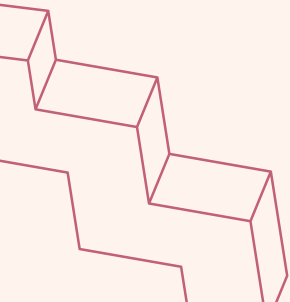


Practicum 2


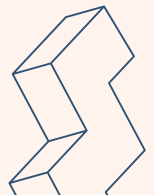
Daniel Sobel, Lauren Foster, Minh Tran





Problem Statement

We want to find the optimal lifestyle and social determinants of health that reduce the chances of having diabetes.

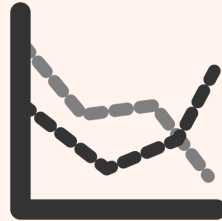


Goals & Objectives

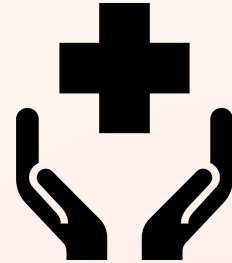
Analyze the complex data set to identify patterns and relationships between lifestyle and demographic factors



Create prediction models to determine diabetes risk based on these variables.



Understanding these the model and its outcomes can help shape public health policies and individual behavior changes.



Describing the dataset

Target variable: Diabetes



Types of factors:

- Demographics
 - Age
 - Income
- Patient's own experiences
 - Mental health score
 - Couldn't seek a doctor due to cost
- Behavioral
 - Physical activity
 - Eating fruits or vegetables
 - Alcohol consumption
- Health history
 - Stroke
 - Heart attack or disease

Looking for patterns

The proportion of individuals with diabetes is lowest for individuals...

With lower BMIs

In lower age groups

Who have a higher income

Do not have difficulty walking

Who gave themselves better general health scores

With better mental health

Without high blood pressure

Who have not had a stroke

Who have not had a heart attack or heart disease

Who do not smoke

Who have done physical activity within the
past 30 days

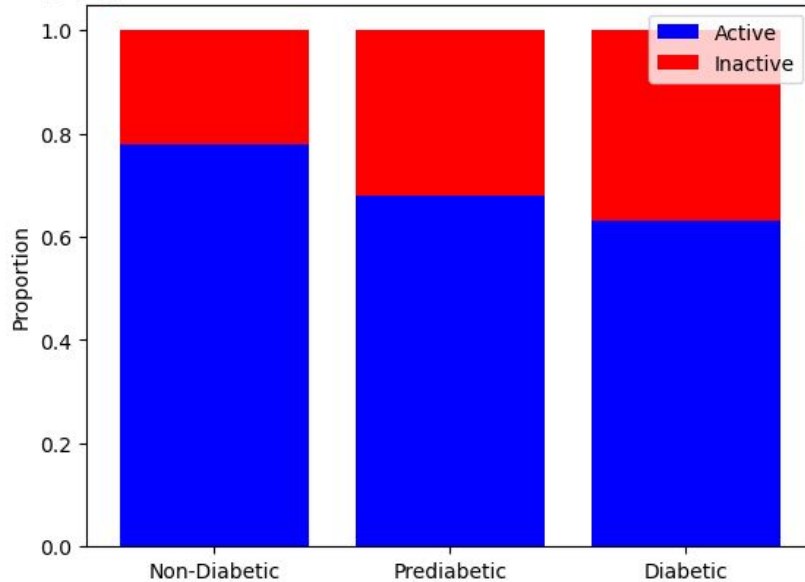
Who eat fruits and vegetables daily

Behaviors

|

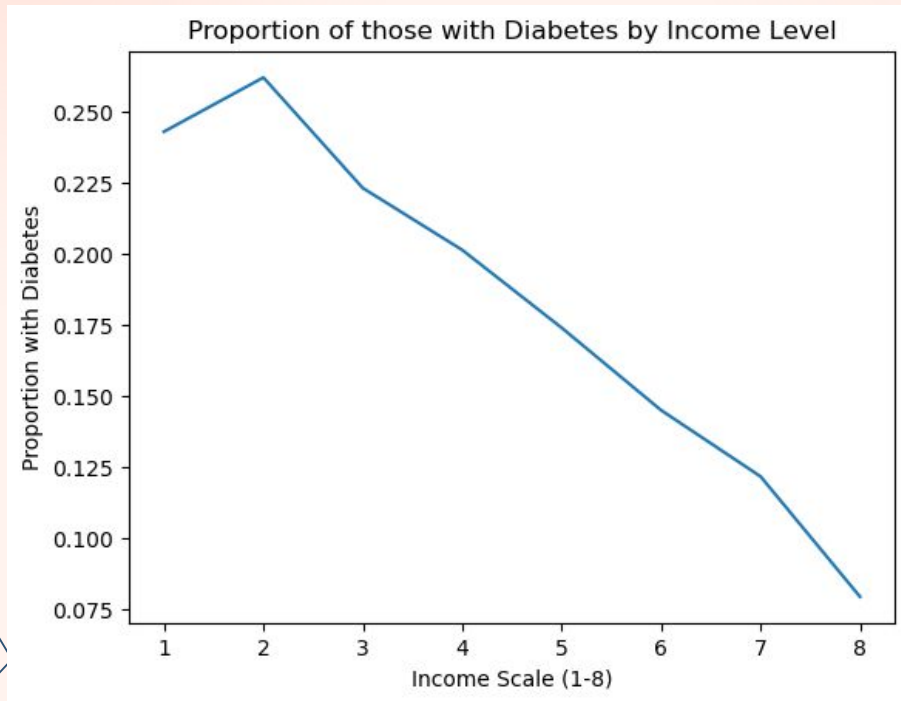
Comparing proportion of active and inactive patients among different groups of diabetes

Comparing Proportion of Active Patients in Diabetics, Prediabetics, and Non-Diabetics



- Criteria for active: completed physical activity apart from their job within the past 30 days
- Observations: As the level of diabetic diagnosis becomes more serious, the proportion of inactive patients increases

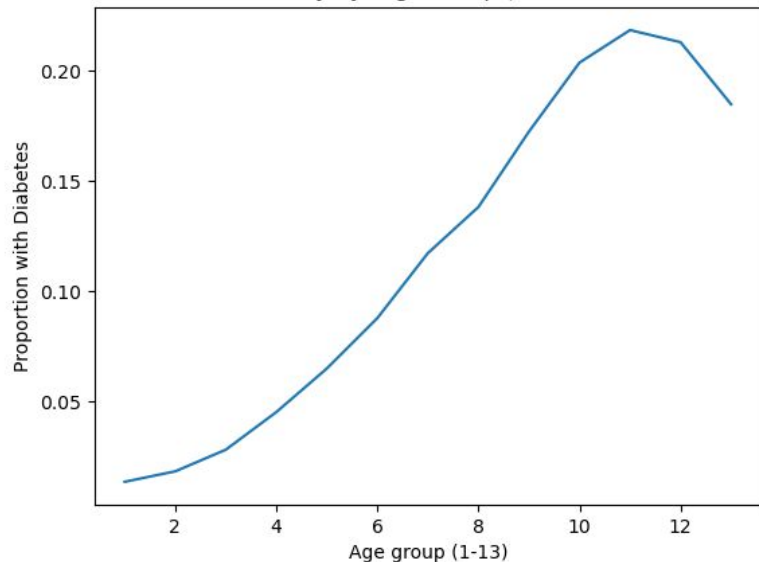
Examining proportion of those with diabetes among different income levels



- Income levels: 1 represents an income less than \$10k while 8 represents an income greater or equal to \$75k
- Highest proportion of diabetes is at income level 2
- Lowest proportion of diabetes is at income level 8
- Negative linear relationship between income and diabetes.

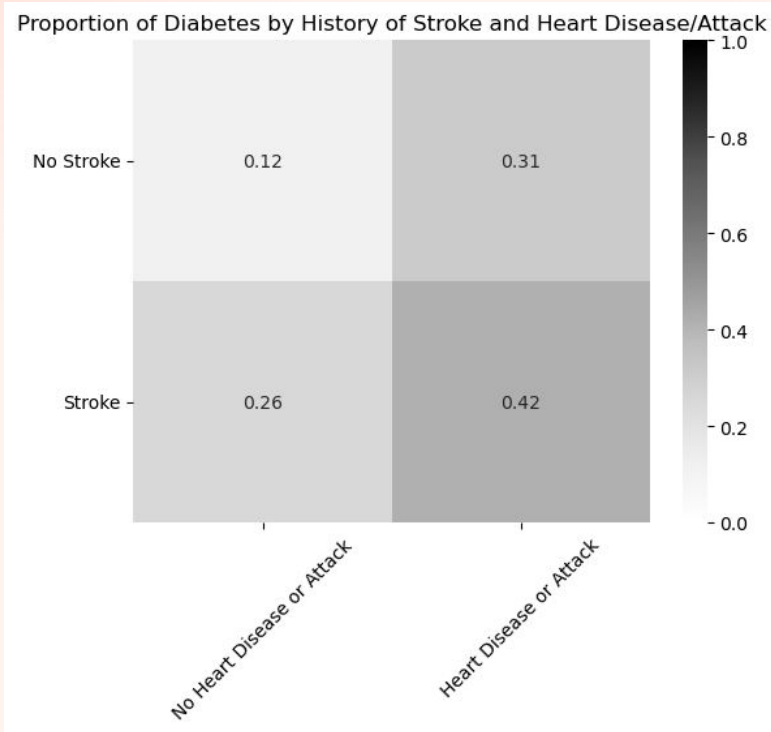
Examining proportion of those with diabetes among different age groups

Proportion of Adults with Diabetes by 5yr Age Group (1 = 18-24 9 = 60-64 13 = 80 or older)



- Age groups: 1 represents adults aged 18-24 while 13 represents 80 or older
- Highest proportion of diabetes is in age group 11 (aged 70-74)
- Lowest proportion of diabetes is in age group 1 (18-24)
- Overall positive relationship between age and diabetes. Proportion decreases after age group 11

Proportion of those with diabetes among those with different health histories



- Looking at all 4 combinations: having had none, one, or both a stroke and heart attack
- Of those in each category, the proportion of those with diabetes is highest for those who have had both a stroke and either heart disease or a heart attack
- Proportion is lowest for those without either medical issue

Predicting diabetes with the best ML model

Data Split: 80% training, 20% testing.

Preprocessing Pipelines:

- **Numerical Features:**
StandardScaler for 'BMI'.
- **Categorical Features:**
OneHotEncoder for 'Sex', 'Age', 'Education', and 'Income'.

- **Models Used:**
 - RandomForest
 - SVM
 - KNN
- **Hyperparameters Tuned:**
 - RandomForest:
n_estimators, max_depth
 - SVM: C, gamma
 - KNN: n_neighbors, weights
- **GridSearchCV:** Used for tuning hyperparameters.

RandomForest Model Results

RandomForest Model

- **Best Model Parameters:**
`RandomForestClassifier(max_depth=10,
random_state=42)`
- **Best GridSearchCV Score:** 0.828
- **Performance Analysis:**
 - **Training Set Accuracy:** 85.57%
 - **Testing Set Accuracy:** 83.33%

The RandomForest model shows good performance on both training and testing sets. However, the significant difference in recall and f1-score for the minority classes (1 and 2) indicates a model bias towards the majority class (0).

This could suggest overfitting towards the majority class and underperformance on rarer outcomes, which is problematic in healthcare settings where detecting less frequent but critical conditions is essential.

SVM Model Results

SVM Model

- **Best Model Parameters:** `SVC(C=0.1, gamma=0.001, random_state=42)`
- **Best GridSearchCV Score:** 0.823
- **Performance Analysis:**
 - **Training Set Accuracy:** 82.31%
 - **Testing Set Accuracy:** 82.29%

The SVM model demonstrates a similar performance on training and testing sets but severely underperforms in predicting minority classes (1 and 2), with both precision and recall being zero. This suggests that the model is entirely biased towards predicting the majority class, ignoring the less frequent ones.

KNN Model Results

KNN Model

- **Best Model Parameters:**
`KNeighborsClassifier()`
- **Best GridSearchCV Score:** 0.812
- **Performance Analysis:**
 - **Training Set Accuracy:** 85.53%
 - **Testing Set Accuracy:** 81.28%

The KNN model shows decent accuracy but like the other models, it performs poorly on minority classes, indicating an inability to generalize well across different health conditions.

Model Quality Analysis

Bias-Variance Tradeoff:

- The RandomForest and KNN models show a slight overfitting tendency with better performance on the training set compared to the testing set.
- The SVM model, while stable across both sets, fails to capture the complexity needed to identify less frequent health conditions, indicating high bias towards the majority class.

Desired Metrics:

- In healthcare data analysis, precision and recall are crucial metrics, especially for imbalanced classes representing different health conditions. High recall is particularly important to ensure that all potential health risks are identified, even if at the expense of some false positives (lower precision).
- Given the context, it might be essential to focus on improving recall for the minority classes (1 and 2) and consider metrics like the f1-score for a balanced view of precision and recall.

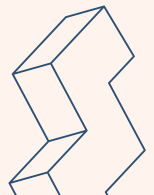



Impacts of Our Diabetes Prediction Solution

Healthcare Providers

- Encourage weight management, physical activity, and a healthy diet
- Holistic approach including mental and physical health
- Promote quitting smoking and drinking

Policy and Community Health

- Policies to provide more resources, education, and affordable healthcare to communities lacking in those areas because those demographics are most at risk
 - Health campaigns for older adults treatment and promote preventive habits for youth
- 
- 

Addressing Bias in Prediction Model

Income Bias

- Lower-income groups show a higher proportion of diabetics

Recommendation:

- Reduce income weight
- Focus on health factors more
- Regularly check for fair categorization by income

Education Bias

- Individuals with lower education levels are more likely to be classified as diabetic.

Recommendation:

- Ensure diverse education levels and accurate health impact features
- Continuously assess predictions for fairness across educational levels.

We know these factors can not directly cause diabetes, so apart from the data, we need to examine why those with low income and low education have diabetes at a disproportionately higher rate. This is a social issue that requires promoting equity and adequate resources.



**Thank
You!**

