Diabetes, Hypertension and Stroke Prediction

Group Members: Jash Jayant Shah (js3297), Kunjan Vaghela (kv353), Hrishikesh Sunil Salunkhe (hs1121), Dhruv Snehal Satyapanthi (ds1990)

Introduction:

In the realm of healthcare, the ability to predict and prevent chronic diseases such as diabetes, hypertension, and stroke is paramount. Our project embarks on this critical journey by harnessing the power of machine learning to develop accurate predictive models tailored to these prevalent health conditions. With a holistic approach encompassing demographic, health, and lifestyle factors, we aim to revolutionize healthcare delivery and empower individuals to take proactive measures toward better health outcomes.

Utilizing the Python library Streamlit, we've crafted a dynamic and intuitive user interface (UI) that caters to patients and doctors alike. All of the pages provide a prediction output per disease given a certain set of inputs in a dynamic form. On top of that, the patient-centric page of the UI provides users with exploratory data analysis (EDA), offering insights into behaviors that can facilitate positive lifestyle changes and improve overall health. Additionally, the doctor-centric pages on the app feature advanced attributes such as cholesterol levels and BMI, helping doctors make informed decisions and deliver personalized care.

Objectives:

In our quest to revolutionize healthcare, our project is guided by three primary objectives:

1. **Developing Accurate Predictive Models:** Our journey begins with developing sophisticated predictive models to identify the likelihood of chronic ailments like diabetes, hypertension, and stroke. By analyzing vast datasets brimming with health indicators and lifestyle factors, we aim to create models capable of accurately forecasting these conditions. Having tried several models for classification, we ultimately conclude that Logistic Regression performs best on the data. We get an accuracy of ~70% for all 6 models.

2. **Identifying Key Risk Factors:** Delving into the intricate web of health data, our mission is to uncover the critical factors driving the onset of these chronic diseases. Through meticulous EDA, we seek to unveil correlations and patterns within the data, shedding light on the pivotal risk factors contributing to these health disparities.

3. **Providing Empowering Healthcare Tools:** Central to our vision is creating user-friendly tools designed to empower patients and healthcare professionals. By integrating our predictive models into intuitive graphical interfaces, we aim to facilitate early diagnosis and preventive interventions, ultimately improving patient outcomes and fostering informed decision-making in healthcare delivery.

Problem Statement:

1. **Predictive Modeling:** Develop accurate models for diabetes, hypertension, and stroke prediction, considering demographic and lifestyle factors mentioned in the dataset.

2. **Key Risk Factors Identification:** Identify significant risk factors contributing to diabetes, hypertension, and stroke through extensive EDA.

3. **User-Friendly Healthcare Tools:** Design intuitive GUI tools for patients and healthcare professionals, enabling early diagnosis and personalized interventions.

4. **Integration of Predictive Models in the UI:** Seamlessly integrate predictive models into healthcare tools for real-time insights and informed decision-making. Trained the ML models and saved them in pickle files for further use for testing the user inputs.

5. **Patient Engagement Enhancement:** Enhance patient engagement with accessible EDA insights and personalized health recommendations on the patient-centric view of the UI.

Data Collection & Analysis:

Dataset Selection:

We finalized the dataset for this project which was sourced in Kaggle, specifically the "Diabetes Health Indicators Dataset" (https://www.kaggle.com/datasets/prosperchuks/health-dataset). This dataset comprises three subsets focusing on diabetes, hypertension, and stroke, each containing relevant health indicators and demographic information.

Dataset Information:

1. **Diabetes:** This subset consists of 70,692 rows and encompasses 18 features detailing various health metrics and lifestyle factors pertinent to diabetes prediction.

2. **Hypertension:** With 26,058 rows, this subset contains 14 features essential for analyzing hypertension risk factors and predictive modeling.

3. **Stroke:** Comprising 40,907 rows, this subset includes 11 features crucial for stroke prediction and understanding related health attributes.

Data Preparation:

Given the dual-purpose nature of the project, where both patient-centric and doctor-centric attributes are considered, the dataset underwent attribute segmentation:

1. **Patient-Centric Attributes:** These attributes are associated with routine behaviors and health indicators directly answerable by patients. For instance, attributes like "Good Mental Health days in a month" are representative of patient-centric data.

2. **Doctor-Centric Attributes:** These attributes encompass medical report inputs typically provided by healthcare professionals. Examples include indicators like "High Cholesterol" or other clinically measured parameters.

Data Analysis:

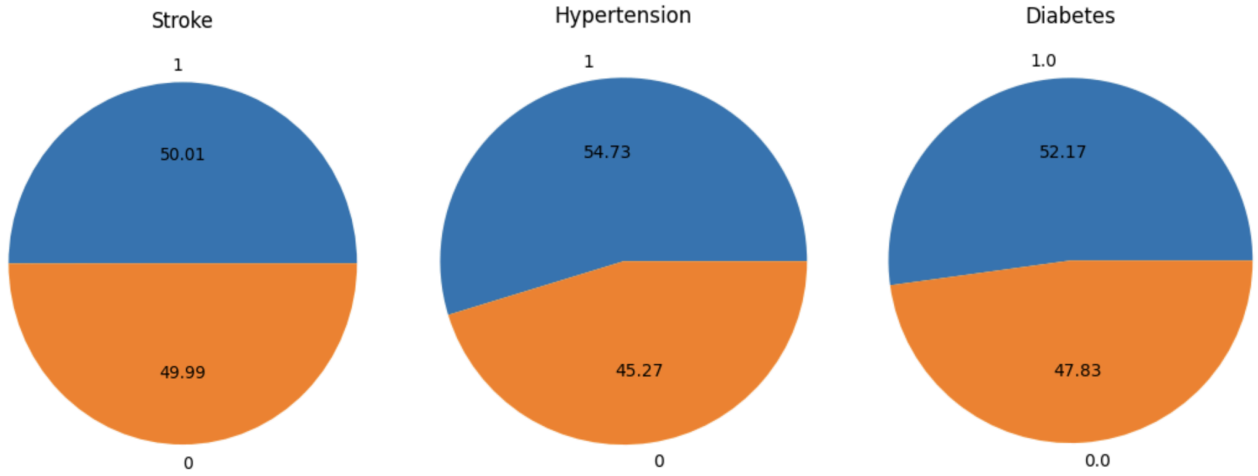
1. **Duplicate and Unique Values:** A preliminary analysis included identification of duplicate entries and assessment of unique values for each feature across the dataset. Notably, the diabetes dataset exhibited approximately 9% duplicate records, while other datasets had no duplicate records.

2. **Null Attribute Handling:** Null or missing attribute values were addressed through appropriate data imputation techniques to ensure dataset integrity and completeness.

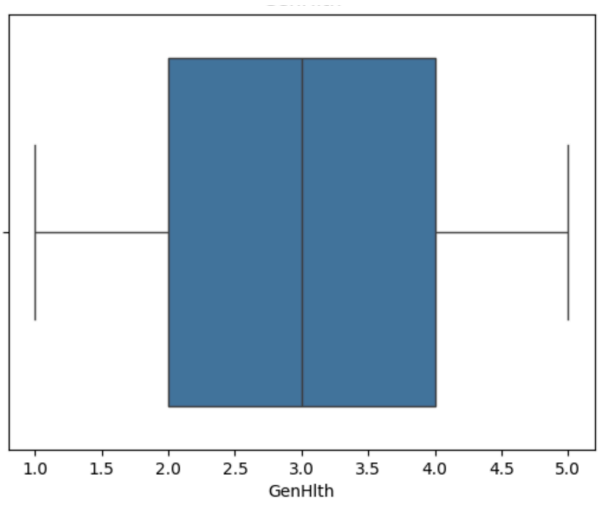
3. **Outlier Detection:** Robust outlier detection methodologies were employed, primarily leveraging Interquartile Range (IQR) and Z-Score analysis for multivariate attributes. This step ensured the identification and treatment of outliers, thus enhancing the reliability of subsequent analyses. For example, outlier detection check done for BMI feature from the Diabetes Dataset:

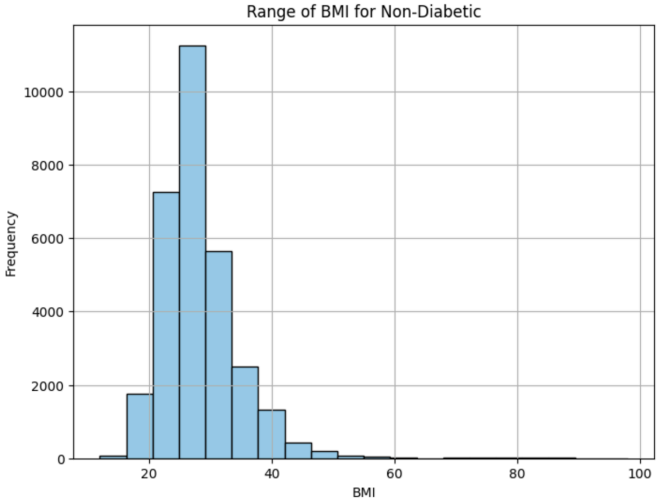
```
Unique values of  BMI  :
[26. 28. 29. 18. 31. 32. 27. 24. 21. 58. 30. 20. 22. 38. 40. 25. 36. 47.
 19. 37. 41. 23. 34. 35. 42. 17. 33. 44. 15. 52. 69. 56. 45. 39. 92. 53.
 98. 50. 46. 79. 48. 16. 63. 72. 54. 49. 68. 43. 84. 73. 76. 55. 51. 75.
 57. 60. 12. 77. 82. 67. 71. 61. 14. 81. 59. 86. 13. 87. 65. 95. 89. 62.
 64. 66. 85. 70. 83. 80. 78. 74.]
Outliers for the column (using IQR range)  BMI  :
[58. 52. 69. 56. 92. 53. 98. 50. 79. 48. 63. 72. 54. 49. 68. 84. 73. 76.
 55. 51. 75. 57. 60. 77. 82. 67. 71. 61. 81. 59. 86. 87. 65. 95. 89. 62.
 64. 66. 85. 70. 83. 80. 78. 74.]
Outliers for the column (using Z-score)  BMI  :
[58. 69. 56. 92. 53. 98. 79. 63. 72. 54. 68. 84. 73. 76. 55. 75. 57. 60.
 77. 82. 67. 71. 61. 81. 59. 86. 87. 65. 95. 89. 62. 64. 66. 85. 70. 83.
 80. 78. 74.]
```

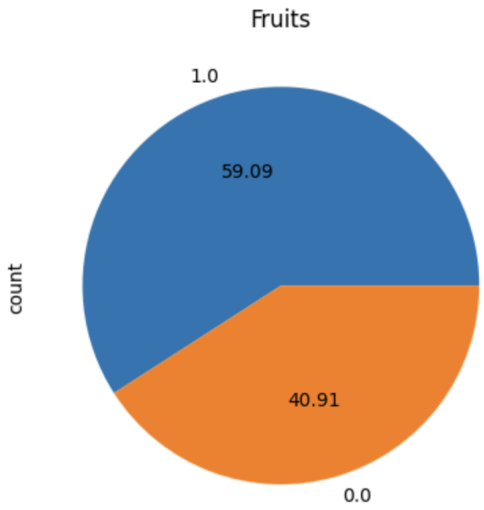
4. **Bias Assessment:** During our analysis, we ensured that our target attributes—stroke, diabetes, and hypertension—exhibited a balanced distribution within the dataset. All three target attributes displayed nearly equal representation of outcomes, indicating minimal bias. This balance assures the fairness and reliability of our dataset for subsequent analyses and model development.

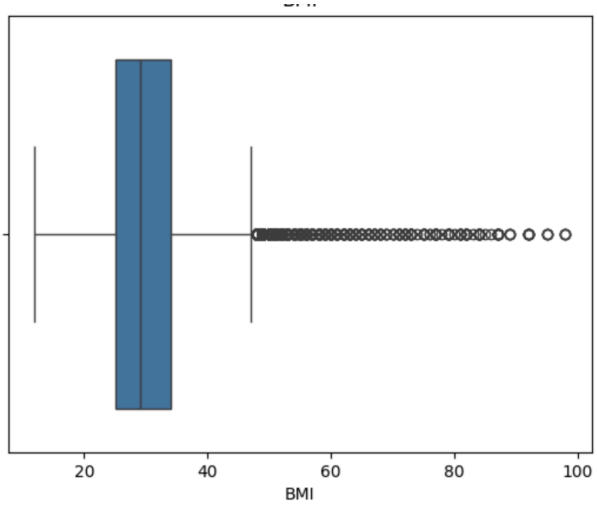


5. **Variability and Spread Analysis:** To gain insight into the distribution and characteristics of various health indicators within the dataset, we conducted a thorough analysis of data variability and spread. For binary attributes, we employed pie charts to visualize the distribution of values, facilitating a clear understanding of the proportion of occurrences for each attribute outcome. Meanwhile, multivariate attributes were analyzed using a combination of boxplots and histograms. These visualizations provided comprehensive insights into the spread and distribution of numerical data, enabling us to identify potential skewness and outliers.

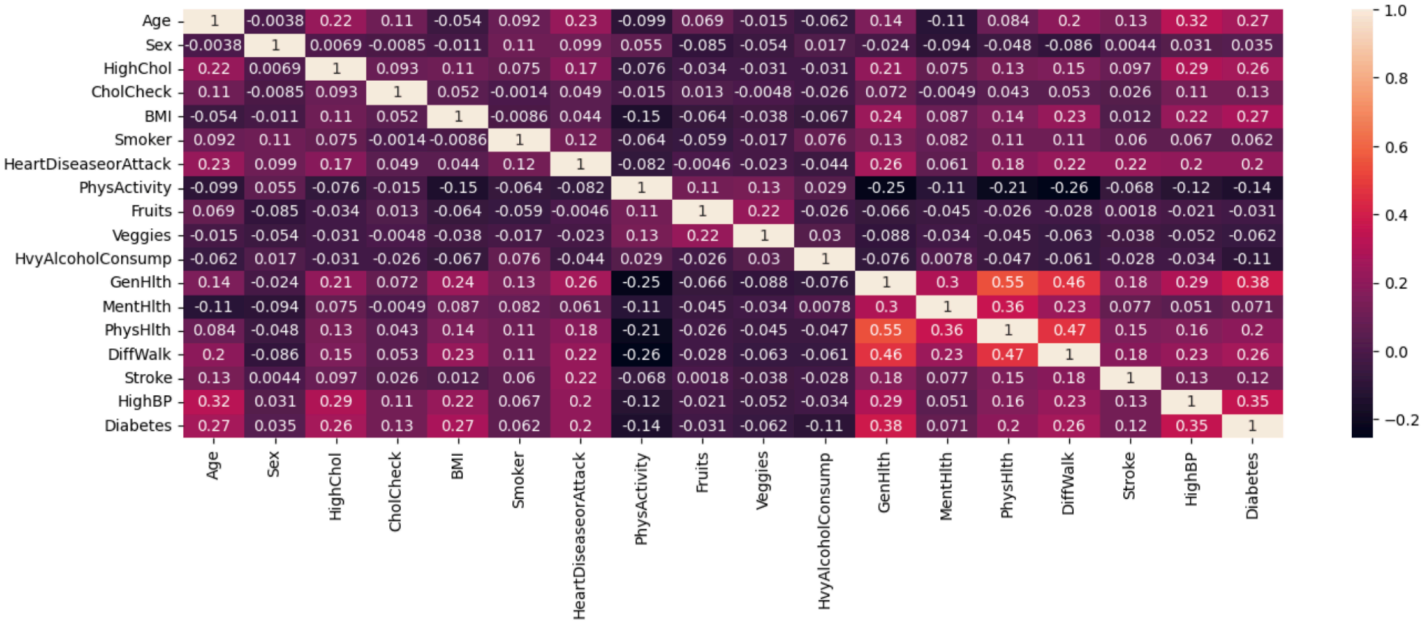








6. **Correlation Analysis:** A correlation check was performed to explore relationships between various attributes, unveiling potential patterns and interdependencies crucial for predictive modeling and risk factor identification. For example within the Diabetes dataset, we found both positive and negative correlations. For instance, attributes like general health and high blood pressure exhibited higher correlations with diabetes, while sex and mental health displayed lower correlations. This analysis guided attribute selection for modeling and provided valuable insights into dataset characteristics.



Methodology:

The process is divided into the following four sections apart from the Data Collection and Analysis process.

Data Preprocessing:

We first preprocess the data by executing the following steps.

1. **Removing null values:** We remove null values from the dataset so that the model does not run into any errors during its training. We achieve this using the pandas.DataFrame.dropna() function

2. **Removing duplicate rows:** We then remove the duplicate rows from all of our datasets so that the training labels do not get biased to any one output. We use the pandas.DataFrame.drop_duplicates() function for this task.

3. **Min-max normalization:** Finally, we scale down the range of all values to the [0, 1] interval using min-max normalization so that all ranges of values are homogenous and easier for the model to train on.

Model Selection:

We explored several different predefined classifiers provided by the sklearn Python library. We compared the different models based on their accuracies and selected the one with the highest accuracy for our task. The different models and their corresponding accuracies are as follows.

1. **sklearn.linear_model.LogisticRegression: 0.7121**

2. sklearn.linear_model.SGDClassifier: 0.6798

3. sklearn.naive_bayes.GaussianNB: 0.6667

4. sklearn.tree.DecisionTreeClassifier: 0.6669

5. sklearn.ensemble.RandomForestClassifier: 0.6820

As observed above, the Logistic Regression classifier provides the highest accuracy to our task. Thus, we go forward with this model for all of the classification tasks of our project.

Model Training:

We do a train-test split of 70% training data and 30% test data for all the datasets.

We experimented with different values of the Logistic Regression classifier's hyperparameters and since it did not lead to a huge difference we decided to move ahead with the default values.

Frontend UI:

Based on our core idea, we have created a UI containing patient-centric view and doctor-centric view. For the patient-centric view, we are providing one page and for the doctor-centric view, we are providing three pages. We are covering identification of Diabetes, Stroke and Hypertension for both views. We are providing buttons to switch the views.

Inputs in the UI are mainly of two types i.e. Slider and Select box. For numeric fields, slider with appropriate step size of the feature and select box for other fields. Outputs in the UI are provided in the form of progress bars filled with blue color. Also, we have printed the score below the progress bar. Score is between 0-100 with 0 being least likely and 100 being most likely. Finally, we have added recommendations in the patient-centric view.

We created the UI using the Streamlit package in Python. We created a dynamic front-end which reflects the changes in real-time. Further instructions for application execution are provided in the instructions.txt. We saved the models in the form of Pickle files and loaded these Pickle files as the models in the application.

Results:

Here are the accuracies of the Logistic Regression classifier that we achieved over each dataset:

1. Patient-centric Diabetes Prediction Score: 0.71
2. Patient-centric Stroke Prediction Score: 0.73
3. Patient-centric HighBP Prediction Score: 0.70
4. Doctor-centric Stroke Prediction Score: 0.68
5. Doctor-centric Diabetes Prediction Score: 0.70
6. Doctor-centric Hypertension Prediction Score: 0.85

The following is the final set of UI pages created for the different views discussed above.

1. **Patient-centric view:** As you can see in figure 1, you can navigate to different pages and on each page, our format is the same with inputs on the left side and score on the right side. Additionally, we have added recommendations in the patient-centric view which can be seen in figure 2.

2. **Doctor-centric view:** Consists of three different pages for Hypertension, Stroke and Diabetes prediction each. As you can see in figure 2, figure 3 and figure 4, inputs are on the left and score is on right.

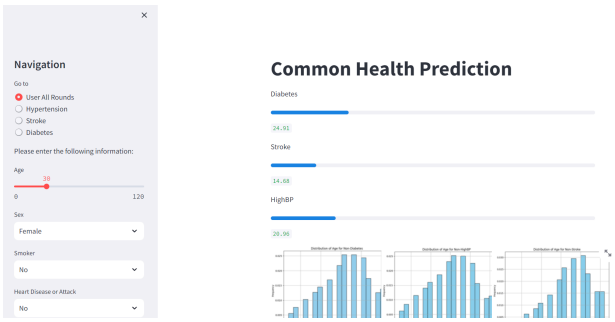


Fig 1

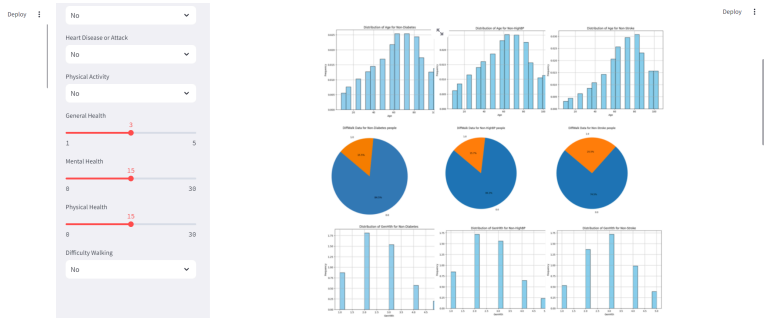


Fig 2

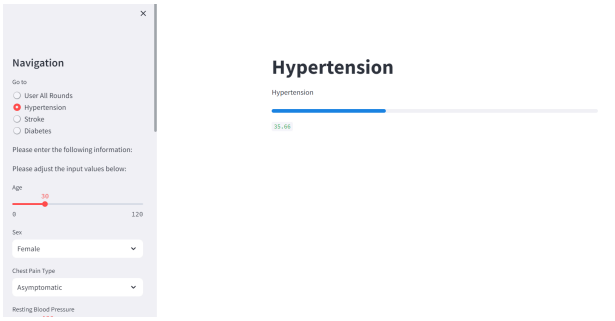


Fig 3

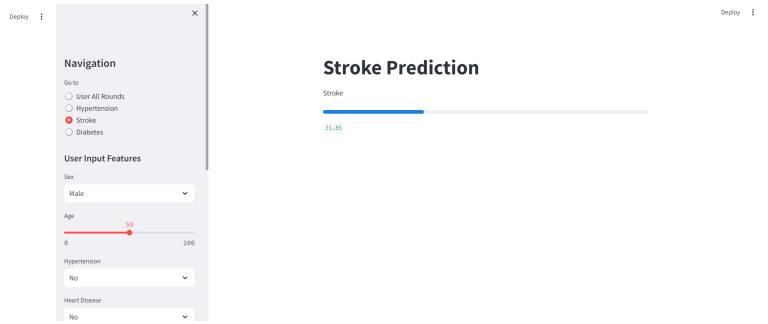


Fig 4

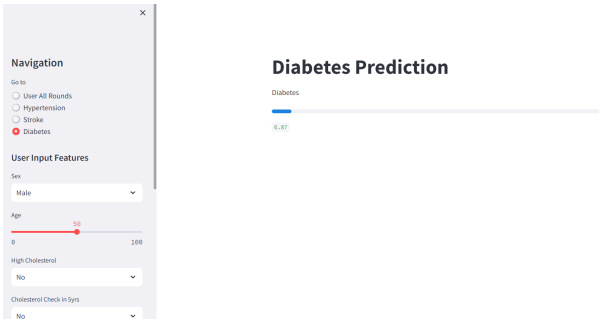


Fig 5

Conclusions and Findings:

We applied our learning from the correlation map and tested it out on the UI to see the effects of each input feature on each output feature. We aimed to find the features that matter the most and these are our findings:

1. **Patient-centric view (Diabetes, Stroke and Hypertension):** Heart Disease or Attack, Difficulty Walking and General Health

2. **Doctor-centric view (Diabetes):** Age, High Cholesterol and BMI

3. **Doctor-centric view (Hypertension):** Chest Pain Type, Exercise Induced Angina, ST Depression Induced by Exercise Relative to Rest, Slope of the Peak Exercise ST Segment, Number of Major Vessels (0-3) Colored by Flourosopy and Thalassemia

4. **Doctor-centric view (Stroke):** Hypertension, Heart Disease, Ever Married and Average Glucose Level

To summarize, these are the salient features of our project.

1. **Accurate models:** We created all the ML based Logistic Regression models using scikit-learn with 70%-80% accuracy.

2. **Dynamic UI:** We created patient-centric and doctor-centric views to provide the chances of Diabetes, Stroke and Hypertension through a dynamic UI with four pages.

3. **Important Features:** Using correlation map and UI, we listed down the important features for each attribute for each view.