Group 26: Benjamin Ecsedy, Andrew Loose, Yash Nalluri

DS4200 Information Presentation & Data Visualization

Prof. Yang

13 April 2025

Website Link: <https://becsedy.github.io/ds4200_final/>

**Project Design Explanations**

**Feature Importance Visualization:**

To understand which factors are most useful in predicting diabetes, we used a Random Forest Classifier to calculate feature importance scores. These scores represent how much each input feature contributes to the model’s ability to make accurate predictions. Once we had these scores, we focused on the top 12 features.

We visualized these 12 features using a lollipop chart, which is a clean and easy-to-read way of showing ranked values. In this chart, each feature is shown as a horizontal line with a dot at the end representing its importance score. The longer the line, the more important the feature is. We ordered the features from least to most important going top to bottom, so it’s simple to compare their impact at a glance.

This style of chart was chosen because it avoids the clutter of bar charts and draws attention to the relative difference between features. It clearly showed that features like BMI, Age, High Blood Pressure, and Cholesterol were among the most influential.

The visualization helped us justify which features to include in our final predictor and also guided how we grouped features into controllable and uncontrollable factors in the rest of our analysis.

**“Uncontrollable/Immutable” Designs**

The Sankey diagram shows the flow of patients between diabetes statuses and various health indicators. The diagram consists of nodes representing different categories: Left-sided nodes show health indicator categories (BMI levels, blood pressure status, cholesterol levels, etc.), and right-sided nodes show diabetes status categories (No Diabetes, Prediabetic, Diabetic). The width of each connecting flow indicates the number of patients in that particular pathway. Thicker flows represent larger patient counts. The nodes are color-coded by category type (blue for diabetes status, different colors for each health indicator type). There is also a menu at the top of the diagram, where users can select which health indicators to include in the visualization using the modern checkbox interface, allowing for focused analysis of specific relationships. The intention with this graph was to show the overall ways that indicators have been represented in the status of diabetes within the graph itself. There is an obvious trend towards not having diabetes, as most of the population does not have the disease, but the flows help us visually represent the ratio of these health indicators as they contribute to a diabetes diagnosis.

For the Correlation Matrix, the design is meant to visualize the relationships between various health indicators and diabetes status through color-coded cells. Each cell represents the correlation coefficient between two variables (ranging from -1 to +1). The variables are displayed on both axes, creating a grid where each intersection shows how strongly two factors relate to each other. The matrix utilizes a color distribution from red to blue. Red indicates negative correlation (as one variable increases, the other tends to decrease), white indicates no correlation (variables change independently), and blue indicates positive correlation (variables tend to increase together). Utilizing the hover functions within D3, the matrix implements a tooltip feature that appears when the mouse is overtop of a particular container, showing the exact correlation value and its strength interpretation. The intention of the correlation matrix is to show which of the immutable indicators are related to each other compared side by side to what their relationship is with the categories of diagnosis themselves. This makes it easier to visually understand both trends between indicators themselves, for which there is merely very weak positive relationships except for a few outliers that reach above +0.3, and allows us to see the directional difference between these categories and their relationships with a diagnosis of non-diabetic and diabetic.

**“Controllable/Mutable” Designs**

Patient Similarity Network

When trying to find a diagnosis for the symptoms of a given patient, it is helpful to rely on previous diagnoses, and compare the features of that patient to those that have already been diagnosed to find similar patterns or trends. Thus, seeing the underlying connections between patients and observing if there are any distinct clusters or other relationships between the different diagnoses can be helpful not only for visually seeing a distinction between those with no diabetes, prediabetes, and diabetes, but also to communicate how the underlying features being studied relate to one another on an underlying level, where more clustering implies a stronger relationship between the features and less clustering implies a weaker relationship. Because each edge represents a connection between a patient and their most similar counterpart based on the mutable features, we can remove excess noise and unnecessary edges in the data, “stripping it down” to only the most helpful information, hopefully encouraging the patient to “learn” from their most similar other patients about the risk they potentially face, thus allowing them to change their lifestyle habits. The circular layout helps distribute nodes evenly and reveals clustering patterns without spatial overlap, since there is no “preferred” or “main” patient; however, if this visualization were to be shown to a specific patient with given features, it would make sense to place them in the middle. By using color to distinguish diabetes status and maintaining balanced group sizes, the visualization highlights how patients with different diabetes outcomes are distributed across the similarity space, and taking an equal sample of diagnoses prevents the visualization from being drowned by an overwhelming sea of blue non-diabetic nodes. The design decisions were made to allow for a clear comparison of how individuals with similar health profiles relate to one another, and it helps identify whether patients with the same diagnosis tend to cluster or mix with others.

Diagnosis Distribution Selectors

This interactive dashboard presents a distribution of diabetes classifications across age groups, filtered by user-selected health and lifestyle factors. The patient’s age was chosen to be on the x-axis because it has the most potential options (8 different age buckets), compared to 2 or 3 in the other categories, creating a more satisfying and informative visualization for the viewer. Below the chart, dropdown menus allow users to adjust key predisposing factors such as education level, mental and physical health, income, diet (fruits and vegetables consumption), exercise habits, and smoking status, and this customization gives the most availability in exploration for the user. Because the chart dynamically updates to reflect these selections, it becomes a powerful visual tool for exploring how different combinations of social and behavioral factors are associated with diabetes risk across age groups, in hopes that a specific user can potentially forecast how their probability of developing diabetes could change over time given their mutable cahracteristics. This design also provides targeted public health insights, enabling users to observe patterns, such as showing which characteristics lead to an increased risk of diagnosing diabetes in specific demographics, while also providing potential protective effects of healthy habits within populations that are especially vulnerable to developing the disease.

**Diabetes Risk Prediction Model (Web-Based Tool)**

To make our project interactive and meaningful for users, we developed a browser-based prediction tool that allows people to input their health data and receive a prediction of their diabetes risk. This tool uses a Random Forest Classifier trained on the same dataset used for our visualizations.

We selected the top 12 most important features from our feature importance analysis and trained the model using 80% of the data, testing it on the remaining 20%. The model achieved an accuracy of 51.85% on unseen data. While the accuracy is modest, it was enough to demonstrate patterns and relationships between health factors and diabetes outcomes.

The model was then converted to JavaScript using a model conversion library, which allowed us to run the prediction directly in the browser without needing any backend server. This made the tool fast and accessible.

The final predictor form includes all 12 features, with clear labels and explanations. When a user submits their input, the model instantly returns a prediction: No Diabetes, Prediabetic, or Diabetic, along with a confidence percentage. The form is styled to match the rest of the site and is fully responsive.

This predictor adds a practical layer to our project, allowing users to engage with the data and understand how their personal health inputs relate to diabetes risk.