DIABETES HEALTH INDICATORS DATASET

Project 4 - Group 7

Sierra Sarkis,
Neyda Morales,
Hannah Lashway,
Thripura Pakala

Project Overview

- **Objective**: Showcase machine learning experimentation, BI visualizations, and app development skills by designing a full-stack AI-powered data application.
- **Dataset:** Diabetes Health Indicators Dataset

Introduction

This project aims to demonstrate data cleaning, machine learning experimentation, data visualization, and web development skills by creating a full-stack AI-powered data application that tells an interactive and compelling data story. The application features a Flask backend, and an HTML/JS front-end to facilitate user interactions and provide a seamless web experience. The dashboard enables personalized exploration of the dataset by allowing individuals to select from various adjustable factors when using our diabetes predictor. Additionally, we created two professional-grade dashboards in Tableau and embedded them into web pages to enhance the data storytelling experience. This project builds on concepts learned over the past six months in the Data Analysis and Visualization Certification Course.

We selected the Diabetes Health Indicators Dataset for its comprehensive health-related information, encompassing demographic details, lifestyle factors, medical history, and diagnostic measurements. This dataset allows us to explore the complex relationships between these variables and the risk of developing diabetes. As part of the project, we cleaned the dataset and ran machine learning experiments to predict diabetes outcomes. On the development side, the web application is powered by Flask, with JavaScript for backend-frontend logic, HTML/CSS for the frontend design, and embedded Tableau visualizations.

Dataset Cleaning

Our initial dataset did not contain any null values. To begin the data cleaning for our machine learning experimentation, we started by transforming the Diabetes_012 column which originally contained three values: 0 - no diabetes, 1 - prediabetic, and 2 - diabetic. By combining prediabetic and diabetic, we made the Diabetes_012 a binary column. Additionally, we scaled our numeric features: BMI and Physical Health. Although our

data was clean, our dataset remained imbalanced as seen below:

```
#check for imbalabce
df_final.Diabetes_012.value_counts()
```

```
Diabetes_012
0.0    213703
1.0     39977
Name: count, dtype: int64
```

After running some ML experiments we also decided to drop 10 features that were not highly correlated to our target, in an effort to reduce noise and increase efficiency.

In Tableau, we began by combining Pre-Diabetic and Diabetic values so we could simplify visualizations and draw more meaningful observations between Diabetic and Non-Diabetic Individuals. We used the '.replace()' method to accomplish this and after consulting the codebook provided by the CDC, we used this same method for fixing the column values in the Education column. Our dataset originally included 6 values in the Education column ranging from some grade-school completed to graduating college, we simplified this to 4 values: 'some high school', 'graduated high school', 'some college' and 'graduated college'. We originally encountered an error when converting the column values so to fix this we used '.fillna()', and converted columns from float to int types with the exception of the BMI column. After successfully converting the rest of the columns to integers, we used dictionaries to individually map out a column's binary values to string values that would translate better in Tableau. We then used the '.replace()' method again to apply the mappings and finish the column values' conversion. Lastly, we converted some of the column names either for readability (Ex. converting 'Diabetes_012' to 'Diabetes Status') or to add more context (Ex. 'CholCheck' to 'Chol. Check, 5 years'). We exported the dataframe to a new CSV and specified that this was for Tableau use and not our original dataset.

Machine Learning

We began with a notebook dedicated to data preprocessing and model experimentation. We labeled ['Diabetes_012'] as our binary target variable (pre-diabetic/diabetic or non-diabetic).

Numeric features:

- BMI
- PhysHlth

Categorical features:

- GenHlth
- Age

Binary Features:

- HighBP
- HighChol
- CholCheck
- Stroke
- HeartDiseaseorAttack
- HvyAlcoholConsump
- DiffWalk
- Sex

We used StandardScaler() to normalize the numeric features (in other words standardize the data format). We combined the scaled data with our categorical and binary features to create a final dataframe. The final dataframe and a subset of the dataframe were split into testing and training groups for machine learning experimentation. We ran multiple classifications along with their corresponding confusion matrix and classification report.

jupyter FocusedML_analysis Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help

Code

[4]:
```
binary_features = ['Diabetes_012', 'HighBP', 'HighChol', 'CholCheck', 'Stroke',
num_features = ['BMI', 'PhysHlth']
cat_features = ['GenHlth', 'Age']
```

[5]:
```
df_num = df.loc[:, num_features]
df_num.describe()
```

[5]:

	BMI	PhysHlth
count	253680.000000	253680.000000
mean	28.382364	4.242081
std	6.608694	8.717951
min	12.000000	0.000000
25%	24.000000	0.000000
50%	27.000000	0.000000
75%	31.000000	3.000000
max	98.000000	30.000000

[6]:
```
#Scaler
#initialize
scaler = StandardScaler()

#fit
scaler.fit(df_num)

#predict/transform
scaled_data = scaler.transform(df_num)
df_scaled = pd.DataFrame(scaled_data, columns=num_features)

df_scaled.head()
```

[6]:

	BMI	PhysHlth
0	1.757936	1.233999
1	-0.511806	-0.486592
2	-0.057858	2.954590
3	-0.209174	-0.486592
4	-0.663122	-0.486592

[8]:
```
#put everything together:
df_clean = df.loc[:, binary_features]
df_clean = pd.concat([df_clean, df_scaled], axis=1)
df_clean = pd.concat([df_clean, df_cat], axis=1)
df_clean.head()
```

[8]:

	Diabetes_012	HighBP	HighChol	CholCheck	Stroke	HeartDiseaseorAttack	HvyAlcoholConsump	DiffWalk	Sex	BMI	PhysHlth	GenHlth	Age
0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.757936	1.233999	5.0	9.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.511806	-0.486592	3.0	7.0
2	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	-0.057858	2.954590	5.0	9.0
3	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	-0.209174	-0.486592	2.0	11.0
4	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	-0.663122	-0.486592	2.0	11.0

[9]:
```
df_clean['Diabetes_012'] = df_clean['Diabetes_012'].replace(2.0, 1.0)

df_final = df_clean.iloc[:, [0] + list(range(1, df_clean.shape[1]))]

df_final.head()
```

[9]:

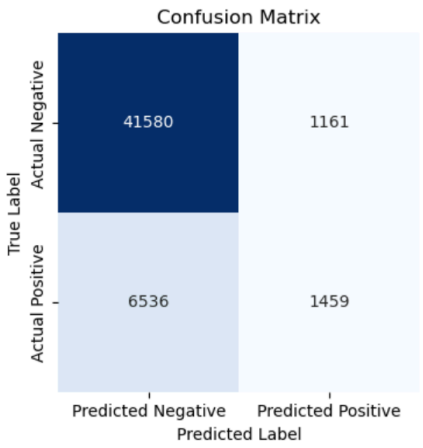
	Diabetes_012	HighBP	HighChol	CholCheck	Stroke	HeartDiseaseorAttack	HvyAlcoholConsump	DiffWalk	Sex	BMI	PhysHlth	GenHlth	Age
0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.757936	1.233999	5.0	9.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.511806	-0.486592	3.0	7.0
2	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	-0.057858	2.954590	5.0	9.0
3	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	-0.209174	-0.486592	2.0	11.0
4	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	-0.663122	-0.486592	2.0	11.0

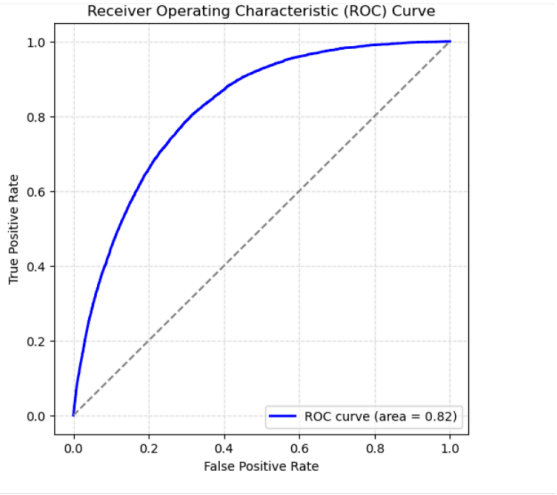
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1, s
```

```
[20]: # Function for Classification
# THE ROC CURVE CODE WILL BREAK FOR MULTI_CLASS PROBLEMS
def doClassification(model, X_train, X_test, y_train, y_test):
    # Step 3: Fit the model
    model.fit(X_train, y_train)

    # Step 4: Evaluate the model
    train_preds = model.predict(X_train)
    test_preds = model.predict(X_test)

    train_proba = model.predict_proba(X_train)[:, 1]
    test_proba = model.predict_proba(X_test)[:, 1]

    # Generate metrics TRAIN
    train_cf = confusion_matrix(y_train, train_preds)
    train_cr = classification_report(y_train, train_preds)
    train_auc = roc_auc_score(y_train, train_proba)

    train_results = f"""TRAIN METRICS
    Confusion Matrix:
    {train_cf}

    AUC: {train_auc}
```

```
[21]: # Step 2: Init the Model
lr = LogisticRegression()

# Do Machine Learning
doClassification(lr, X_train, X_test, y_train, y_test)
```

```
TRAIN METRICS
    Confusion Matrix:
    [[166304   4658]
 [ 26281   5701]]

    AUC: 0.8148835683776046

    Classification Report:
                  precision    recall  f1-score   support

         0.0       0.86      0.97      0.91    170962
         1.0       0.55      0.18      0.27     31982

    accuracy                           0.85    202944
   macro avg       0.71      0.58      0.59    202944
weighted avg       0.81      0.85      0.81    202944
```





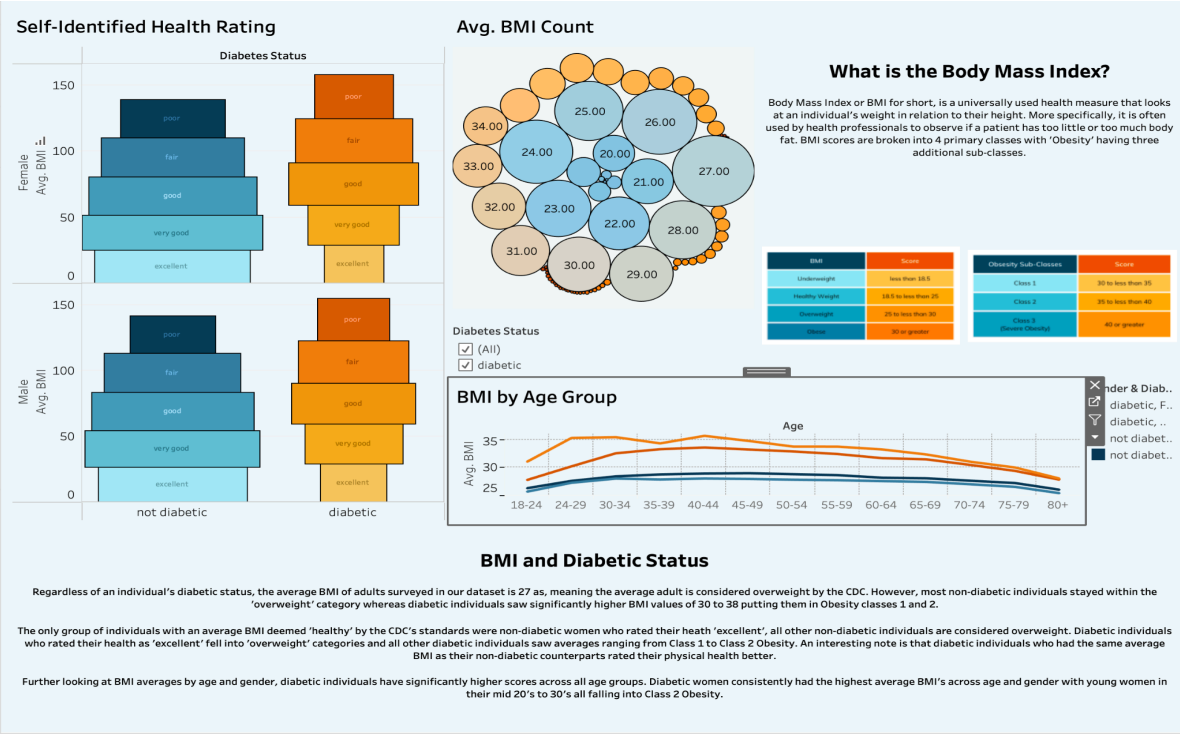
We settled on an XGB model as the best fit model however needed to pivot to using an ADA model because the XGB model was not compatible when deploying our model with PythonAnywhere. Finally, we pickled off our model and our scaler to be used in out Inference notebook. This allowed us to test our model's predictions for when we inserted it into our web app.

Tableau

Dashboard 1: Exploring the Relationship Between Diabetes and BMI

How do health conditions relate to diabetes status and management?

We focused on the Body Mass Index (BMI) and how an individual's BMI score and subsequent weight category relate to their gender, age, and diabetic status.



The infographic explores the relationship between diabetic status and BMI scores by looking at factors such as age, gender and how individuals rated their own physical health.

- Avg. BMI Count: Overall average BMI is 27; non-diabetic trends to 'overweight' (BMI over 25, less than 30). While diabetic trends to 'obese' (BMI 30+).
- BMI by Age: Diabetic individuals have significantly higher BMI averages, specifically women aged mid 20's to 30's. There was not a huge difference between non-diabetic men and women's average BMI scores.
- Self-ID Health Rating: Non-diabetic individuals most likely to rate health 'very good' or 'good' and diabetic individuals most likely to rate health 'good' or 'fair'. Non-diabetic women with 'excellent' health were the only group with healthy BMI. All diabetic groups had 'obese' BMI averages except for those who rated their health 'excellent'.

Average BMI Count

Regardless of diabetic status, the average BMI of our dataset was 27. More plainly, the average BMI of individuals surveyed for our dataset would qualify as overweight. The CDC considers any BMI score in the 25 to 30 range 'overweight'. When we filtered to only look at non-diabetic individuals, the majority stayed within the 'overweight' range. While diabetic individuals still had significant counts in the 'overweight' BMI range, there were much higher counts of BMI scores ranging from 30 to 38 which would fall under the Class 1 and 2 'obesity' ranges.

Individual Health Assessment in Relation to Gender, Diabetes Status, and BMI

When we segmented individuals in our dataset by their gender and diabetic status and looked at how they rated their own physical health, the only group that had a 'healthy' BMI score was women who rated their health 'excellent'. All other non-diabetic groups had 'overweight' BMI averages and all diabetic groups had 'obese' BMI averages with the exception of diabetic men and women who rated their health 'excellent'; their BMI averages fell into the 'overweight' category. Another potentially significant observation is the fact that non-diabetic individuals were more likely to rate their physical health *worse* than diabetic individuals that shared the same average BMI score.

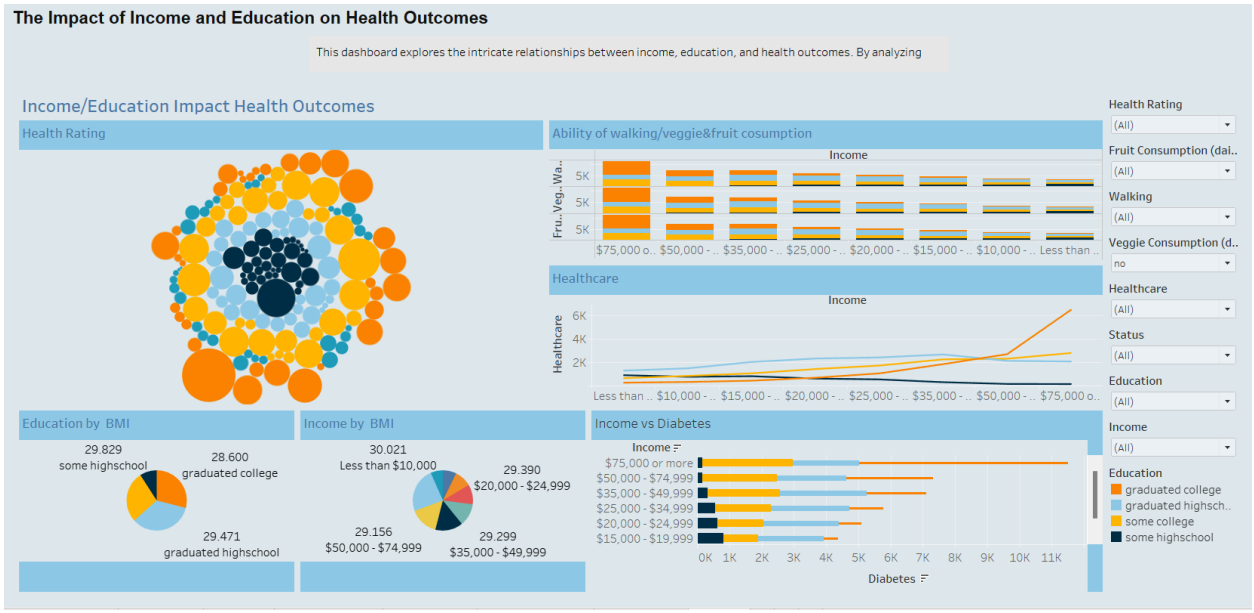
BMI by Age Group

Diabetic women consistently held the highest BMI averages over all other groups. Furthermore, the difference between non-diabetic and diabetic women's BMI averages is greater than that non-diabetic and diabetic men. The highest BMI averages were for women in their mid 20's to 30's.

Dashboard 2: The Impact of Income and Education on Health Outcomes

What is the impact of socioeconomic factors (education and income) on diabetes health indicators(Health Rating, Healthcare, Income)?

How do lifestyle factors (such as physical activity and diet) correlate with diabetes prevalence across different demographics?



This dashboard explores the intricate relationships between income, education, and health outcomes. By analyzing various visualizations, we can uncover key insights into how these factors interplay to influence overall health.

- Income and Health: Higher income is consistently linked to better health ratings and lower prevalence of diabetes.
- Education and Health: Higher education levels correlate with better health ratings and healthier dietary habits.
- Dietary Habits: Consumption of fruits/veggies and ability to walk varies across income levels, providing insights into lifestyle choices.

Income and Education Effects on Health Rating

The bubble chart reveals a clear trend: Higher income and education levels are associated with better health ratings. Larger and more intense bubbles indicate a higher number of individuals within each income bracket who report better health, underscoring the positive impact of socioeconomic status on health.

Consumption of Veggies/Fruits and Ability to Walk by Income Level

The bar graph shows the frequency of consuming fruits/veggies and ability to walk across different income levels. This visualization highlights dietary habits and suggests that income influences healthy choices, which can have broader implications for health.

Health Care Correlation with Income and Education

The line graph illustrates the correlation between healthcare and both income and education levels. It shows that as income and education increase, having a health care plan improves, reinforcing the importance of socioeconomic factors in determining health outcomes.

Income Levels and Diabetes

The bar graph comparing income levels with diabetes prevalence reveals that lower income groups have higher rates of diabetes. This highlights the health disparities that exist across different income brackets and the need for targeted interventions to address these inequalities.

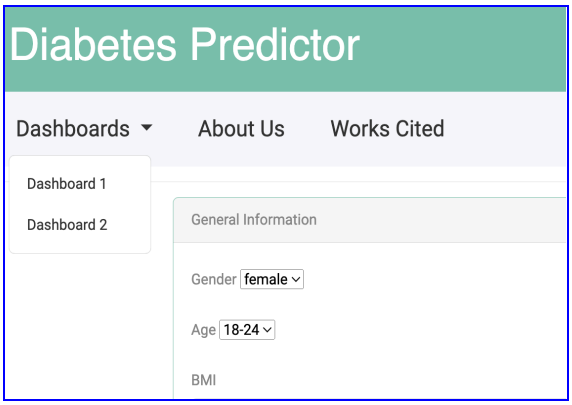
- Income and Health: Higher income is consistently linked to better health ratings and lower prevalence of diabetes.
- Education and Health: Higher education levels correlate with better health ratings and healthier dietary habits.
- Dietary Habits: Consumption of fruits/veggies and ability to walk varies across income levels, providing insights into lifestyle choices.

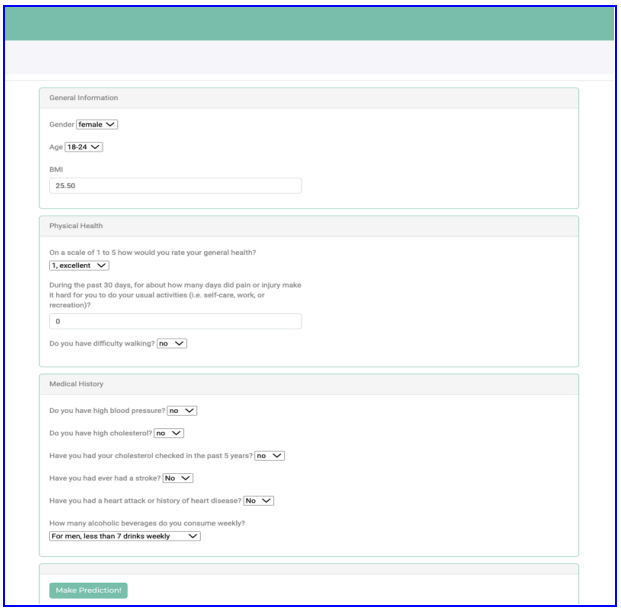
This story emphasizes the critical role of income and education in shaping health outcomes. It provides a comprehensive view of how socioeconomic factors influence health, offering valuable insights for public health strategies and policies.

Web App

Front-end development consisted of building five HTML pages and a supporting CSS stylesheet.

Home Page





The home page serves as an interactive Diabetes Predictor which allows users to input and select different metrics such as medical history, physical health, gender, age, and BMI to utilize our trained ML model to make a prediction on whether an individual is diabetic or not. The home page consists of a simple and easy form, a clear navigation bar, and a robust backend functionality that ensures that users have a seamless and informative experience. This tool can be a valuable resource for individuals looking to assess their risk of diabetes and take proactive steps towards managing their health.

Navigation Bar

The top of the page features a navigation bar with tabs to our other webpages and allows users to explore different sections of the website seamlessly. A dropdown menu was added for the ‘Dashboard’ tab as it looked better visually.

User Input Form

The primary purpose of the homepage is to host our Diabetes prediction form where users can input their health information. The form includes various fields such as:

- Gender: dropdown for selecting gender
- Age: dropdown for selecting age range

- BMI: textbox for entering a specific BMI score
- General Health: dropdown for selecting physical health rating from 1 - 5
- Physical Health: dropdown for selecting number of poor physical health days
- Difficulty Walking: dropdown for selecting if walking is difficult or not
- High Blood Pressure: dropdown for selecting high blood pressure or not
- High Cholesterol: dropdown for selecting high cholesterol or not
- Cholesterol Check: dropdown for selecting if cholesterol has been checked in past 5 years or not
- Previous Stroke: dropdown for selecting stroke history
- Heart Attack or Disease: dropdown for selecting heart attack/disease history
- Alcohol Consumption: dropdown for selecting weekly alcohol consumption

Prediction Button

At the bottom of the form, there is a prominent button labeled “Predict Diabetes”. Once the user fills in their information and clicks this button, the backend processes the data using our selected machine learning model to generate a prediction.

Backend Functionality

The backend is built with JavaScript and CSS files, along with an `index.html` file. When the user submits their information, the backend processes the data and runs the prediction algorithm.

Interactive and Responsive Design

The web page is designed to be interactive and responsive, ensuring a smooth user experience across different devices. The use of JavaScript and CSS enhances the functionality and aesthetics of the page.

About Us Page

About Us

We are a team of passionate data enthusiasts committed to gaining a deeper understanding of Type 2 Diabetes through observing individuals' behavior and health trends.

Our Mission

Our mission is to raise awareness about the prevalence of Type 2 Diabetes. By thoroughly analyzing our dataset and training a machine learning model, our goal is to identify both positive and negative behaviors and health trends that help individuals understand how to prevent or manage Type 2 Diabetes.

The Team


Sierra Sarkis


Hannah Lashway


Neyda Morales


Thripura Pakala

Data Sources

The data used in this project is sourced from Kaggle. We are committed to using reliable and up-to-date information to ensure the accuracy of our visualizations.

Contact Us

If you have any questions contact us in the DATA-PT-EAST-APRIL-041524 slack workspace.

Works Cited Page

Works Cited

The data used in this project is sourced from Kaggle. This dataset takes data from a 2015 survey conducted by the Behavioural Risk Factor Surveillance System, a health survey system within the CDC. Throughout our data cleaning and analysis, we utilized the guidebook provided by CDC to better understand the original data as many of our columns were binary.

We further referenced the CDC's website to gain more insight into BMI metrics.

Biases and Limitations

Biases

- Our dataset was imbalanced in a few ways; even when combining the pre-diabetic and diabetic values, our dataset was still heavily composed of non-diabetic individuals. Additionally, our dataset was economically imbalanced with the majority of individuals making over $75,000.
- Data was collected from a 2015 survey, making it almost 10 years old and data may not be entirely reflective of current Diabetes and health trends.

Limitations

- The dataset we used was already a clean version of the original survey dataset, because of this there were many columns such as pregnancy and ethnicity that were not included and could have provided additional important insights.
- We did not have any time-frame information. This made it impossible to observe trends in behavior changes in relation to a pre-diabetic, diabetic or heart disease diagnosis. More specifically, there was no way to observe if individuals who had a serious health flag such as a heart attack or diabetes diagnosis began to make positive health choices like physical activity after a diagnosis.

Future Work

Addressing Imbalanced Datasets: Future work will focus on tackling the issue of imbalanced datasets to ensure more accurate and reliable model predictions.

Implementing Pipelines in Inference Notebooks: We plan to incorporate pipelines in our inference notebooks to streamline the workflow and enhance reproducibility.

Exploring Alternative Models: We will experiment with alternative models, such as the Balanced Random Forest, to improve performance and robustness.

Reducing Dataset Noise: Efforts will be made to better tailor the dataset, aiming to reduce noise and enhance the quality of the input data.

Conclusion

Overall Findings

- Our study highlights the importance of monitoring BMI as a critical factor in managing diabetes and overall health.
- BMI generally increases with age, with diabetic individuals showing higher average BMIs compared to non-diabetic individuals across all age groups.
- Diabetic individuals consistently have higher BMIs, which correlates with poorer self-identified health ratings.
- Non-diabetic individuals with high BMIs tend to rate their health better than diabetic individuals with similar BMIs, suggesting that diabetes significantly impacts perceived health quality.

Machine Learning Integration

- The application of machine learning models allows for more accurate predictions and deeper insights into the relationships between BMI, age, health status, and diabetic status.
- These models can identify patterns and trends that may not be immediately apparent through traditional analysis, providing valuable information for public health strategies and personalized healthcare interventions.

Web Page Implementation

- The creation of interactive web pages based on these dashboards can enable users to explore the data dynamically, offering a more engaging and informative experience.
- Users can filter and drill down into specific data points, advancing their understanding of how various factors influence health outcomes.

Final Reflection

The integration of machine learning models with comprehensive data visualization tools like Tableau provides a powerful approach to analyzing health data. This combination not only improves our ability to predict health outcomes but also supports the

development of targeted interventions to improve public health. By making these insights accessible through interactive web pages, we can empower users to make informed decisions about their health and well-being.