Diabetes, Hypertension, Stroke Analysis

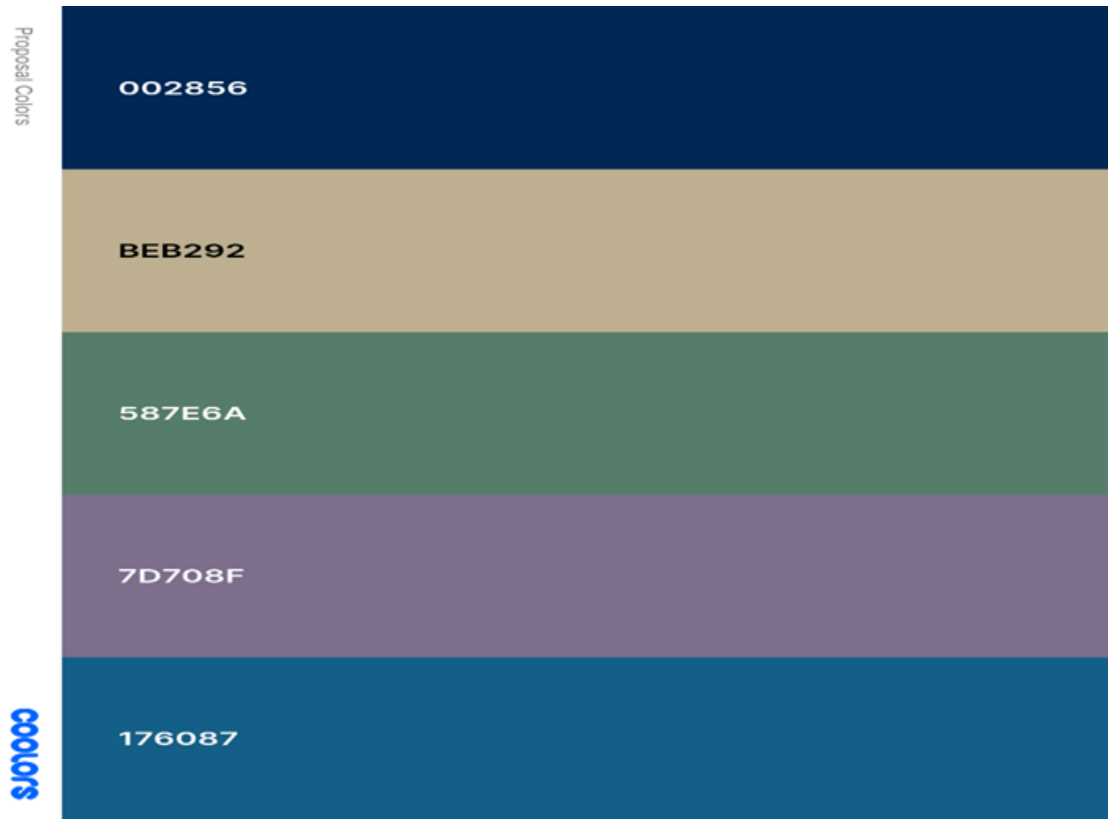
Project 4 Group 16:

Introduction

In this project, we aimed to analyze and predict diabetes, hypertension, and stroke. Our dataset is sourced from Kaggle. Our primary goal was to create interactive visual dashboards and machine learning models that would give us a better understanding of these illnesses. We utilized a Flask on the backend to manage API requests and handle data querying, while the frontend was built using BootStrap and JavaScript. This project provides insights into multiple different features due to the utilization of three notebooks being used.

Data Cleaning, Tableau and Model Creation

The dataset was initially provided as a CSV file and needed minimal cleaning, as it was already well-structured. The main modifications required were dropping unwanted columns that maximized our machine learning predictions and also formatting columns for clarity to visualize in tableau. After cleaning, we worked on creating the machine learning models and importing the CSV files into tableau. For the visual design, we chose a color palette :



Visuals used in tableau were:

a. Scatter Plot - For the diabetes dataset, age and BMI were compared . This was used to effectively show where people with a diabetes diagnosis were on the BMI scale according to their age group. This showed us conclusions based on gender and BMI by age. For the stroke and hypertension dataset, this was used as a way to be able to consistently observe the difference between sexes and show the strongest correlations between healthy patients and patients who did in fact have said ailment.

b. Bar Chart - Used to visualize age breakdown by physical health for our diabetes dataset to show the distribution of diagnosis throughout the age brackets of our data set. This was shown to make the colors dynamic to more quickly convey to the reader the ages that needed to be highlighted the most. For stroke and hypertension, the visual comparison between the count of said ailment and healthy patient is to be able to use the filter and observe the change on the graph to see a correlation in it being higher or lower depending on your own condition and filter out. This was to be able to observe the correlations between the filters for many of the columns.

c. Cluster Plot - The diabetes cluster plot was used to compare mental and general health and if there was a related diabetes diagnosis. Due to the quantity of data points on this visual, it was initially ineffective as something that a conclusion could be drawn from. So the data was aggregated in Tableau using the (AGG) function to produce a visual that was clearer in its results.

Website Architecture

Our website is structured into seven main pages:

1. Homepage: Provides an overview of the project and research questions.
2. Stroke_Hypertention Dashboard: Displays various visualizations, allowing users to filter data.
3. Diabetes Dashboard: Displays various visualizations, allowing users to filter data.
4. Predictions: Predicts if your chances of having diabetes, hypertension, or stroke
5. About Us: Offers background information about the team and the project.
6. Works Cited: Lists the sources and tools used in the project.
7. Write Up: Embedded copy of our write up on the site.

This structure ensures that users can easily navigate through the site and access the information they need.

Machine Learning

Machine Learning for the diabetes dataset utilized 'Age', 'Sex', 'HighChol', 'BMI', 'Smoker', 'GenHlth', 'MentHlth', 'PhysHlth', 'HighBP', and 'Diabetes' as features selections to make

predictions. These columns were chosen after looking at the value counts of all of the columns and dropping the columns that were imbalanced. There was no need for a scaler being all data was relatively around the same ranges. Next, we finalized the data frame that would be used for predictions. We used the train_test_split function to separate our data. “doRegression” and “doClassification” were two functions we created to render the test results. “doClassification” was the better of the two at producing productive models. Testing multiple algorithms, we found that AdaBoostClassifier was the best model based on its performance. AdaBoostClassifier had an AUC score of 0.82 which tied other models. The model fit almost perfectly showing no signs of over or under fitting. The deciding factor was the feature importances. AdaBoostClassifer had Age, BMI, and GenHlth as its drivers for training the model. These features made the most sense compared to other models' important features.

For the stroke model, all features from the dataset were kept since there were too few to drop. These were: 'age', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', and 'smoking_status'. We decided that a forest classifier would be the best type of model to use since the data is too complex for a typical logistic regression to effectively categorize. We created several tree-based models and compared their performance. We based our assessments on the accuracy of the model with special consideration paid to false negatives. Because this is health data, it can be a major issue to miss a diagnosis. With all of this in mind, the Gradient Boosted model performed the best while also being compatible with our production pipeline. It has an accuracy of 81% and a recall of 80%. The LightGB and XGBoost models performed better with an accuracy of at least 98%, but were incompatible with our method of deployment. What we found, both in the data analysis and in our model, is that the marriage status is the most important indicator of stroke risk. The Gradient Boosted model uses average glucose level as its primary feature, however.

The hypertension model was designed in a similar way. The features from this dataset are: 'age', 'cp' (chest pain), 'trestbps' (resting blood pressure), 'chol' (cholesterol), and 'thal' (Thalassemia). Several of the features needed to be dropped since they would not have made sense for this project’s use-case. They included health data that most people visiting our website would not have access to or understand. We also decided to test tree-based models for this dataset and found that a Gradient Boosted Tree model worked the best while also limiting false negatives (accuracy=93%, recall=97%). For this model, chest pain and cholesterol level are the most important features which align with what we know about hypertension and heart disease.

Dashboard Design Concepts

Our dashboard design concepts were based around other Tableau Public visualizations that we found for healthcare based topics. From this, we looked for visuals that were clear and effective in describing the intended message of the corresponding dataset. This helped us with the next steps of our design, while also considering color palette, the layout of the dashboards were made to be similar for the ease of viewing for the reader. While giving them an on screen overview of what the visual represents and can do, the variety of filters gives the user an opportunity to draw their own conclusions and data-based conclusions.

Answering the Research Questions

The dashboard allows us to explore several research questions:

1. What are some of the most common risk factors for diabetes/hypertension/stroke?

    - Age, BMI, and GenHlth for diabetes
    - Married, Hypertension, and Heart Disease for stroke
    - Chest Pain, Cholesterol, and Heart Rate for hypertension

2. How accurately are we able to predict someone's risk of developing one of these conditions?

    - 82% for diabetes
    - 98% for hypertension
    - 99% for stroke

3. Are there specific age groups that see a higher rate of development of a specific condition?

    - 55 – 70 is the peak where you see more of these illnesses. After 70 years of age there is a drop off.

4. Does fruit/vegetable consumption have an impact on diagnosis rates?

    - These columns were dropped due to imbalance.

5. Does gender have an impact on diagnosis rates?

    - Gender only played a role in the diabetes model.

Bias and Limitations

Some of the data is generated, referred to in the description as 'augmented'. Certain columns within the data are based off of the patient/users discretion, which would leave room for people to not be giving the most accurate reflection of their actual health, highlighted in the 'GenHlth', 'HvyAlcoholConsumption', 'MentHlth', 'PhysHlth' columns. Most of the columns measured are binary, 1 for yes, 0 for no for the corresponding column. Age is bracketed, meaning 1 = ages 18-24 and so on. Columns will be dropped based on a heatmap of correlation.