple Regression	42	N/A	False

Table 1. Hyperparameters by Model

3.1. Random Forest for Blood Pressure Category

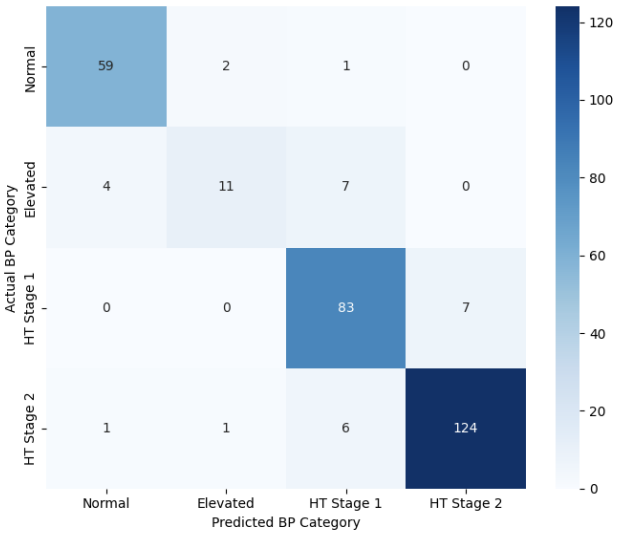

Figure 1. Random Forest - Confusion Matrix

3. Results Bacheler Burt

Data cleaning procedures addressed missing values in continuous variables using mean imputation, and original categorical features were encoded numerically and subsequently removed to ensure compatibility and robustness in machine learning modeling. Modeling efforts centered on predicting blood pressure outcomes, with particular attention to systolic and diastolic blood pressure as markers of cardiovascular health. The results of this study are as follows. Two models, random forest and multiple regression, were evaluated for predicting blood pressure outcomes. Given the varying evaluation metrics used between classification and regression models, they will need to be evaluated within the appropriate context and using different metrics.

Over the course of this model's analysis, the models that performed with It should be noted that while a scientific basis for feature engineering exists (as reflected in this analysis' references), models built with a smaller, clinically grounded feature set are more efficacious in this analysis. As will be seen, each model has its own unique strengths and drawbacks. Together, they create a useful, composite image of how overall health relates to blood pressure.

Exploratory visualizations demonstrate variation within the sample, supporting the evaluation of model performance in this dataset. Together, these findings indicate that non-linear models provide stronger predictive performance for clinical outcomes in this dataset. The sequence of cleaning, feature engineering, model selection, and evaluation was performed in accordance with best practices in data science.

Figure 1 provides a detailed overview of the random forest classifier's performance for predicting blood pressure categories in the clinical dataset. The confusion matrix visualizes model accuracy by showing the number of true and false predictions for each blood pressure group—Normal, Elevated, Hypertension Stage 1, and Hypertension Stage 2—with correct classifications concentrated along the diagonal. The vast majority of cases are classified correctly, including 59 Normal, 83 HT Stage 1, and an exceptionally high 124 HT Stage 2 participants, reflecting the model's ability to differentiate higher-risk profiles with strong precision.

One reason why the model underperforms with respect to classification accuracy for the "Normal" and "Elevated" categories is that they are much smaller in terms of sample size than HT Stage 1 and HT Stage 2. Over 600 of the subjects in the sample belong to the HT Stage 2 category; just 100 subjects belong to the Elevated category. This discrepancy explains in part the model's varied performance and offers a justification for an expanded version of the study in future, as we note in the conclusion.

Table 2 displays model evaluation scores sorted by target variable (Blood Pressure Category). These scores are indeed reflective of the results of the confusion matrix. The scores for all metrics (Precision, Recall F1-Score, and Accuracy) are lower in the "elevated" class than anywhere else. The model appears to be the most efficacious in its predicting power for Stage 2 Hypertension (HT Stage 2), achieving an accuracy of 0.939.

The "Elevated" category is noted as more challenging for

Class	Precision	Recall	F1-score	Accuracy
Normal	0.922	0.952	0.937	0.952
Elevated	0.786	0.5	0.611	0.5
HT Stage 1	0.856	0.922	0.888	0.922
HT Stage 2	0.947	0.939	0.943	0.939

Table 2. Random Forest - Classification Metrics

the model, with the greatest confusion occurring between adjacent clinical categories (Elevated and HT Stage 1), a known difficulty in medical datasets where category boundaries can be subtle. Nevertheless, the model minimizes critical misclassifications, rarely mistaking severe hypertension for less serious states, which is clinically significant. For example, the model shows weaker performance for the 'Elevated' class, with an accuracy of 0.50, but its precision remains relatively high (0.78), meaning that most patients labeled as 'Elevated' truly belonged to that category. The lower accuracy simply means, overall, roughly half of the predictions were incorrect.

Feature	Importance
Diastolic BP	0.342
Systolic BP	0.248
CVD Risk Score	0.062
BMI	0.051
Fasting Blood Sugar (mg/dL)	0.049
Weight (kg)	0.049
Height (m)	0.047
Total Cholesterol (mg/dL)	0.047
Age	0.045
Physical_Activity_Level_Encoded	0.015
CVD_Risk_Level_Encoded	0.013
Diabetes_Status_Encoded	0.009
Smoking_Status_Encoded	0.009
Sex	0.008
Family_History_CVD_Encoded	0.008

Table 3. Random Forest - Feature Importance

Feature importance analysis (Table 3) reveals that both systolic and diastolic blood pressure are the strongest predictors, aligning with established risk factors from medical literature. Auxiliary features such as encoded lifestyle and demographic variables contribute meaningfully but to a lesser degree. In summary, these results demonstrate the random forest classifier's robust performance, especially for detecting higher-risk blood pressure categories, and confirm its suitability as a screening tool in resource-limited clinical settings. This multi-metric evaluation provides a comprehensive view of model behavior across classes.

3.2. Multiple Regression Mia Cachion

Figure 2 presents the regression model's predictive results for systolic blood pressure, visualized as a scatterplot com-

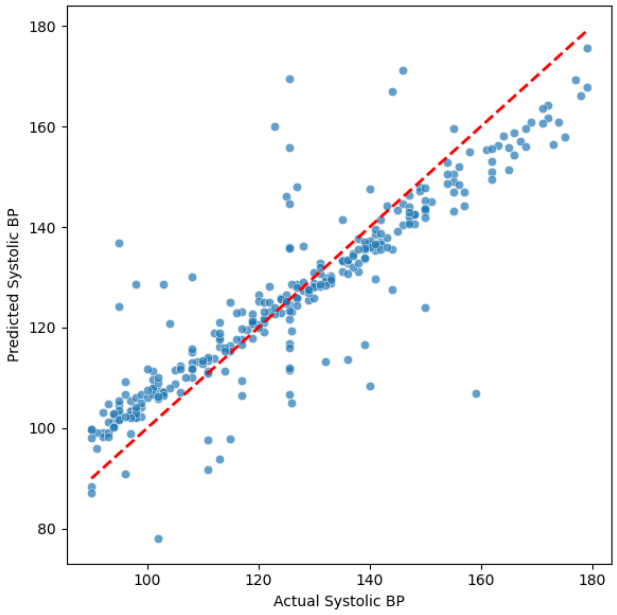

Figure 2. Multiple Regression - Systolic Blood Pressure

paring predicted values with actual measurements from the dataset. The majority of data points cluster tightly along the dashed 45-degree line, signaling a strong relationship between the model's predictions and true systolic readings. The calculated R^2=0.804 demonstrates that the model explains about 80 percent of the observed variance in systolic blood pressure, indicative of strong predictive performance. On average, predictions differ from true values by about 10 units, as reflected by the RMSE of 9.87.

Feature importance analysis further refines the interpretation. These values are seen in Table 4, sorted by highest to lowest coefficient value. Diabetes status emerges as the most influential variable, with an absolute regression coefficient of 27.072, followed by CVD risk score (coefficient 14.001). Other features yield much smaller coefficients, suggesting limited independent contribution to the explained variance. This finding is clinically relevant, reinforcing established medical knowledge: diabetes and overall cardiovascular risk are major drivers of elevated systolic pressure in the patient sample from Jamalpur.

Compared to the more flexible random forest model, multiple regression preserves interpretability and transparency, but may slightly underperform when modeling non-linear and interaction effects present in real-world health data. The plot also highlights some heteroscedasticity and outlier patterns, indicating that further research—such as residual analysis or inclusion of interaction terms—could refine future predictions. Overall, Figure 2 validates the effectiveness

Feature	Importance
Diabetes_Status_Encoded	27.072
CVD Risk Score	14.001
BMI	2.551
BP_Category_Label	2.185
CVD_Risk_Level_Encoded	1.345
Smoking_Status_Encoded	1.311
Height (m)	0.988
Family_History_CVD_Encoded	0.808
Total Cholesterol (mg/dL)	0.261
Sex	0.235
Weight (kg)	0.046
Age	0.012
Fasting Blood Sugar (mg/dL)	0.008
Physical_Activity_Level_Encoded	0.007

Table 4. Systolic Blood Pressure Feature Importance

of multiple regression for systolic blood pressure prediction, while the feature-importance results underscore the dominant role of diabetes and vascular risk in determining systolic values in this population.

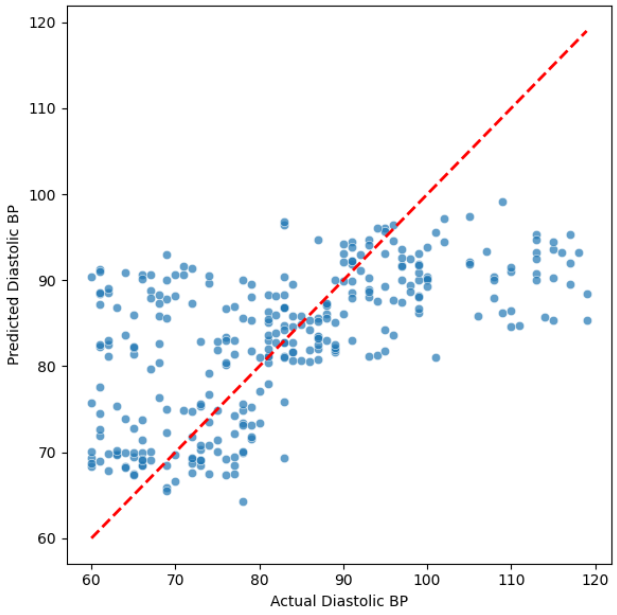

Figure 3. Multiple Regression - Diastolic Blood Pressure

Figure 3 displays the regression model's predictive performance for diastolic blood pressure, comparing predicted values with actual measurements from the dataset. While the plot shows a positive relationship between predicted and true diastolic values, the scatter is wider than in the systolic model, with more points deviating above and below the 45-degree reference line. This indicates that although the model captures general trends, its precision is noticeably lower. The performance metrics support this interpretation: the model yields an R^2 of 0.356, meaning it explains roughly 36 percent of the variance in diastolic blood pressure, and an RMSE of 12.19, indicating greater average prediction error than in the systolic model. Compared to the systolic model, the diastolic regression shows reduced accuracy and greater heterogeneity. This may reflect the known clinical complexity of diastolic blood pressure, which is often more sensitive to measurement noise, non-linear interactions, and individual variability. The wider scatter and presence of outliers suggest that model extensions such as including interaction terms, non-linear transformations, or alternative algorithms, could improve performance in future work. Overall, Figure 3 indicates that while multiple regression provides interpretable insight into diastolic blood pressure patterns, its predictive power is more limited than for systolic values.

Model	R^2	RMSE
Systolic BP	0.804	9.87
Diastolic BP	0.356	12.19

Table 5. Systolic vs. Diastolic Regression Model Evaluation Comparison

As evident in Table 5, the model is stronger in predicting systolic blood pressure than diastolic blood pressure. The model targeting systolic blood pressure has an R^2 value of 0.804, a score significantly higher than that of diastolic blood pressure (0.356). Additionally, the model also yields a lower RMSE score of 9.87 (as opposed to 12.19), meaning that, on average, there is less error in the model's predictions than those of diastolic blood pressure.

In summary, this analysis demonstrates that machine learning models, most notably the random forest classifier and multiple regression, are highly effective at predicting blood pressure outcomes and classifying risk among patients in the Jamalpur dataset. However, several limitations must be acknowledged. The cross-sectional nature of the data precludes causal inferences, and selection bias may arise from the geographic and temporal focus on a single region and year. Data quality, such as missing values and imputation choices, could introduce modeling artifacts, and some relevant risk factors may be unmeasured or insufficiently captured. Despite these potential flaws, the modeling strategy is methodologically sound, employing rigorous data preprocessing, robust train-test splits, and comprehensive evaluation metrics to ensure generalizability. The strong alignment between predictive results and clinical assessments supports the practical merit of these models for both risk stratification and public health planning, though con-

tinued validation on larger and more diverse datasets will further enhance their reliability.

4. Conclusion Lily Rademacher

This analysis examined the relationships between key biometric, lifestyle, and clinical variables and blood pressure outcomes using a dataset of 1,529 patient samples from Jamalpur Medical College Hospital in Bangladesh. By focusing on predictors such as Diabetes Status, Smoking Status, Physical Activity Level, Sex, Age, Weight, Height, Body Mass Index (BMI), Total Cholesterol, Fasting Blood Sugar, and both Systolic and Diastolic Blood Pressure (measured in mmHg) the study employed complementary modeling approaches: random forest classification for categorical blood pressure categories and linear regression for continuous systolic blood pressure predictions. Data preparation involved rigorous cleaning—handling missing values via mean imputation, standardizing continuous features, and one-hot encoding categorical variables—to ensure compatibility with machine learning pipelines implemented in Python using scikit-learn, pandas, and visualization libraries like seaborn and matplotlib. Model training followed an 80/20 train-test split, with performance evaluated through domain-appropriate metrics: accuracy, precision, recall, and F1-score for classification, alongside R^2 and root mean squared error (RMSE) for regression, enabling direct cross-model comparisons.

The random forest classifier demonstrated robust performance, achieving an overall accuracy of 0.905, with high precision and recall particularly for higher-risk categories like Hypertension Stage 2 (124 correct predictions out of relevant cases), and minimal clinically-adverse misclassifications between severe and milder states. The test set R^2 of 0.893 and RMSE of 0.37 confirmed strong predictive alignment, outperforming linear regression's R^2 of 0.799 and RMSE of 10.00 for systolic blood pressure. Despite the dataset's modest size, several methodological choices mitigated risks of overfitting and unstable estimates. Simpler model configurations, such as shallow trees, constrained complexity relative to sample availability. Dimensionality reduction through targeted feature selection (dropping redundant variables like height and weight in favor of BMI) maintained a favorable predictor-to-sample ratio, enhancing stability without information loss. Multi-metric evaluation, including both classification-specific (F1-scores) and regression-standard R^2 measures normalized for comparability, provided comprehensive benchmarking that transcends task boundaries, reinforcing the findings' reliability. Exploratory visualizations of the heterogeneous patient cohort further supported generalizability, as diverse demographic and clinical distributions reduced variance inflation common in small, non-representative samples.

While this project successfully demonstrated predictive utility for blood pressure risk, several avenues for extension lie beyond its scope. Firstly, replicating this study on a larger sample size would have huge positive implications for the scale and depth of possible analysis. A larger dataset would strengthen generalization and allow for complex machine learning processes, like neural networks or tabular foundation models. Future studies could also separate a larger body of samples into subgroups, allowing for specific analysis based on age, gender, diabetes status, or other characteristics that create heterogenous results not captured in our study. The drivers of hypertension risk may behave differently across different demographics—for example, perhaps age has a stronger effect on hypertension in men than women; alternatively, the impact of weight could differ depending on diabetes status. These are questions that certainly merit investigation in future studies. With expanded data, models could also detect rare events (e.g., extreme hypertension) reliably, calibrating for imbalanced classes via stratified sampling. This expansion of the study would be critical to high-risk cases and has the potential to drive significant public health benefits. It would also be beneficial to expand the subject to non-Bangladeshi people, whether via collating data from numerous hospitals across the world or using data from a hospital in another country or region. It is important to verify whether these results are specific to Bangladesh or the region of Jamalpur, a textile-industry region in the north of the country. Another study might examine subjects from the capital city, Dhaka, instead, or neighboring India or Myanmar. Additionally, future work could incorporate longitudinal data to model progression dynamics, or integrate external cohorts for external validation and transfer learning. Advanced techniques like synthetic oversampling or generative augmentation could enrich the dataset, though careful fidelity checks would be essential to preserve clinical realism.

These enhancements, while valuable, exceed the exploratory aims of this academic analysis, which prioritizes foundational modeling and interpretability in a constrained data environment. This study advances the understanding of hypertension risk factors in an underserved Bangladeshi cohort, showcasing machine learning's potential for accessible clinical screening even with limited data. The impressive model performance signals vast prospects for further investigation, where scaled datasets and advanced tools could transform predictive hypertension models into deployable public health interventions worldwide.

Impact Statement All Group Members

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, especially because it involves data pertaining to health. Health information is complex and personal, so any predictions must be interpreted cautiously. Overgeneralizing or making strong claims from limited data can affect people's lives in harmful ways. For this reason, it is important to acknowledge the limits of the model and avoid drawing conclusions that go beyond what the data can support.

References

Couch, C. A., Mascarenhas, R., Stierman, B., Fryar, C. D. (2025). Prevalence of cardiovascular disease risk factors in adults: United States, August 2021–August 2023. NCHS Data Brief, (540), 1–9. https://doi.org/10.15620/cdc/174623

Nirob, M. A. S., Bishshash, P., SIAM, A. K. M. F. K., Haque, M. A., Assaduzzaman, M. (2025). CAIR-CVD-2025: An extensive cardiovascular disease risk assessment dataset from Bangladesh (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/d9scg7j8fp.1

Qiao, W., Zhang, X., Kan, B., Vuong, A. M., Xue, S., Zhang, Y., Li, B., Zhao, Q., Guo, D., Shen, X., Yang, S. (2021). Hypertension, BMI, and cardiovascular and cerebrovascular diseases. Open Medicine, 16(1), 149–155. https://doi.org/10.1515/med-2021-0014

Singh, A., Vishal, Kumar, A., Sonam, Kumar, N., Sharma, N. (2024). A basic review on hypertension, its nonpharmacological and pharmacological treatment. Journal of Drug Delivery and Therapeutics, 14(3), 202–209.

Wang, C., Ye, D., Xie, Z., Huang, X., Wang, Z., Shangguan, H., Zhu, W., Wang, S. (2021). Assessment of Cardiovascular Risk Factors and Their Interactions in the Risk of Coronary Heart Disease in Patients with Type 2 Diabetes with Different Weight Levels, 2013-2018. Diabetes, metabolic syndrome and obesity : targets and therapy, 14, 4253–4262. https://doi.org/10.2147/DMSO.S335017

Zhu, N., Suarez-Lopez, J. R., Sidney, S., Sternfeld, B., Schreiner, P. J., Carnethon, M. R., Lewis, C. E., Crow, R. S., Bouchard, C., Haskell, W. L., Jacobs, D. R., Jr. (2010). Longitudinal examination of age-predicted symptom-limited exercise maximum HR. Medicine Science in Sports Exercise, 42(8), 1519–1527. https://doi.org/10.1249/MSS.0b013e3181cf8242

Appendix Tables/Figures: Bacheler Burt

Feature	Importance
BP_Category_Label	8.293
Diabetes_Status_Encoded	5.5
CVD Risk Score	2.5
Height (m)	2.391
Sex	1.055
CVD_Risk_Level_Encoded	0.927
Family_History_CVD_Encoded	0.676
BMI	0.61
Physical_Activity_Level_Encoded	0.563
Smoking_Status_Encoded	0.316
Total Cholesterol (mg/dL)	0.054
Weight (kg)	0.036
Fasting Blood Sugar (mg/dL)	0.027
Age	0.003

Table 6. Diastolic Blood Pressure Feature Importance

EDA

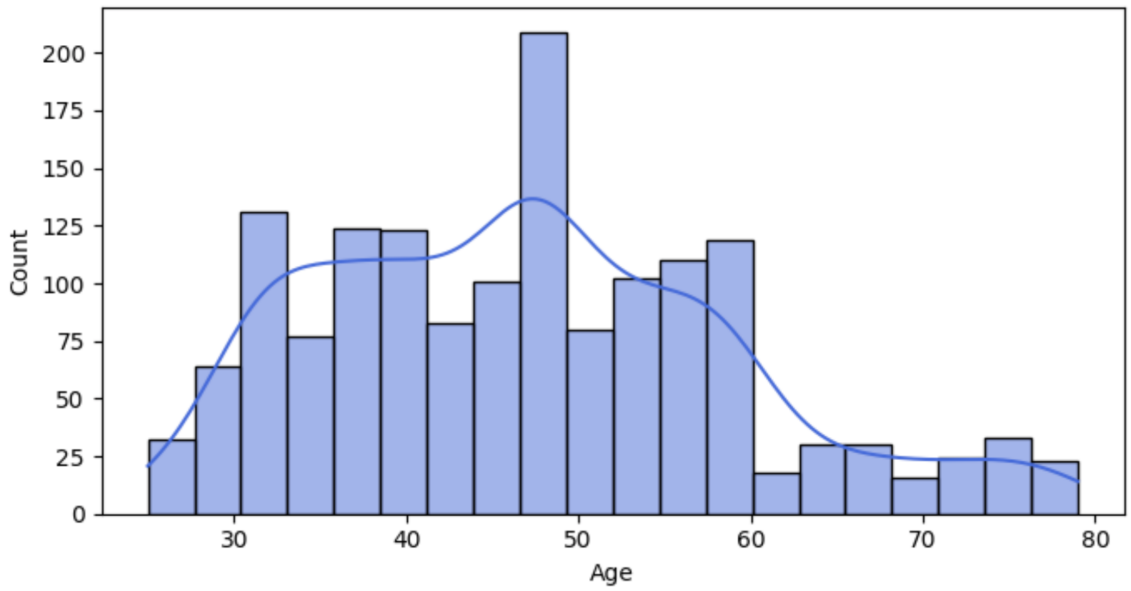

Figure 4. Age Distribution

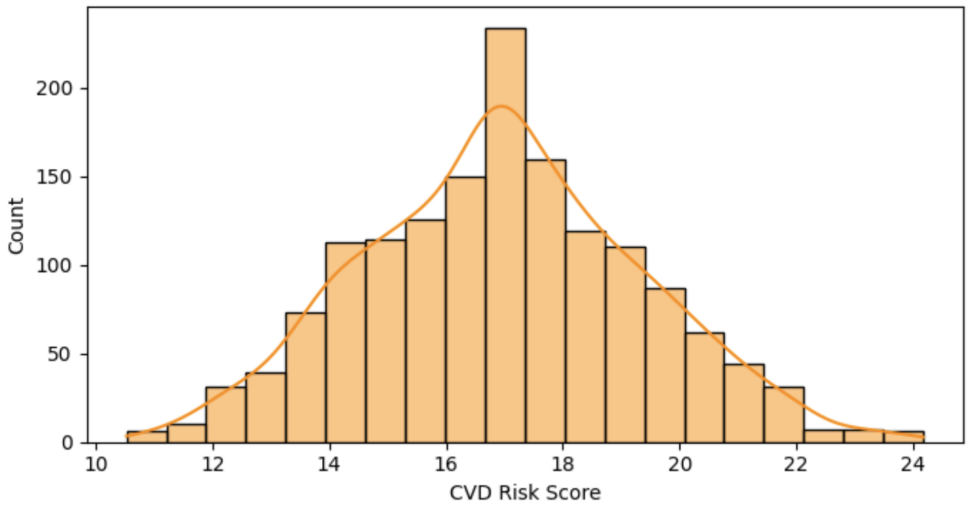

Figure 5. CVD Risk Score Distribution

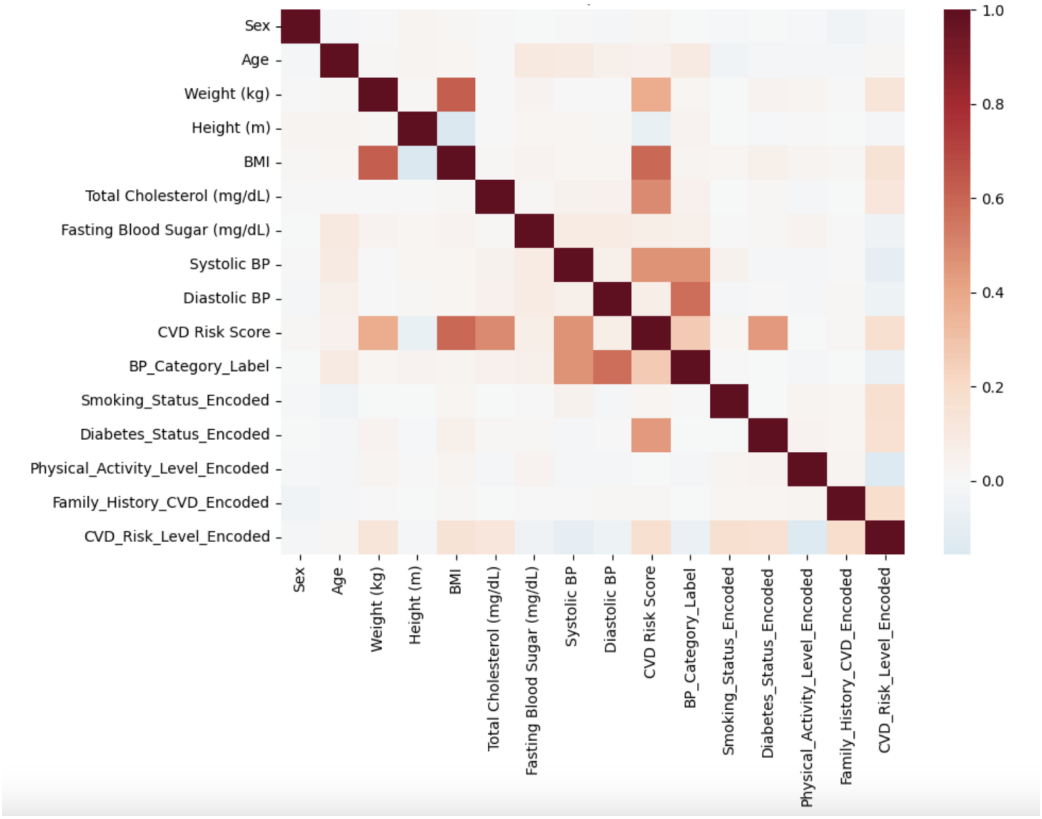

Figure 6. Correlation Heatmap

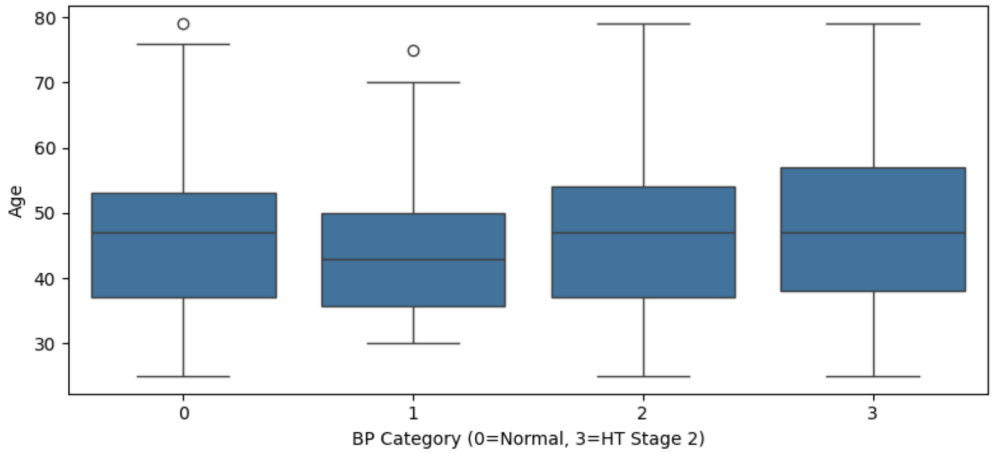

Figure 7. Age by Blood Pressure Category

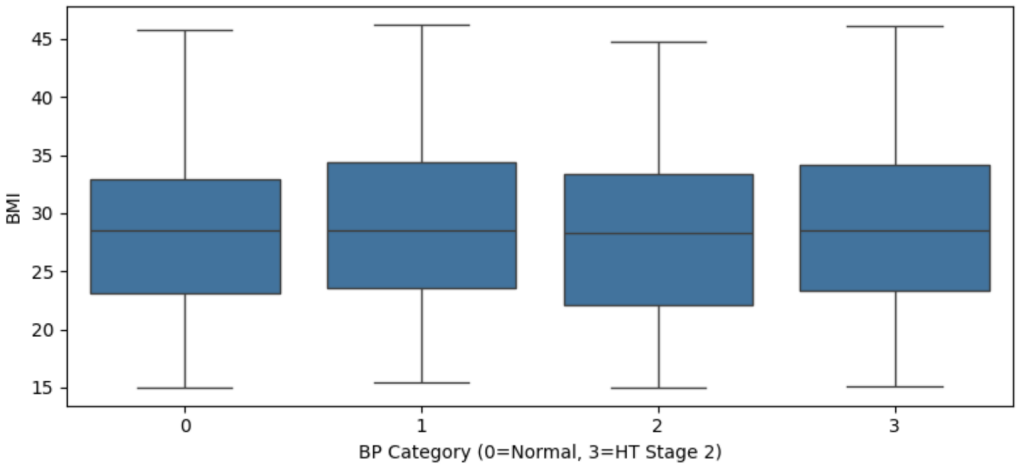

Figure 8. BMI by Blood Pressure Category

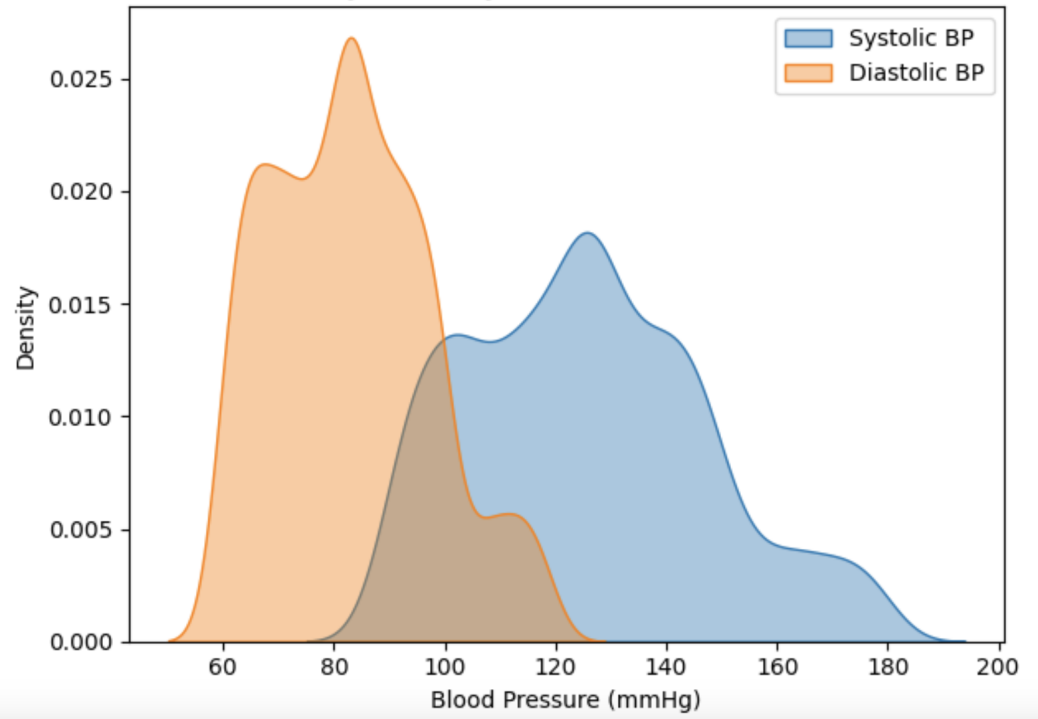

Figure 9. Kernel Density Plot of Systolic and Diastolic Blood Pressure

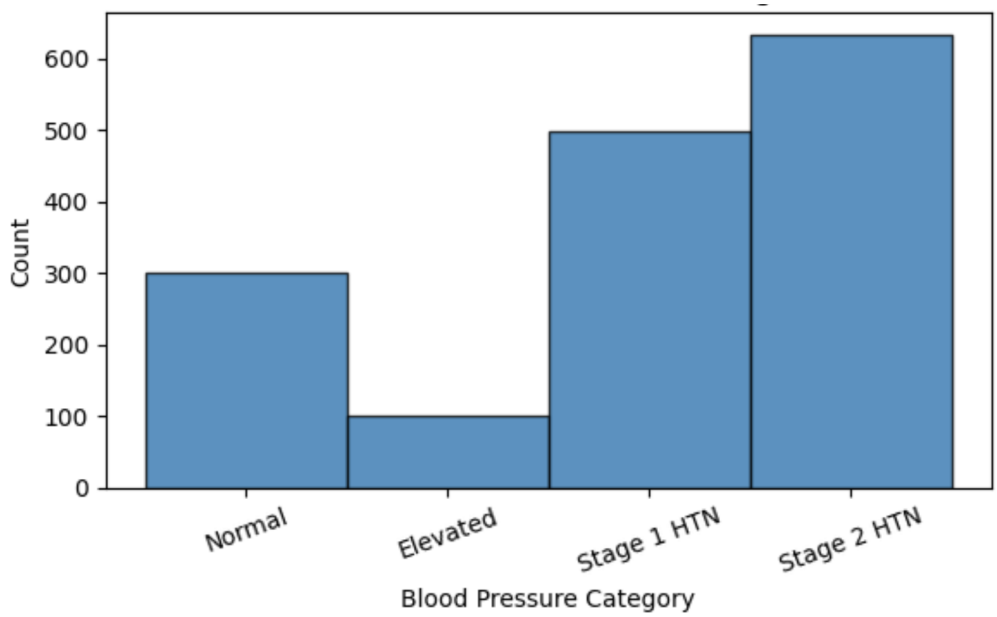

Figure 10. Distribution of Blood Pressure Categories

Overleaf Commits: Mia Cachion