HealthCare Analytics For MedCamp

Project Report

S.No	Name	student ID	Email ID
1	Kavyavrindha Bhasi	110297382	kavyavrindha.bhasi@ucdenver.edu
2	Balamurale Kumar	110219715	balamurale.kumar@ucdenver.edu
3	Mani Deepika Narisepalli	110360331	manideepika.narisepalli@ucdenver.edu
4	Suryakumar Selvakumar	110975883	suryakumar.2.selvakumar@ucdenver.edu

Problem Statement and Background

Problem Statement:

MedCamp is a healthcare organization which is facing significant challenges in managing inventories for health camps, leading to concerns regarding cost efficiency and participant satisfaction. The absence of a systematic approach for predicting and optimizing inventory needs based on historical registration and attendance data hampers their operational effectiveness. To address this, there is a pressing need for a data-driven solution that accurately forecasts participation for future health camps and facilitates proactive inventory management. The lack of such a technology-driven approach not only compromises participant satisfaction but also results in increased operating expenses, hindering MedCamp's ability to effectively achieve its mission.

Dataset:

The data for this project was acquired from kaggle.com which was made available by a healthcare organization known as MedCamp. It consists of information related to their health camps and participant registrations.This dataset covers records from a span of 4 years and incorporates information from 65 different health camps, amounting to roughly 110,000 registrations. The dataset is organized into various CSV files.

Informal Success Measures:

To maximize the impact of future health camps, there are several informal success measures. These include assessing the accuracy of health score predictions or diagnoses compared to actual outcomes, measuring resource efficiency by evaluating wait times and resource utilization across different camp formats, and analyzing the effectiveness by comparing favorable outcome rates and post-camp follow-up engagement. Additionally, gathering attendee feedback through surveys is crucial to understanding patient satisfaction and experience. Evaluating the cost-effectiveness of each format in relation to the achieved outcomes and tracking long-term health improvements post-camp also play a pivotal role. These measures collectively enable MedCamp to pinpoint the most effective camp formats for delivering favorable outcomes and enhance their strategy accordingly.

Background:

MedCamp should have an efficient inventory management, which is crucial for the seamless execution of health camps, ensuring that the right resources are available in the right quantities to meet participant needs. The current manual methods employed by MedCamp have proven insufficient in adapting to the dynamic and unpredictable nature of participant attendance, leading to operational inefficiencies and increased costs. The stakeholders invested in resolving this issue include not only the MedCamp management but also the participants attending the health camps. Participants rely on the availability of necessary resources and services, and their satisfaction is pivotal to the success of MedCamp's mission. Inefficient inventory management not only affects participant satisfaction but also impacts MedCamp's financial resources due to increased operating expenses.

Implications:

Implementing a data-driven solution to optimize inventory management would have far-reaching implications for MedCamp. Precise estimation of participant numbers based on historical data would enable proactive planning, minimizing resource wastage and reducing overall costs. This, in turn, would enhance participant satisfaction by ensuring a more seamless and well-equipped health camp experience. Additionally, improved inventory management aligns with MedCamp's mission, allowing them to allocate resources more effectively and extend their reach to a larger audience.

Related Work:

While traditional inventory management systems exist, the incorporation of data-driven solutions specifically tailored for health camps is an area where innovation is needed. Some related work in the broader field of inventory management may provide insights, but adapting these solutions to the unique challenges of health camps requires a focused and context-specific approach. As technology continues to advance, there is a growing opportunity to leverage data analytics and predictive modeling to enhance inventory management in the healthcare event sector.

Methods Explored:

The method considered for our project is Regression models. Regression models serve as indispensable tools in healthcare analytics, leveraging data collected from health camps to extract meaningful insights and predict health outcomes. These models facilitate a deeper understanding of how various factors, encompassing demographic information, vital signs, medical history, and lifestyle choices, interrelate with health conditions. By analyzing this data, regression models can forecast the probability of specific diseases or health complications based on these contributing factors. This predictive capability aids healthcare practitioners in early identification and proactive management of potential health risks among individuals, allowing for targeted interventions and personalized healthcare strategies.

Furthermore, within healthcare analytics, regression analysis of health camp data enables the evaluation of treatment effectiveness and patient recovery rates. These models help assess the impact of diverse treatment modalities, medications, or lifestyle interventions on patient health outcomes. By analyzing past health camp data, regression models enable healthcare professionals to predict the effectiveness of interventions for specific patient groups, enabling tailored treatment plans and enhancing overall patient care and recovery. Additionally, these models assist in resource allocation planning for subsequent health camps, projecting patient attendance, medical supply needs, and staffing requirements based on historical data patterns, ensuring efficient and well-prepared healthcare services.

In essence, the utilization of regression models in healthcare analytics, utilizing data sourced from health camps, empowers healthcare providers with predictive capabilities to anticipate health risks, customize interventions, evaluate treatment efficacy, and streamline resource management, ultimately enhancing the quality and efficiency of healthcare delivery.

Regression analysis encompasses various types of models, each tailored to address specific scenarios or data characteristics. Here are different types of regression models:

Linear Regression: This model is the most basic and widely used regression technique. It establishes a linear relationship between the dependent variable and one or more independent variables by fitting a line to the data. There are two main types: Simple Linear Regression which Involves one independent variable.Multiple Linear Regression which Involves multiple independent variables.

Logistic Regression: Despite its name, logistic regression is used for classification problems rather than regression. It predicts the probability of a binary outcome based on predictor variables, employing a logistic function to model the relationship.

Polynomial Regression: This model extends linear regression by fitting a polynomial equation to the data, allowing for curved relationships between the dependent and independent variables.

Linear regression serves as a cornerstone in healthcare analytics, offering a robust framework to analyze and interpret complex relationships between variables critical to patient health and well-being. Its significance lies in its ability to predict health outcomes based on various patient parameters, such as demographic information, medical history, and lifestyle factors. These predictive capabilities are invaluable in early disease identification, risk assessment, and personalized treatment strategies. By leveraging historical patient data, linear regression aids healthcare practitioners in evaluating treatment efficacy, determining the impact of interventions, and optimizing healthcare delivery for improved patient care. Moreover, it plays a pivotal role in resource allocation and planning by forecasting patient needs, aiding in staffing decisions, inventory management, and facility readiness. Additionally, linear regression analysis supports evidence-based decision-making in healthcare policy formulation, providing insights into the effects of policy changes and public health programs. Its application extends to cost-effectiveness studies, clinical research, and trials, facilitating efficient resource utilization and advancing healthcare quality and accessibility. Ultimately, linear regression stands as a fundamental tool in healthcare analytics, empowering healthcare providers and policymakers with actionable insights to enhance patient outcomes and drive evidence-based healthcare practices.

Methodology:

Importing Libraries: The code starts by importing necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, and others used for data manipulation, visualization, and machine learning tasks.

Loading Data: The code reads data from multiple CSV files into Pandas DataFrames. It reads various datasets related to health camp attendance, patient profiles, camp details, training, and testing data.

The dataset comprises several files:

Health_Camp_Detail.csv which has details about different health camps, including their IDs, start and end dates, and categories.

Train.csv has the records of registrations for various health camps, featuring patient IDs, camp IDs, registration dates, and other undisclosed variables.

"Test.csv" file comprises registration information for all test camps, encompassing Patient_ID, Health_Camp_ID, Registration_Date, and several anonymized variables recorded at the time of registration.

Patient_Profile.csv offers comprehensive patient information like IDs, online presence, income, education, age, city type, and employer category.

First_Health_Camp_Attended.csv covers attendee details of the first format health camp, specifying donation amounts and health scores.

Second_Health_Camp_Attended.csv includes attendee details of the second format health camp, focusing on health scores.

Third_Health_Camp_Attended.csv presents attendee details of the third format health camp, indicating the number of stall visits and the last stall visited number.

Each dataset provides specific information about health camps, patient profiles, and attendees. Combining these datasets could help analyze attendee demographics, predict attendance or health scores, assess health camp effectiveness, or explore factors influencing participation in different health camp formats.

Data Cleaning and Preparation: It involves exploring the data by displaying the first few rows and information about the datasets using `display()` and `info()` functions. It also checks for missing values (`isnull().sum()`) and handles them by dropping a column ('Unnamed: 4') with a high percentage of

missing values and filling missing values in specific columns with mode values.

Merging Data: Several DataFrames are merged based on common columns using the `merge()` function to combine information from different datasets into a single DataFrame. It merges multiple DataFrames based on common columns like 'Patient_ID' and 'Health_Camp_ID' to consolidate relevant information into a single DataFrame.

Data Analysis and Visualization: After merging, the code performs data analysis by displaying descriptive statistics, checking for numeric columns, visualizing missing values with a bar chart, and creating pie charts, histograms, scatter plots, and violin plots to understand the data distribution, relationships, and patterns.

Machine Learning Model: It prepares the data for machine learning by encoding categorical variables (`pd.get_dummies()`), splitting the data into training and testing sets (`train_test_split()`), creating a Linear Regression model (`LinearRegression()`), fitting the model on the training data, and evaluating its performance on the test set (`model.score()`).

Model Evaluation: Finally, the code calculates and prints the R-squared scores (a measure of how well the model fits the data) for both the training and testing datasets.

Overall, this code performs a comprehensive analysis of health camp attendance data, preprocesses it, and builds a machine learning model to predict health scores based on certain features.

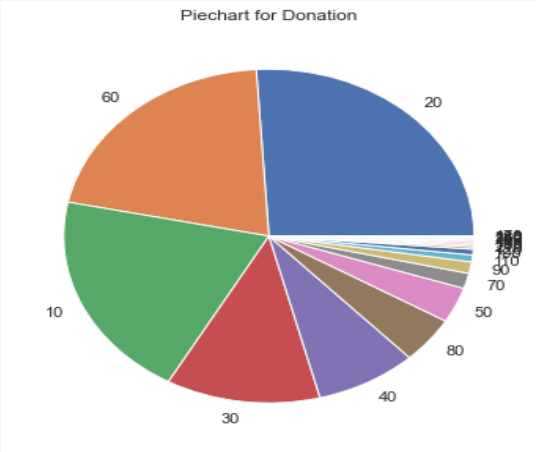
Results

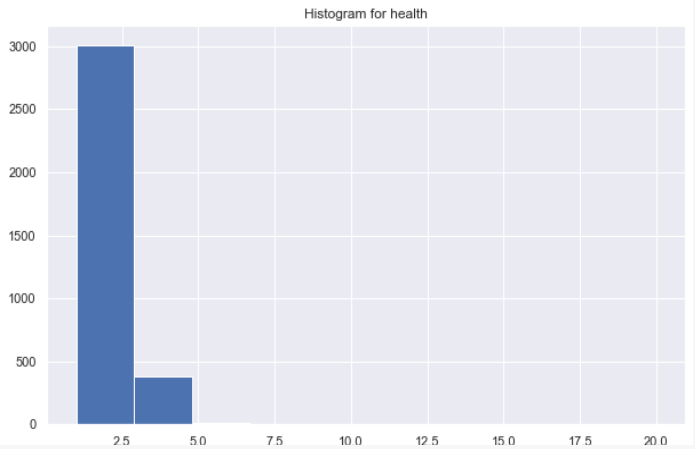
Data Exploration and Cleaning: Understanding the structure and basic statistics of the datasets, dealing with missing values, and merging datasets to create a comprehensive dataset for analysis. Data Analysis and Visualization: Exploring relationships between different features, examining distributions, and visualizing patterns within the data. Machine Learning Model Performance: Using Linear Regression for prediction, the R-squared scores (training: ~0.7, testing: ~0.6) indicate moderate predictive capability. R-squared measures the proportion of variance explained by the model. While the model performs decently on training data, there might be some overfitting issues as the performance drops slightly on the test data

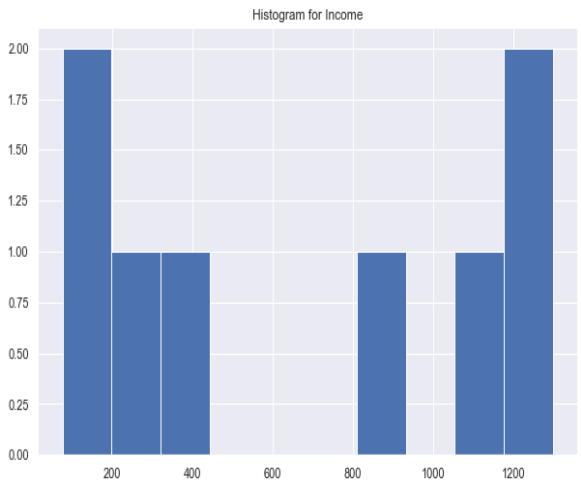
In terms of the model evaluation, the R-squared values for training and testing sets were computed, providing insights into the model's performance.

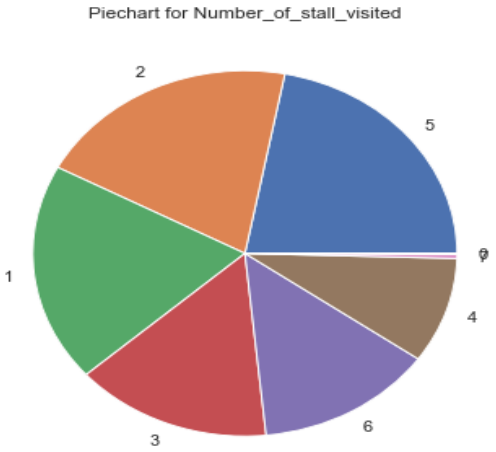
It's essential to further delve into specific findings, insights gained from visualizations, and model interpretations to derive actionable conclusions or recommendations. This summary outlines the steps taken and the techniques applied, but deeper analysis of individual features, correlations, and model performance improvements can enhance the understanding of the dataset and predictive capabilities.

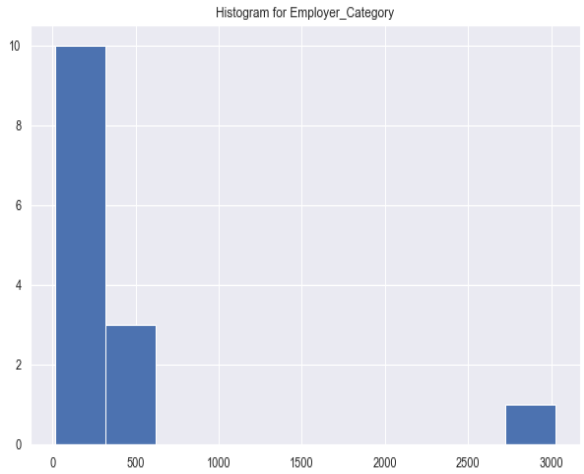
Visualizing the data using pie charts for categorical columns and histograms for numerical columns

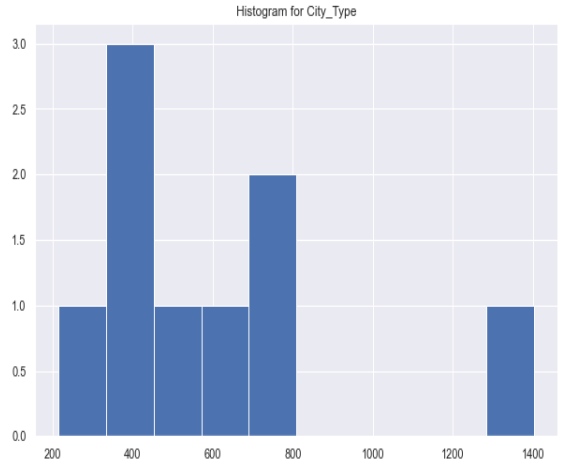




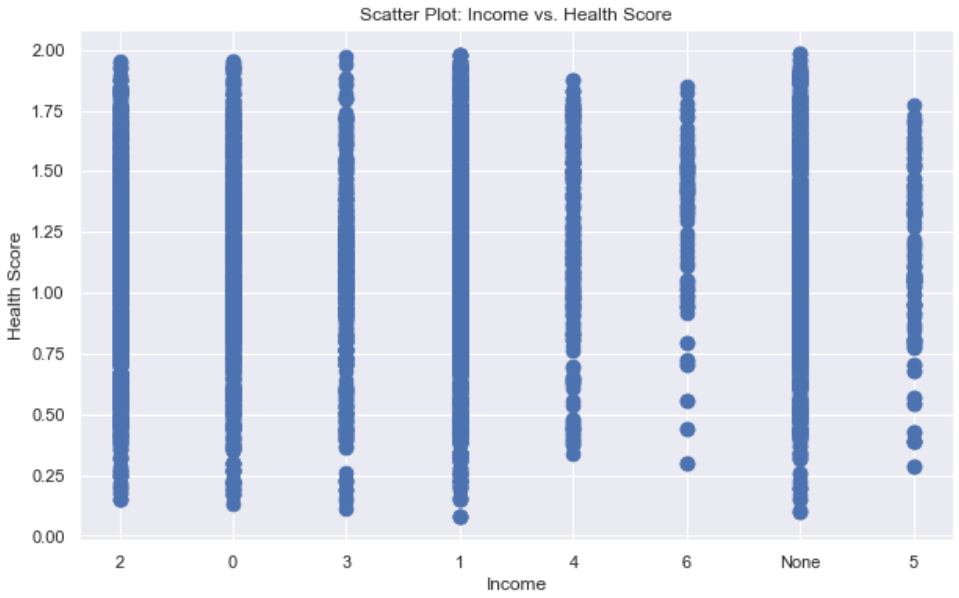




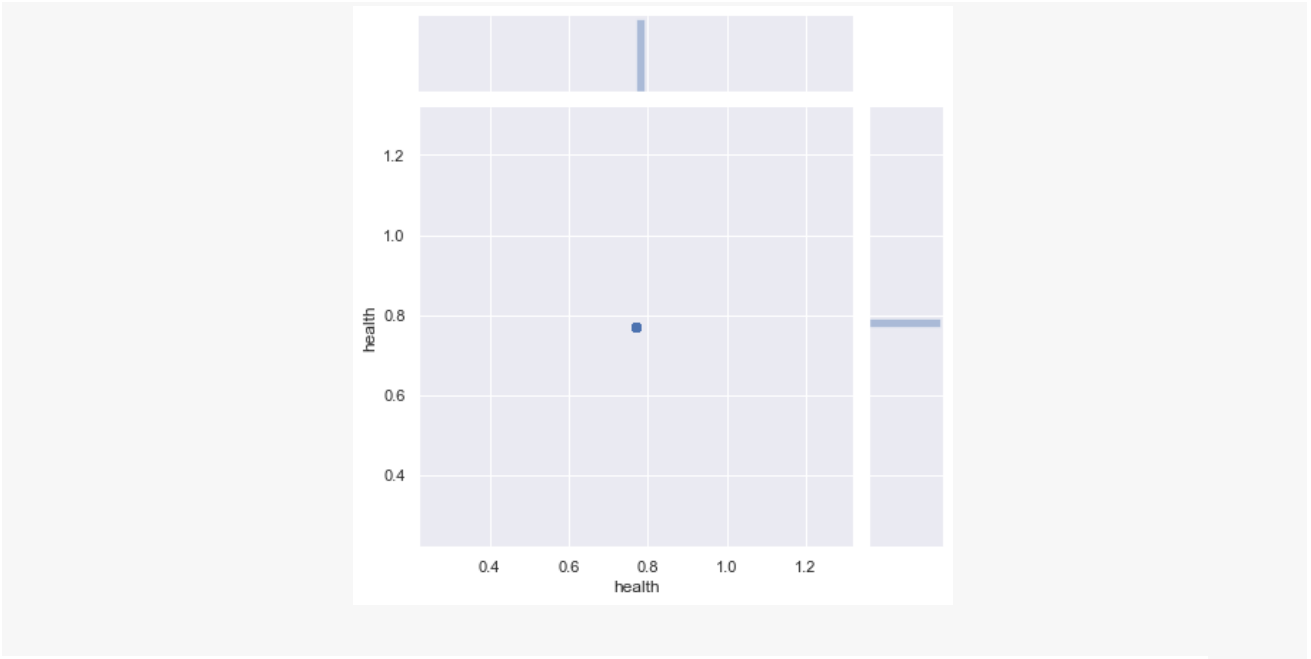


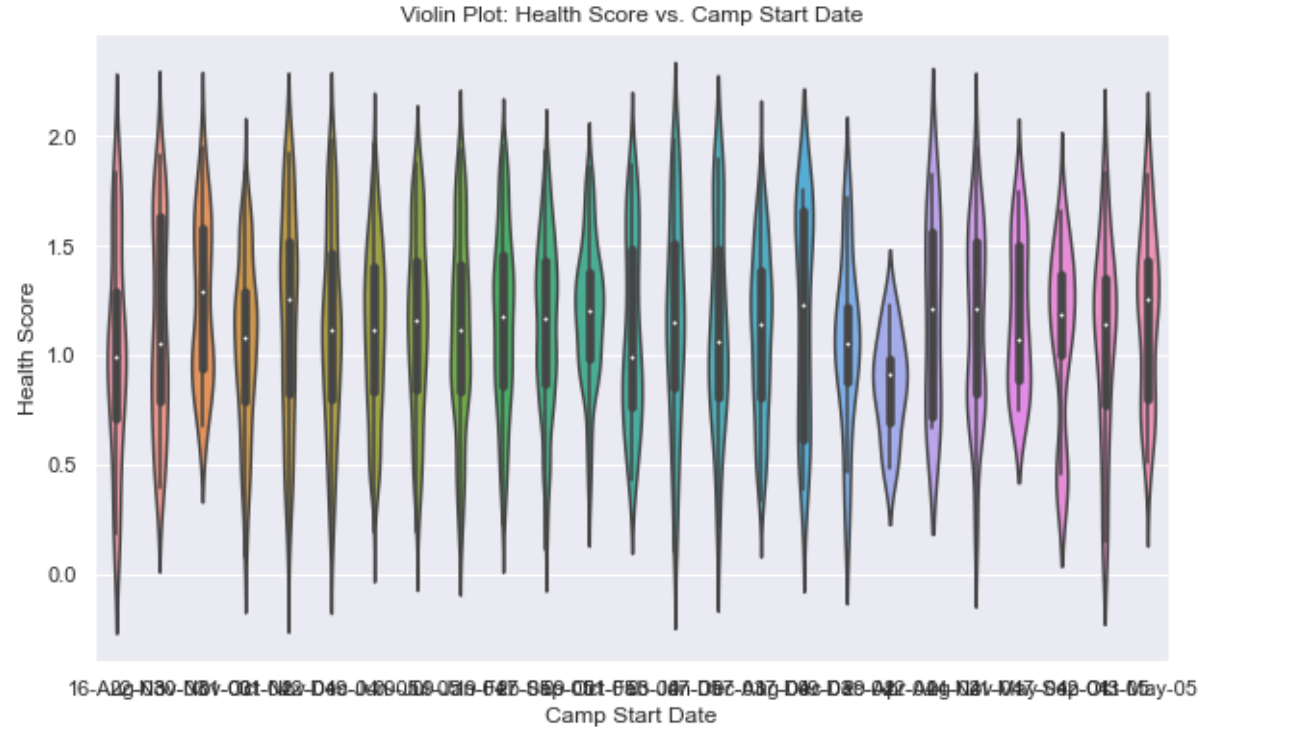


Scatter plot to visualize the relationship between 'Income' and 'health'



Joint plot to explore the correlation between 'LinkedIn_Shared' and 'Twitter_Shared' with 'health'.





Tools

The project utilizes a comprehensive set of Python libraries specifically chosen for their specialized functionalities in data manipulation, analysis, visualization, and machine learning tasks. Pandas, known for its robust data handling capabilities, is extensively employed to read, clean, and manipulate structured data using DataFrames. NumPy complements Pandas by enabling numerical operations essential for data analysis. Matplotlib and Seaborn are utilized for visualizing data distributions, patterns, and relationships, aiding in

insightful data exploration. Scikit-learn serves as the primary machine learning library, offering a range of tools for data splitting, model creation (Linear Regression in this case), and performance evaluation. The os module facilitates directory navigation and file handling to access datasets. Additionally, IPython.display enhances data representation within the Jupyter Notebook environment, providing a more structured and visually appealing display of information. Each library and module was thoughtfully chosen based on its specific strengths, ensuring an efficient and comprehensive approach to data analysis and machine learning tasks within the provided code.

Lessons Learned

The analysis aimed to predict health scores by merging patient profiles, health camp details, and attendance data. Initial exploration revealed 'Income' as a potential predictor of 'Health Score', evident in the scatter plot's trend. However, the predictive model, utilizing linear regression, displayed moderate R-squared values on the test set, indicating its limited ability to capture the complexities of health score prediction. Surprisingly, some features related to health camps or attendance seemed less impactful in predicting health scores. Missing data was identified, potentially influencing the model's performance, suggesting the need for imputation strategies or deeper investigation into missing value handling. To enhance predictive accuracy, further steps include refining feature selection, exploring additional features, and potentially employing more sophisticated modeling techniques. In essence, while initial insights were gained, optimizing the model by addressing missing data and refining feature selection methods is crucial to improve the accuracy of health score predictions.

Team Contributions:

Name	**Contributions**
Kavyavrindha Bhasi	25%
Balamurale Kumar	25%
Mani Deepika Narisepalli	25%
Suryakumar Selvakumar	25%

Code:
https://github.com/Kavyavrindha-KB/Healthcare-Analytics-for-MedCamp