Members: Alex Nunez, Trinity Turner, Carson Hinson, Arman Kaur, Treasure Smith

Introduction:

This project will solve a relevant healthcare problem: how to evaluate and classify levels of risk for patients with machine learning. Identifying whether a patient is low, medium, or high risk can facilitate patient care, improve resource allocation, and potentially improve patient outcomes. This project aims to test whether we can appropriately utilize classification techniques to assess risk categories for patients based on their medical and lifestyle factors. In this project, we will apply classification algorithms to a patient dataset, create a model that can classify risk levels based on observed patterns, and evaluate the model's performance. The broader objective is to demonstrate clinical decision-making informed by data.

Data:

In this project, we used a dataset from Kaggle called "Patient Dataset for Clustering—Raw Data" by Arjunn Sharma. While this dataset was meant for clustering, it has structured health-related data appropriate for classification purposes after adding additional labels.

Dataset Details:

Number of records: 1000 patients

Number of features: 14

File format: CSV

Target variable (created for this project/Pilot-on): RiskLevel (Low, Medium, High)

Features consist of:

Demographics of the patients: Age, Gender

Lifestyle of the patients: Smoking, Alcohol, Activity Level

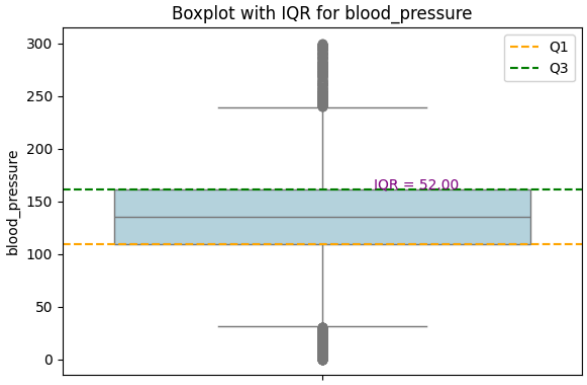
Vitals and Biometrics of the patients: Blood Pressure, Glucose Level, BMI, Cholesterol
Symptoms and Conditions of the Patients: Hypertension, Diabetes, Heart Disease, etc.
A new target variable of RiskLevel (Low, Medium, and High) was created for the supervised classification of the health-related data using clinical thresholds based on multiple health indicators (e.g., BMI, blood pressure, and comorbidity flags). Fundamental exploratory data analysis (EDA) was held to identify descriptive statistics and visualizations (histograms, correlation matrix and class balance chart) to understand the features' distribution and find areas with imbalances or deviations. Overall, the dataset represents how health-related data patterns are used to support training classification algorithms in a health-related context and begin assessing algorithms (i.e., decision trees, random forests, and logistic regression).

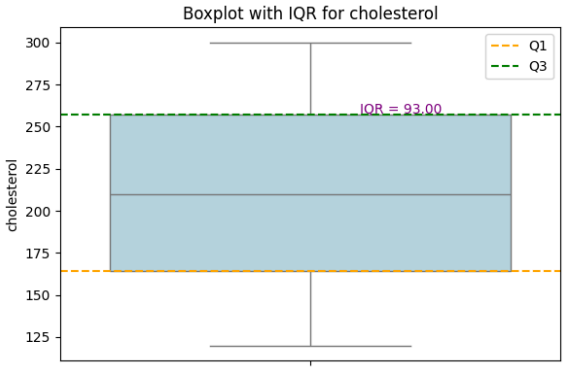
Methodology Plan:

Trinity:

Since the goal is to evaluate the risk of having high or low cholesterol, we will perform a K clustering algorithm to segment patients that have a low to high risk of having high cholesterol. The purpose behind this is to identify any potential causes that are unknowingly associated with having high cholesterol. Since the data is labeled, we also plan to compare that to a logistic regression and see if there are any algorithmic differences between the two. In the case that neither performs well, it is also appropriate to apply a random forest or decision tree to this type of data, especially due to the amount of binary variables in the set. The dataset is medium sized with 6000 entries and 2718 null values, almost half of the dataset. Many of these belong to plasma glucose and skin thickness, so in an effort to compensate for it, we will impute these

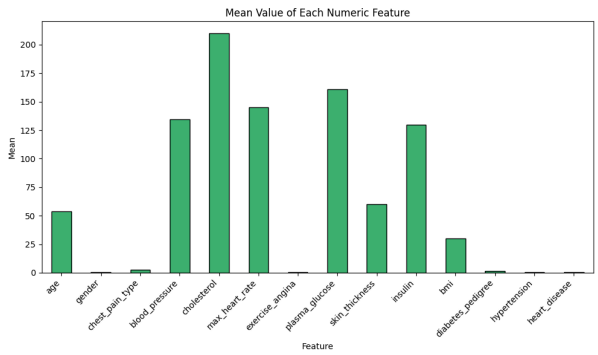
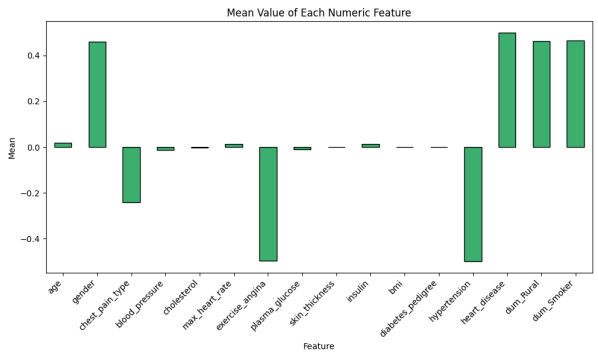
nulls using the median due to its resistance to outliers. Here is the dataset distribution before removing null values from columns with the most variability-





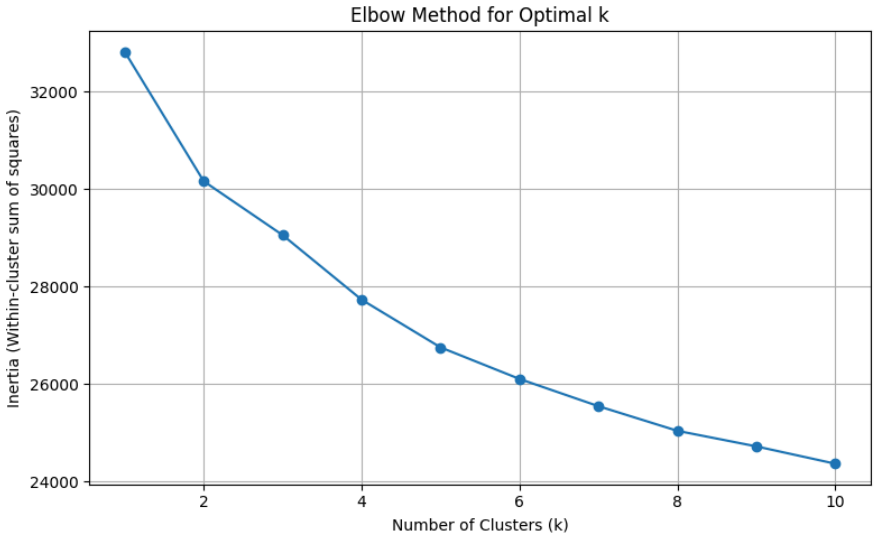
After imputing null values with the median, there was no change to the mean, standard deviation or outliers. Many features have no significant outliers, however each physiological measurement is placed on a different scale, which could affect K clustering and Logistic regression algorithms as they are sensitive to outliers. We then had to create dummy variables for features like 'Smoker Status' and 'Residence Type'. Next, we will apply a robust scaling method because it maintains relationships between data points while normalizing the data for clustering.

Here is a distribution of the dataset before and after null imputation, and scaling.

Before	After
	

The scaling had a profound effect, as the distribution of non-binary features was about normal, and has changed drastically – however it should not affect the outcomes of our logistic regression as it does not assume normal distribution.

Next, to find the optimal number of clusters we plotted the inertia/variability to identify where the variance between groups slows down, otherwise known as the Elbow Method.



As shown above, the optimal number of clusters would be 3 or 4.

Carson:

To investigate whether early signs of burnout can be detected through daily wellness patterns, we applied unsupervised clustering techniques to a time-series behavioral dataset. The dataset, collected via personal fitness tracking, contains over 500 daily records, including features such as step count, sleep hours, active minutes, water intake, calories burned, and mood. These features were chosen because they represent accessible and commonly tracked indicators of overall wellness. The project's core goal was to identify sequences of days in which individuals

exhibited low physical activity, insufficient sleep, and negative emotional states—patterns often associated with mental and physical decline. Our hypothesis was that by clustering these multi day periods, we could uncover subtle trends signaling increased burnout risk before it becomes apparent through clinical symptoms.

The methodology began with extensive preprocessing. Text-based mood entries were converted into numerical values using ordinal encoding to maintain their rank order (e.g., “happy” to “stressed” to “sad”). The date column was parsed into datetime format to support chronological operations, including calculating rolling averages. To reduce short-term noise and capture meaningful behavior shifts, we computed three-day rolling averages for sleep, active minutes, steps, and mood. This smoothing helped identify prolonged low-wellness periods rather than isolated bad days. These rolling averages became the core features for modeling. StandardScaler was then used to normalize all features, ensuring that no single variable dominated the distance-based clustering algorithms.

We implemented both K-Means and Agglomerative Clustering to group sequences of low-wellness days and compare model performance. K-Means was selected for its simplicity and speed, using the Elbow Method and KneeLocator to determine the optimal number of clusters (k = 3). This revealed one group strongly associated with reduced sleep, lower activity, and negative moods—matching our initial expectations for a burnout-prone cluster. Agglomerative Clustering, a hierarchical technique, was also used to capture potentially more nuanced trends. It identified smaller, more concentrated clusters with even lower wellness metrics, offering deeper insight into behavioral decline. Together, these two approaches provided a more complete view of wellness risk patterns, validating the use of clustering as an early-warning tool for detecting burnout through common, trackable daily behaviors.

Evaluation Plan:

To evaluate the performance of the classification models, we will use a combination of statistical metrics and visual tools that provide a comprehensive understanding of model effectiveness. After splitting the dataset into training and test sets, each model, Decision Tree, Random Forest, and Logistic Regression, will be trained on the training data and tested on unseen test data. Evaluation metrics will include accuracy to measure overall correctness, precision and recall to assess performance on each risk class, and the F1 score, which balances precision and recall. A confusion matrix will be used to visualize the distribution of true vs. predicted classifications across the three risk levels. Additionally, for models where it applies, the ROC-AUC score will be used to measure how well the model separates the classes. To ensure model reliability and generalizability, k-fold cross-validation will be employed during training, especially when tuning hyperparameters. This multi-metric evaluation approach will allow us to select the most effective model not just by overall accuracy, but by its ability to identify high-risk cases correctly, which is most critical in a clinical context.

Github Repository and Initial Tasks:

https://github.com/trin4turner/DM-Proj-Group-5

Group Expectations:

Failure to meet the group expectations can result in a letter grade reduction or removal from the team after many failures. Not communicating with the group or not completing a person's assigned project task will result in a warning. Each person will receive 3 warnings in total before being removed from the team. The first will be in our text group chat detailing where the

disconnect is and how our group should proceed moving forward. The second warning will be an email copying the entire group and the professor/TAs, again detailing the issue and how the group would like to move forward. The third and final warning will be a meeting, either on zoom or in person, between the team and the professor/TAs. This meeting will discuss all the issues with the group member and the group member will have a letter grade reduction on the final project. Lastly, if any more warnings are given, the group member will be removed from the group.