Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts

TEAM NUMBER 21

Gurwinder Kaur: 50363756

Dishank Jagadeeshnaidu Karampudi: 50559618

Rishi Shanthan: 50560689

Manasa Lakhsmi Gunampalli: 50559593

Introduction

- Our project integrates multiple health indicators and lifestyle factors into a single comprehensive health index.
- This index classifies individuals into three risk categories : high, medium, low
- Our project employes K-means clustering.
- This project is significant as it serves the needs of various stakeholders, including healthcare providers, insurers, policymakers, and individuals.
- Our Dataset: EFFECTS OF SMOKING AND DRINKING ON HEALTH

Why K-Means for health risk prediction?

- k-Means clustering is chosen for its ability to handle complex patterns in health data, which often exhibit non-linear relationships. This makes it more suitable for predicting health risks in our project.
- k-Means does not assume linear relationships. k-Means adapts flexibly to the data structure, making it robust.
- k-Means provides clear and distinct clusters for multi-classification tasks
- The decision-making process in k-Means is straightforward and based on the proximity of data points to cluster centroids.

Data collection and Preprocessing

- Data collection: Comprehensive dataset including health metrics such as blood pressure, cholesterol levels, BMI, dietary habits, and physical activity levels.
- Preprocessing steps:
  - Missing values: missing data handled through imputation to maintain dataset integrity
  - Normalization: Standardization of all features to ensure a uniform scale across the dataset.
  - Outlier Removal: Enhancing data quality by identifying and removing outliers.

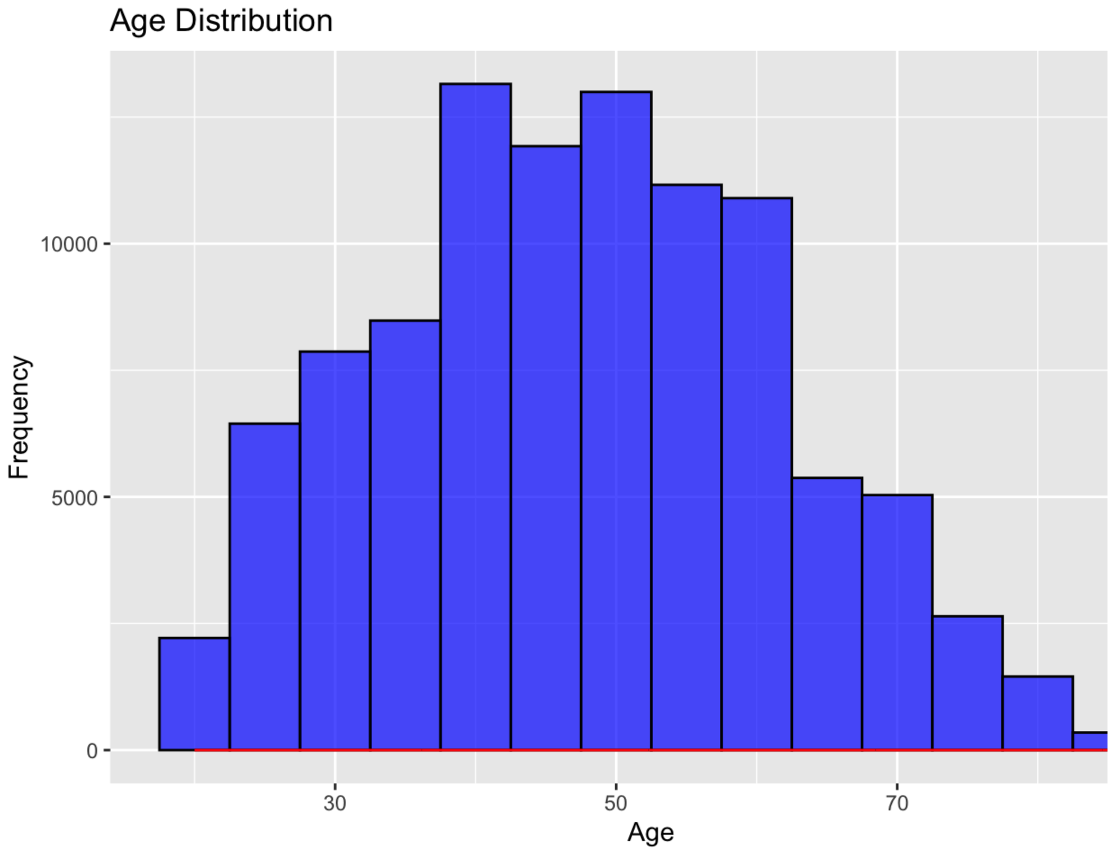
Feature Engineering and dimensionality reduction

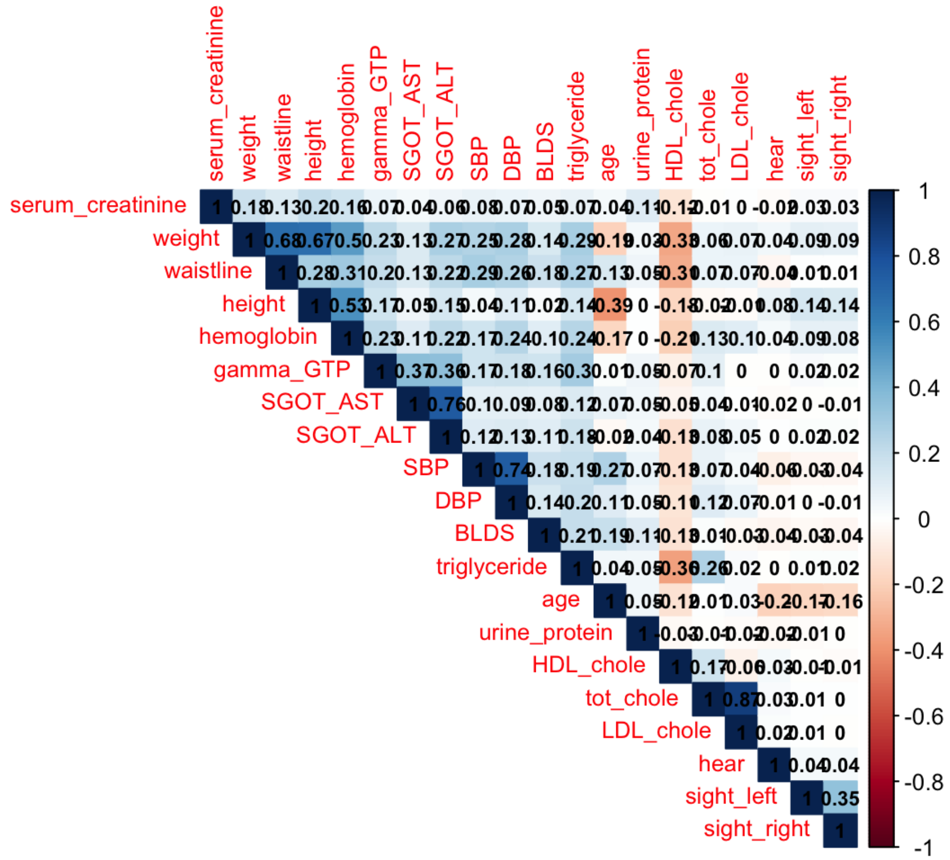
- Feature Engineering Techniques:
  - Label Encoding: Ordinal categorical variables are transformed into numerical codes
  - One-Hot-Encoding: Non-ordinal categorical variables are converted into binary columns.
  - Interaction Terms: created to capture synergistic effects

- Dimensionality Reduction:
  - Principal Component Analysis (PCA) : reduces the dimensionality while retaining significant variance, optimizing the K-Means model's efficiency.

Model validation, Optimization, and Interpretability

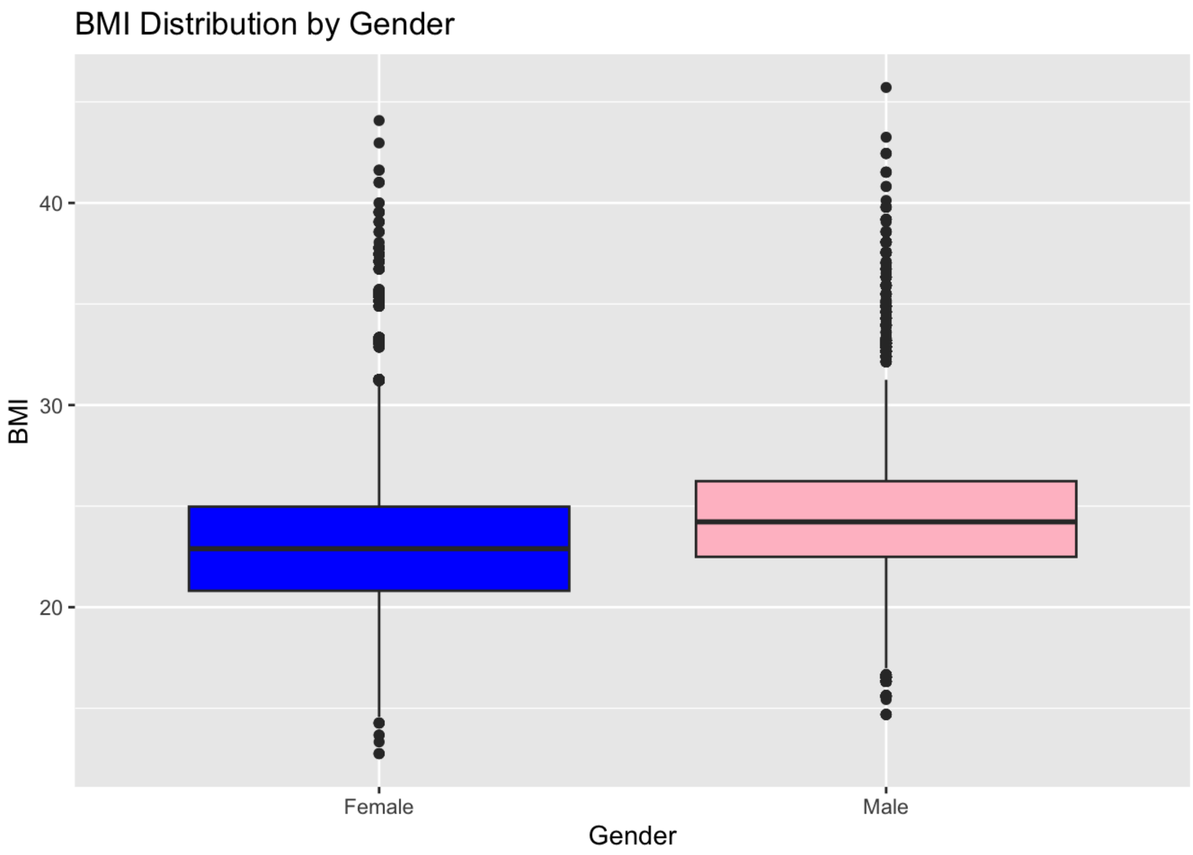
- Cross-validation: utilized to prevent overfitting and ensure that the model generalizes effectively to unseen data.
- Metrics Used: accuracy, precision, confusion matrix
- Hyperparameter Tuning: Both grid search and random search



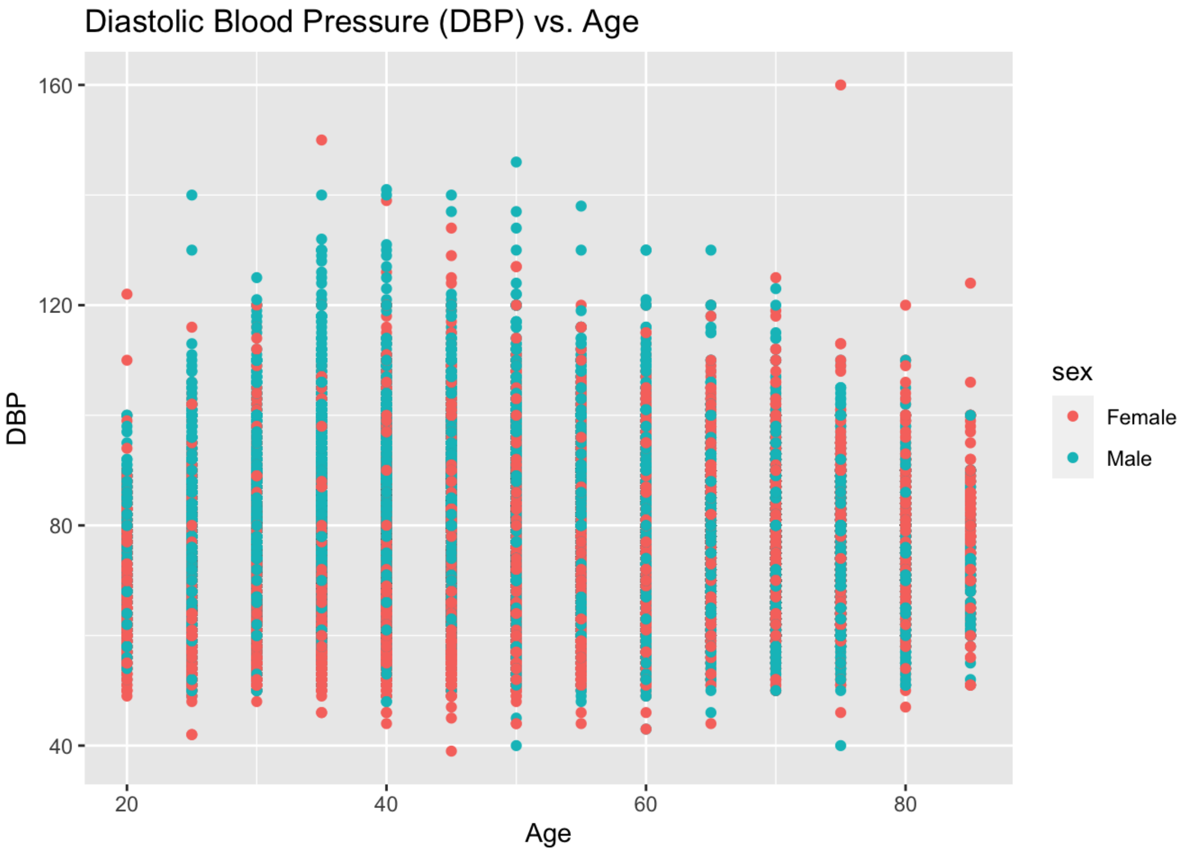


The population is predominantly middle-aged, with the largest group around 45-55 years, indicating a dataset centered around working age-adults

Highlight the overall health metrics such as average BMI, typical blood pressure ranges, and cholesterol levels.

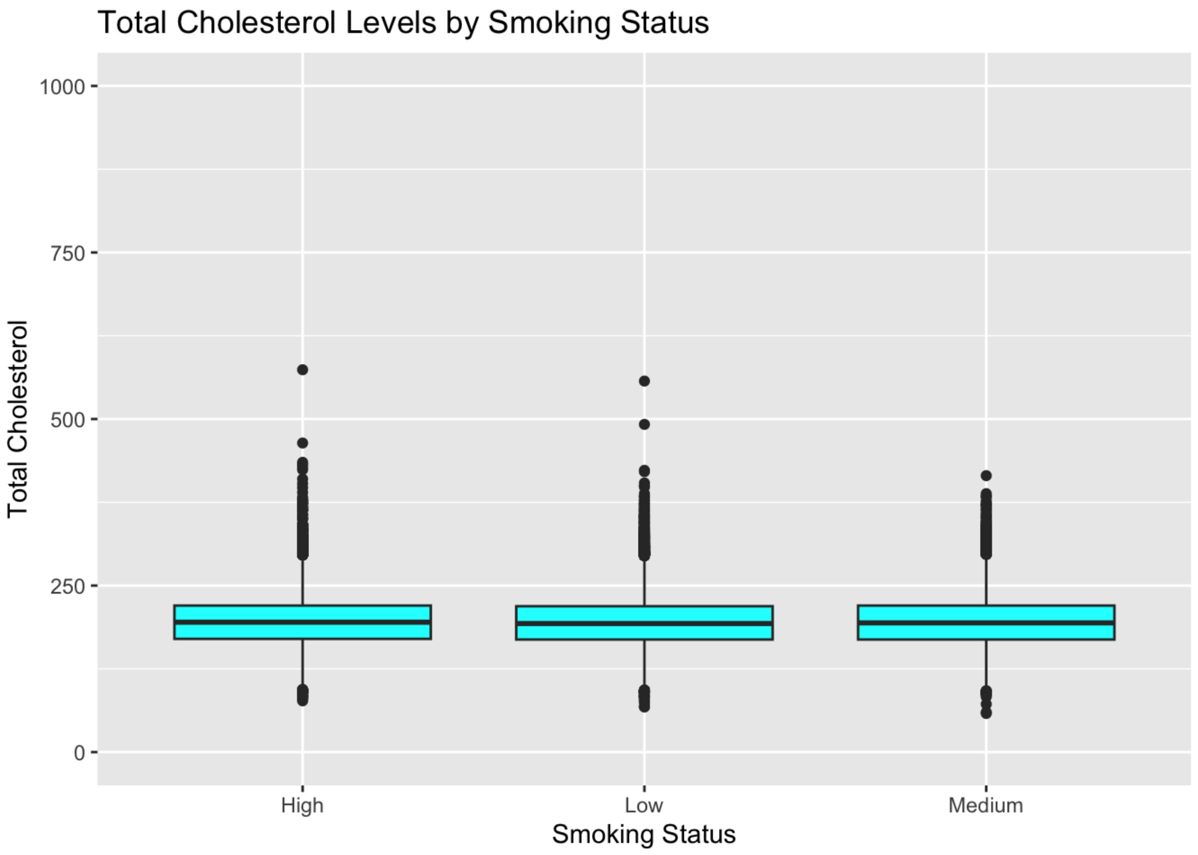


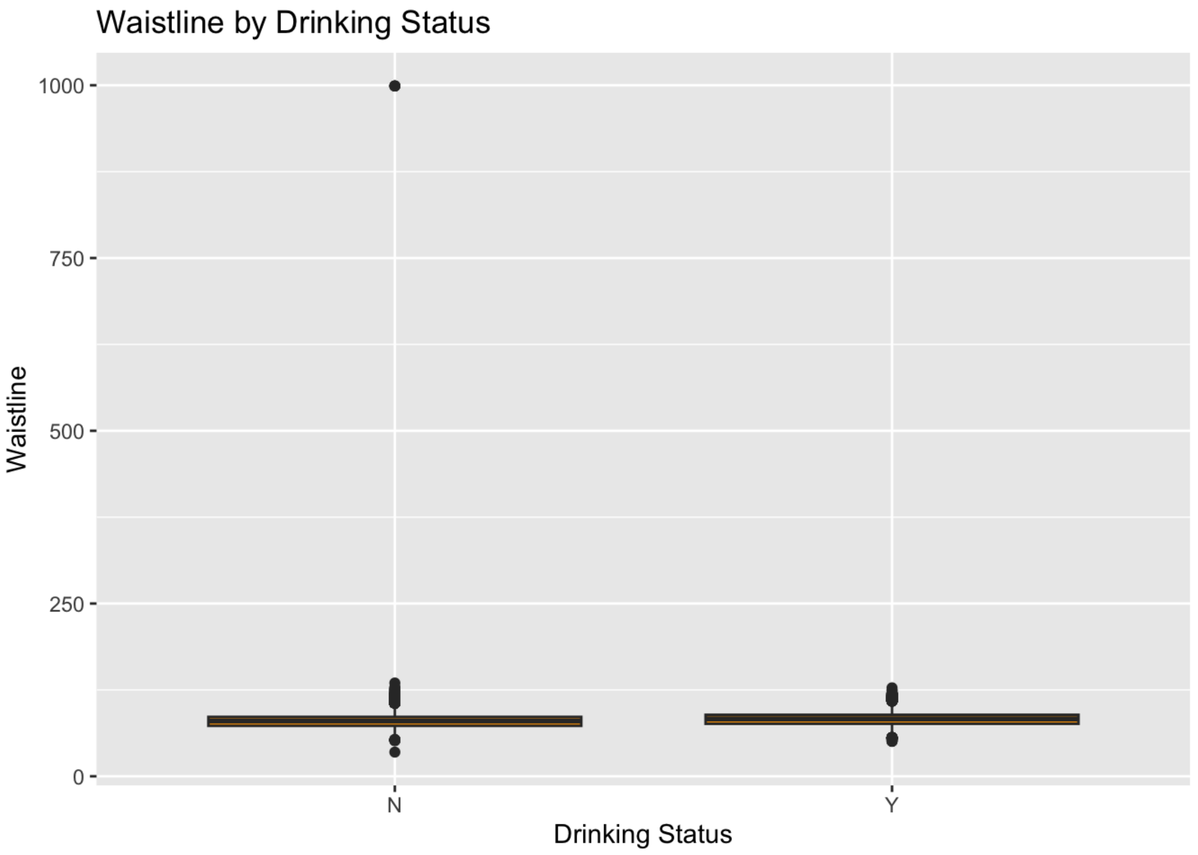
Both gender show a median BMI around 25, but males exhibit a slightly wider range. This suggest that there is a variability in health risk factors between genders.



SBP usually increases with age, with males showing a slightly higher blood pressure. This potentially results in greater cardiovascular risk.

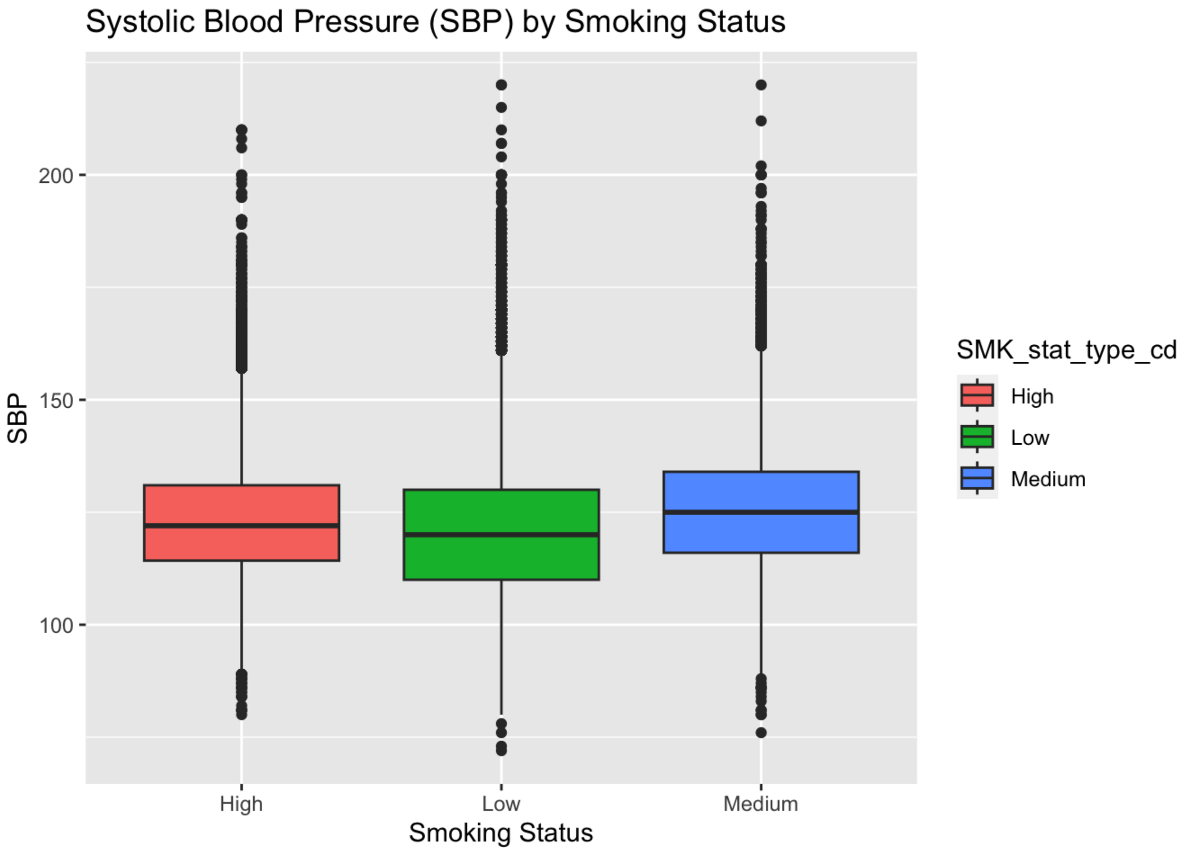
DBP is similar to SBP. As people get older, their diastolic blood pressure generally tends to increase, and the range of blood pressure readings becomes broader.

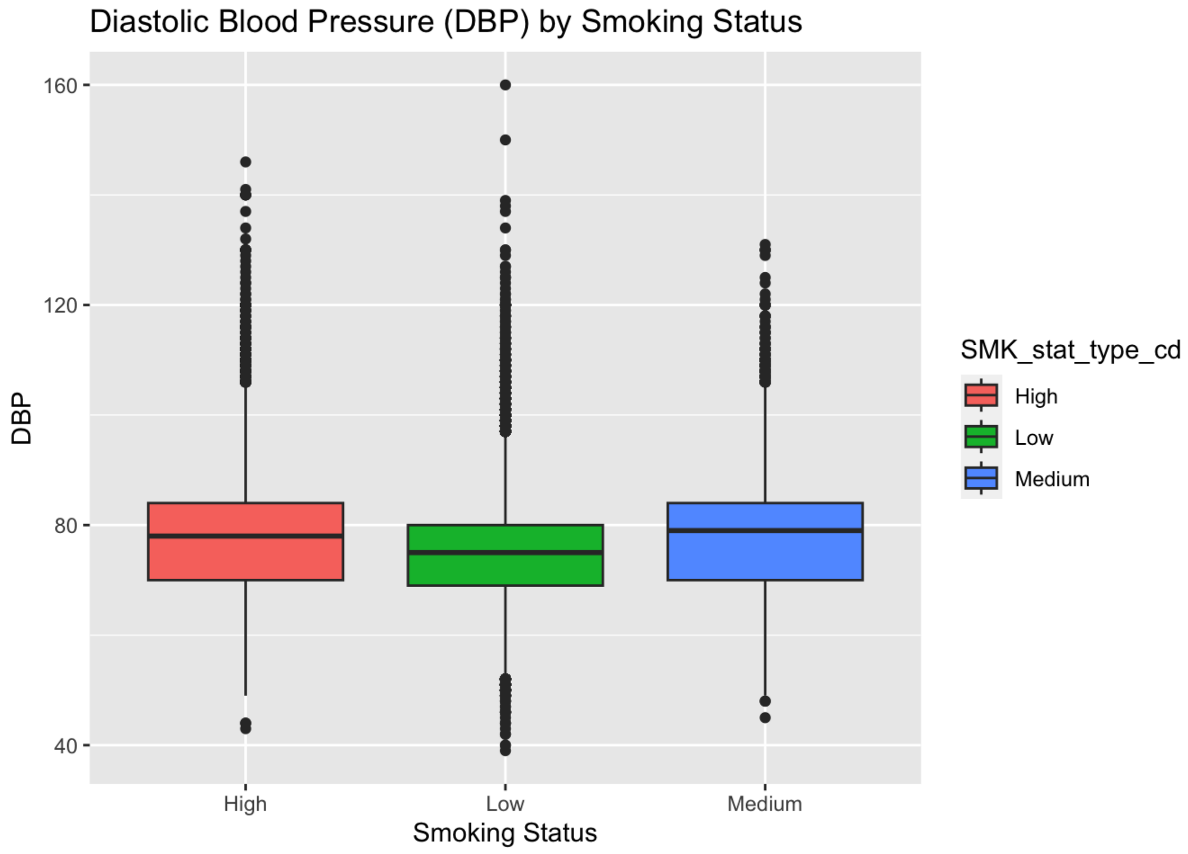




Cholesterol levels do not show significant differences across smoking status .

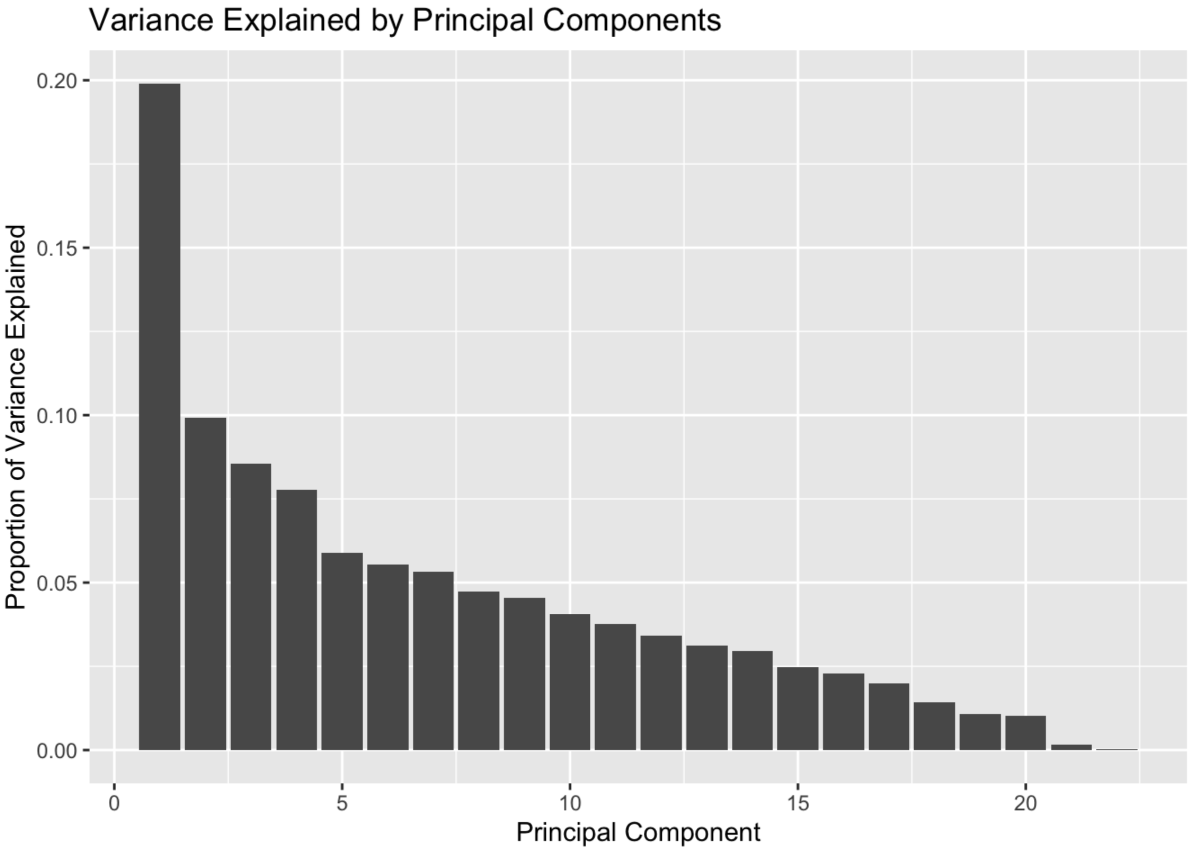
Individuals who drink have a slightly higher average waistline.





There is a higher SBP in individual with high smoking status indicating a direct health risk from heavy smoking.

Have similar pattern as the SBP. It follows the same pattern.



The first key components capture the majority of the data variance.

This clustering demonstrate how different health metrics cluster together resulting in having distinct health profiles within the dataset.

This shows the effectiveness of the clustering model across different data subsets.

Results

- Accuracy: k-Means showed superior accuracy in predicting health risks over linear and logistic regression models.
- Feature Impact: Significant predictors include BMI, blood pressure, and cholesterol levels.
- Optimal k: Identified optimal clusters balancing bias and variance.
- Healthcare Interventions: Provides insights for personalized healthcare, improving outcomes.
- Policy Making: Informs public health policies on lifestyle impacts.
- Patient Empowerment: Offers interpretable health risk assessments for informed choices. Computational Intensity: Requires significant computation, especially with large datasets.
- Feature Engineering Dependence: Performance heavily depends on quality of preprocessing.
- Scalability: Efficiency decreases with larger datasets, needing careful resource consideration.

Implications

- Healthcare interventions: The model provides insights that can guide personalized healthcare interventions, potentially improving health outcomes.
- Policy making: Insights into lifestyle impacts on health can inform public health policies and take initiatives.
- Patient Empowerment: The model offers interpretable and actionable health risk assessments, empowering patients to make informed lifestyle choices.

References

- Gladence, L. Mary, M. Karthi, and V. Maria Anu. "A statistical comparison of logistic regression and different Bayes classification m ethods for machine learning." *ARPN Journal of Engineering and Applied Sciences* 10.14 (2015): 5947-5953.
- Issabakhsh, Mona, et al. "Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study." *PLoS One* 18.6 (2023): e0286883.
- Song, Yu-xiang, et al. "Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study." *CNS Neuroscience & Therapeutics* 29.1 (2023): 158-167.
- Yang, Li, et al. "Study of cardiovascular disease prediction model based on random forest in eastern China." *Scientific reports* 10.1 (2020): 5245.
- Siddiqui, Muhammad Aadil, Abdul Samad Khan, and Gunawan Witjaksono. "Classification of the factors for smoking cessation using logistic regression, decision tree & neural networks." *AIP Conference Proceedings*. Vol. 2203. No. 1. AIP Publishing, 2020.
- Sykes, Alan O. "An introduction to regression analysis." (1993).
- Greenland, Sander, Judith A. Schwartzbaum, and William D. Finkle. "Problems due to small samples and sparse data in conditional logistic regression analysis." *American journal of epidemiology* 151.5 (2000): 531-539.
- Zeka, Ariana, Rebecca Gore, and David Kriebel. "Effects of alcohol and tobacco on aerodigestive cancer risks: a meta-regression analysis." *Cancer Causes & Control* 14 (2003): 897-906.
- Nishida, Nobuko, et al. "Determination of smoking and obesity as periodontitis risks using the classification and regression tree method." *Journal of periodontology* 76.6 (2005): 923-928.
- Bayaga, Anass. "Multinomial Logistic Regression: Usage and Application in Risk Analysis." *Journal of applied quantitative methods* 5.2 (2010).
- Cheng, A. C. K., et al. "The association between cigarette smoking and ocular diseases." *Hong Kong Medical Journal* 6.2 (2000): 195.
- Warnet, Jean-Michel, et al. "Relation between consumption of alcohol and fatty acids esterifying serum cholesterol in healthy men." *Br Med J (Clin Res Ed)* 290.6485 (1985): 1859-1861.