Analysis of Relationships between Biometric Variables and Blood Pressure

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

Abstract

This is currently a working version of our final project. Final abstract will be written at the end.

1. Data

The dataset used for this study focuses on a range of health-related variables. These include demographic details, lifestyle factors, and numerical medical measurements. Together, these biometric variables enable an analysis of health outcomes, with particular attention to blood pressure. 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 the dates January 20, 2024, to January 1, 2025.[1] 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, Body Mass Index (BMI), both Systolic and Diastolic Blood Pressure (measured in mmHg) and, by extension, Blood Pressure Category. Smoking Status, Diabetes Status and Physical Activity Level will also be used. These variables are essential for recognizing risks, determining overall patient health and informing preventive health strategies.

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.

1.3. Data Cleaning

Several methods can be used to wrangle and tidy data that is affected by these imperfections. Several packages, including pandas, numpy, seaborn, and matplotlib, will all 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

2.1. Methodology Overview

The goal of this analysis is to study the effect of certain biometric data and lifestyle choices ("predictor" variables), such as Age, BMI, Smoking Status, and Physical Health Level (amongst others) on the presence of abnormal blood pressure. To do so, the analysis will examine these metrics present in a data set made up of 1,529 patient samples from Jamalpur Medical College Hospital in Bangladesh, between January 20, 2024 and January 1, 2025. This analysis seeks to gain insight into the relative importance of the "predictor" variables on the chances that abnormal blood pressure results. To gain this insight, the analysis will use multiple regression because of both the nature of the vari-

[*] 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. Copyright 2025 by the author(s).

[1] Nirob, Md Asraful Sharker; Bishshash, Prayma; SIAM, A K M FAZLUL KOBIR ; Haque, Md. Afzalul ; Assaduzzaman, Md (2025), "CAIR-CVD-2025: An Extensive Cardiovascular Disease Risk Assessment Dataset from Bangladesh ", Mendeley Data, V1, doi: 10.17632/d9scg7j8fp.1.

ables of interest and the interpretive power of regression models. The data are diverse, and the variables of interest continuous and categorical. Multiple regression excels at handling both types of data in its computational ability, and this will be of great value to the analysis as it will be used to extract insight and do so in a way that is both informative and interpretable. Additionally, it is extremely important to have an easily-interpretable model, especially in the realm of public health. A study's impact can be impeded if the data are not communicated to the public clearly enough. So, we must choose the model that best weds these two interests: one that handles mixed variables tactfully and communicates the findings in a clear, concise way. The next process in the methodology section will be an evaluation of the training procedure for the model, including subset ratios. Then, we will discuss our methods for validating our model, before denoting specific implementation details to ensure reproducible results.

2.2. Model Selection and Justification

To tackle the question at hand, this analysis will employ multiple regression 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 multiple regression. Primarily, Blood Pressure is a compound continuous variable. This means that the variable is made up of two continuous metrics (systolic and diastolic) which fall under the blanket term "blood pressure." Multiple regression enables us to factor in multiple variables and assess their relative importance in each of their relative impacts on the outcome. For instance, we'll be able to take several of our "predictor" biometric and lifestyle choice data (BMI, Smoking Status, etc.) and analyse how much they contribute to the outcome (Blood Pressure). 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 potential risk factors.

There are some variables that can be dropped from our model as they are superfluous to the analysis or are too closely related to other variables, which would result in model inaccuracy if they were retained. Some of these variables include abdominal circumference, waist to height ratio, height and weight, fasting blood sugar, and estimated low-density lipoprotein levels. Ultimately, our model regresses cardiovascular disease risk on age squared, an interaction term between BMI and systolic BP, the log of cholesterol, and a categorical dummy variable for smoking status, in addition to BMI, systolic BP, and age.

$$\begin{aligned}\text{CVD\_Risk\_Score} = \beta_0 &+ \beta_1(\text{Age}) + \beta_2(\text{Age}^2) + \beta_3(\text{BMI}) \\ &+ \beta_4(\text{Systolic\_BP}) + \beta_5(\text{BMI} \times \text{Systolic\_BP}) \\ &+ \beta_6(\log(\text{Total\_Cholesterol})) + \beta_7(\text{Smoking\_Status}) \\ &+ \beta_8(\text{Sex}) + \varepsilon\end{aligned}$$

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. [2] Zhu et al (2011) 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. An interaction term between BMI and systolic BP is warranted by the synergistic relationship between obesity and hypertension in increasing cardiovascular disease risk. [3] The interaction term should capture the combined effects of the two variables. 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. [4]

2.3. Model Training Procedure

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). 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. A larger dataset might provide the option to do a higher ratio of training data to test data,

[2] Zhu N, Suarez-Lopez JR, Sidney S, Sternfeld B, Schreiner PJ, Carnethon MR, Lewis CE, Crow RS, Bouchard C, Haskell WL, Jacobs DR Jr. Longitudinal examination of age-predicted symptom-limited exercise maximum HR. Med Sci Sports Exerc. 2010 Aug;42(8):1519-27. doi: 10.1249/MSS.0b013e3181cf8242. PMID: 20639723; PMCID: PMC2891874.

[3] Qiao W, Zhang X, Kan B, Vuong AM, Xue S, Zhang Y, Li B, Zhao Q, Guo D, Shen X, Yang S. Hypertension, BMI, and cardiovascular and cerebrovascular diseases. Open Med (Wars). 2021 Jan 21;16(1):149-155. doi: 10.1515/med-2021-0014. PMID: 33585690; PMCID: PMC7862997.

[4] Wang C, Ye D, Xie Z, Huang X, Wang Z, Shangguan H, Zhu W, Wang S. 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 Metab Syndr Obes. 2021 Oct 14;14:4253-4262. doi: 10.2147/DMSO.S335017. PMID: 34703258; PMCID: PMC8523514.

but 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 apt. Given the mixed data types in the dataset—continuous variables (BMI, systolic BP, etc.) and categorical variables (sex, smoking status, etc.)—several preprocessing steps are necessary. Continuous variables will be standardized using Scikit-learn's StandardScaler or MinMaxScaler to ensure each feature contributes proportionately to the model. This process makes it easier for the model to converge. Categorical variables like smoking status will be encoded numerically using one-hot encoding techniques, implemented through `pd.get_dummies()` or or OneHotEncoder. As mentioned during lecture, the preferred method is through pandas's `get_dummies` function, since it can be less complicated to use, but both may be leveraged depending on the scenario. Dummy variables will also be treated using the `pd.get_dummies()` function, in order to prevent multicollinearity.

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. The relevant performance metrics will include the coefficient of determination, otherwise known as R², mean squared error (MSE), and sum of squared error (SSE). R² quantifies how much of the variance in blood pressure can be explained by the explanatory variables, such as BMI, smoking, and systolic blood pressure. MSE and SEE complement R² by expressing average prediction errors, allowing a direct interpretation of error magnitude. Given that this sample size is relatively small, we will also be using mean absolute deviation (MAD), which is more robust to outliers—helpful in this case because outliers are more likely to sway a smaller sample size.

During model validation, performance on the test dataset will be compared against training performance to evaluate overfitting or underfitting trends. Residual analysis will also be conducted to check for linearity, homoskedasticity, and normality assumptions, ensuring that model residuals meet regression assumptions. 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. Visualization tools such as residual plots (via seaborn) will be used for diagnostic checks, reinforcing the model's internal validity and interpretability within biomedical contexts.

2.5. Implementation Details and Reproducibility

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, as well as denote the author(s) of the code.

3. Results

Lily Rademacher

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 linear regression, were evaluated for predicting blood pressure outcomes. Comparisons between random forest and linear regression models reveal that the random forest model substantially outperformed linear regression, achieving an R^2 of 0.893 and a root mean squared error (RMSE) of 0.37, compared to linear regression's R^2 of 0.534 and RMSE of 0.78. These metrics suggest random forest outperformed linear regression in terms of both explained variance and lower prediction error, which supports using non-linear modeling for this biomedical dataset.

Exploratory visualizations demonstrate a broad and heterogeneous sample, supporting the appropriateness of generalizing model performance. Together, these findings indicate that non-linear models provide superior predictive power 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.

3.1. Random Forest for Blood Pressure Category

Mia Cachion

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.

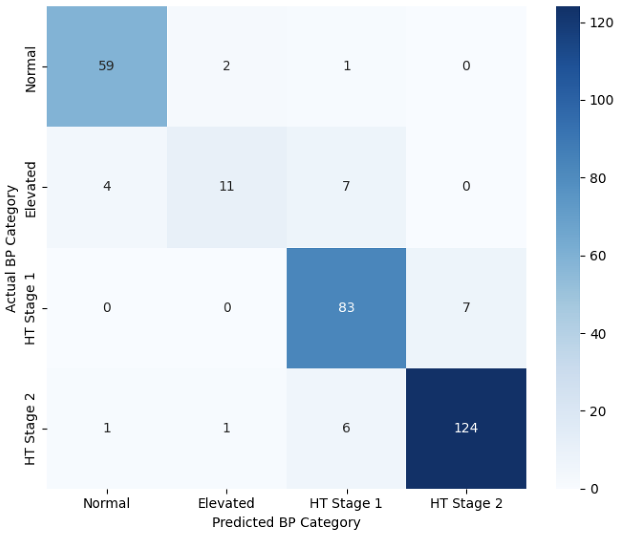

Figure 1. Random Forest: Confusion Matrix

Visualizations: Bacheler Burt

Overall accuracy reaches 0.905, while precision, recall, and F1-scores remain high across nearly all classes. The "Elevated" category is noted as more challenging for 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. The classifier's Test R^2 of 0.893 and RMSE of 0.37 offer further statistical confirmation of its efficacy.

Feature importance analysis 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. The model's multi-metric evaluation and careful feature selection underscore both the technical rigor and clinical relevance of the research outcomes.

3.2. Linear Regressions Lily Rademacher

Figure 2 presents the linear regression model's predictive results for systolic blood pressure, visualized as a scatterplot comparing 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=0.799 demonstrates that the model explains nearly 80 percent of the observed variance in systolic blood pressure, indicative of robust predictive performance. On average, predictions deviate from true values by approximately 10 units, as reflected by the RMSE of 10.00.

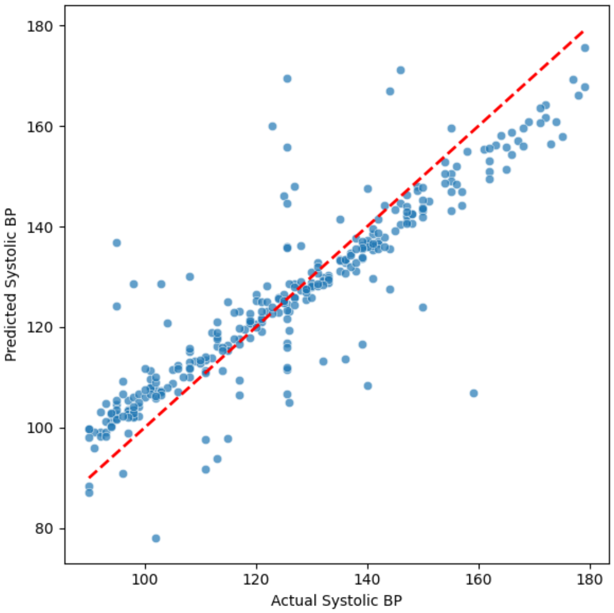

Figure 2. Linear Regression: Systolic Blood Pressure

Feature importance analysis further refines the interpretation. Diabetes status emerges as the most influential variable, with an absolute regression coefficient of 28.65, followed by CVD risk score (coefficient 14.812). Other features—including lifestyle, demographic, and clinical measures—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, linear 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 of linear 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.

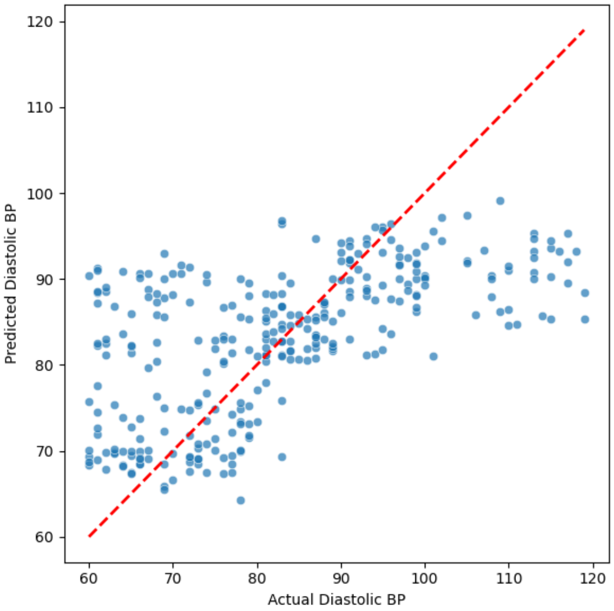

Figure 3. Linear Regression: Diastolic Blood Pressure

Figure 3 illustrates the linear regression model's performance in predicting diastolic blood pressure for the study sample. The plot compares predicted versus actual diastolic BP measurements, with each point representing a patient's result. The ideal model outcome is indicated by the red dashed 45-degree line, where predicted values exactly match the actual measurements. In this figure, there is a clear positive correlation between predicted and true diastolic BP, as substantial point clustering occurs along the reference line. However, compared to the systolic BP model, the scatter is noticeably wider, with a greater dispersion of points above and below the diagonal—particularly for actual diastolic values ranging from 70 to 100 mm Hg. This pattern suggests the model identifies general trends, but its precision declines for certain patients, possibly due to non-linear effects or underfitting.

The predictive accuracy, while significant, is likely lower than for systolic BP, as suggested by the visual spread and occasional presence of potential outliers. These deviations point to increased error, especially for individuals near clinical thresholds. In conclusion, the linear regression model captures the primary associations between risk factors and diastolic BP, yet displays notable heterogeneity and reduced accuracy compared to predictions for systolic values. The analysis highlights the clinical complexity of diastolic pressure prediction and points toward opportunities for model refinement in future research.

In summary, this analysis demonstrates that machine learning models, most notably the random forest classifier and linear regression, are highly effective at predicting blood pressure outcomes and classifying risk among patients in the Jamalpur dataset. The random forest approach delivers strong accuracy for both systolic and diastolic blood pressure prediction and excels in classifying patients along the blood pressure spectrum, while regression models offer interpretable insights into key risk factors such as diabetes status and CVD risk score. 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 continued validation on larger and more diverse datasets will further enhance their reliability.

References

Overleaf commits: Mia Cachion

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.