Insight into Health Indicators for Diabetes

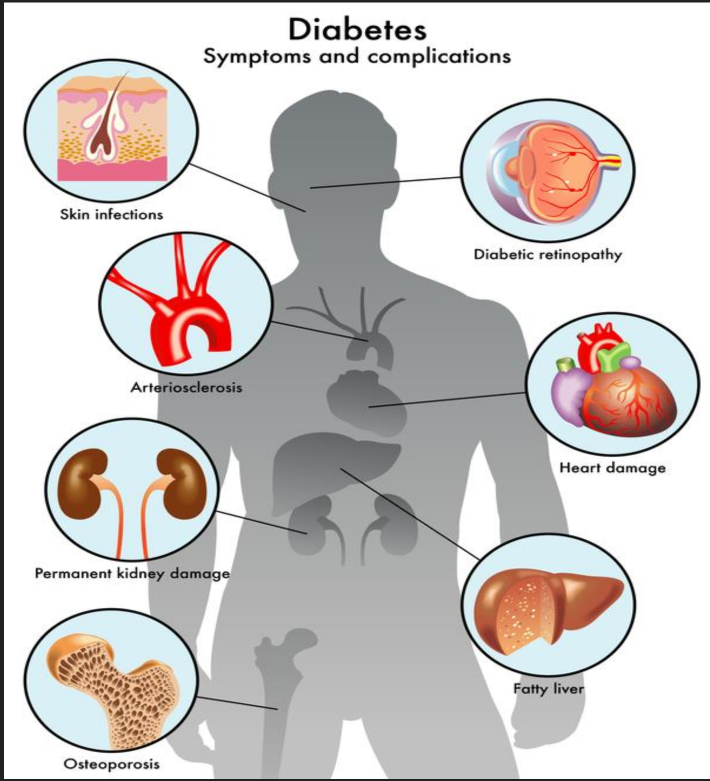
Genevieve Ferguson
Nitin Pagarani
Cristian Biondi

Mentor: Giri Narasimhan

Overview & Background

- **Diabetes** - body doesn't produce enough insulin or cannot regulate insulin produced

- **Chronic health condition** - when untreated, can affect the heart, kidneys, brain, eyes, etc.



Overview & Background

- Globally-1.5 Million deaths, 48% before age 70 (WHO, 2019)
- USA- 38 Million adults have diabetes (CDC, 2022)
- USA- Of these cases, Estimated 91% (Type 2), 6% (Type 1), 3% (Gestational & Other)
- Known **Risk Factors** include: Age, Family History, Physical activity, Diet

Diabetes (who.int)

National Diabetes Statistics Report | Diabetes | CDC

Goals & Aims

- Problem Description: Leveraging data to understand and address diabetes
- Aim: To create predictive models and use correlations to provide insight into diabetes
- Objectives:
  - Build **Classification Models**.
  - Identify risk factors.
  - Make recommendations.
- Goals:
  - Identify most salient features.
  - Find correlations between health indicators.
  - Recognize the distribution of diabetes across Florida.

Problem Motivation

- Very prevalent in US
- Causes serious health complications
- Several **risk factors** (Ex: family history, lifestyle, nutrition, etc.)
- Key to a healthy life and wellbeing
- Increase awareness and prevent the risk of developing diabetes

Data Sources

- Description: Diabetes Health Indicators and their diagnosis
  - Size: 253,680 x 22
  - Features: Diabetes_binary, Age, Income, Smoker, Education, Sex, High BP, Stroke
  - Weaknesses: **Imbalanced Data**
  - Source: UCI Machine Learning Repository

- Description: City Census Data & Health Indicators 18+
  - Size: 29006 x 24
  - Features: State, City Name, Geolocation, Data Value, Population Count
  - Weaknesses: Crude Data Values
  - Source: CDC

Methodology

- Data Preprocessing
  - Removing duplicates & null values
  - Missing value imputation
  - Inserting categorical values

- Exploratory Data Analysis:
  - Heatmaps, countplots, barplots, pointplots, US map using Folium Library

Methodology

- Balancing Dataset
  - Near Miss -Undersampling the majority class

- Selecting Relevant Features
  - Correlations
  - Random Forest (Feature Importance)
  - Select K Best (16 Best Feature Scores)
    - Feature selection method – using chi squared

Methodology

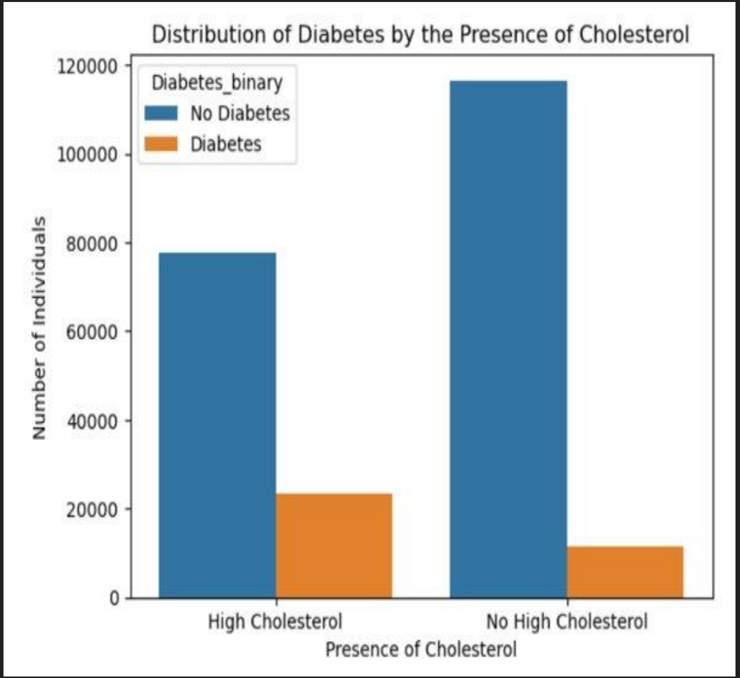
Classification Models:

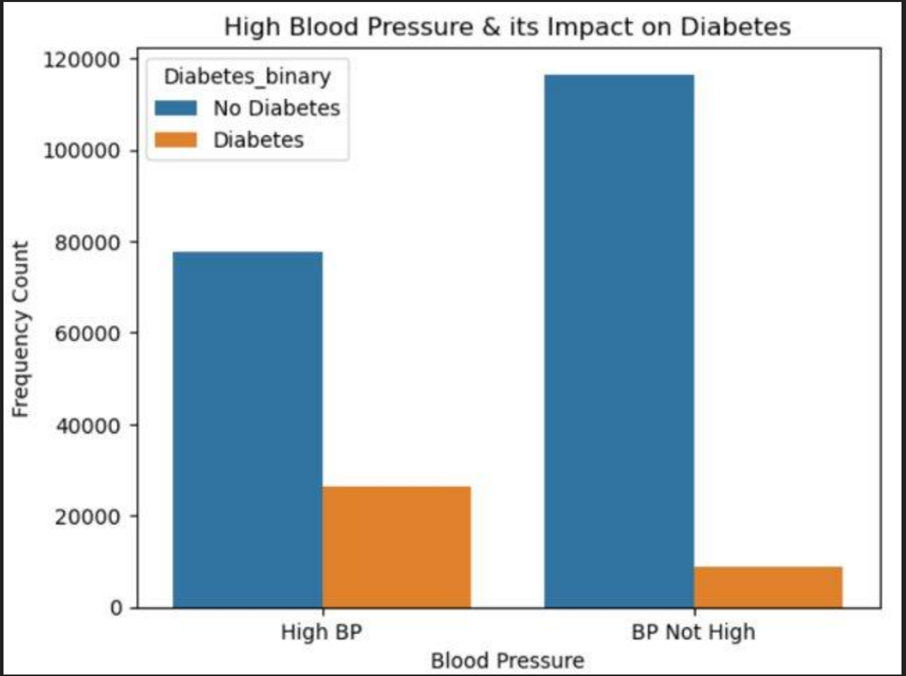
- KNN
- Random Forest
- Decision Tree
- SVM
- XGBoost
- MLP Classifier
- Ridge Classifier
- Logistic Regression
- Passive Aggressive Classifier
- ANN

Packages used:

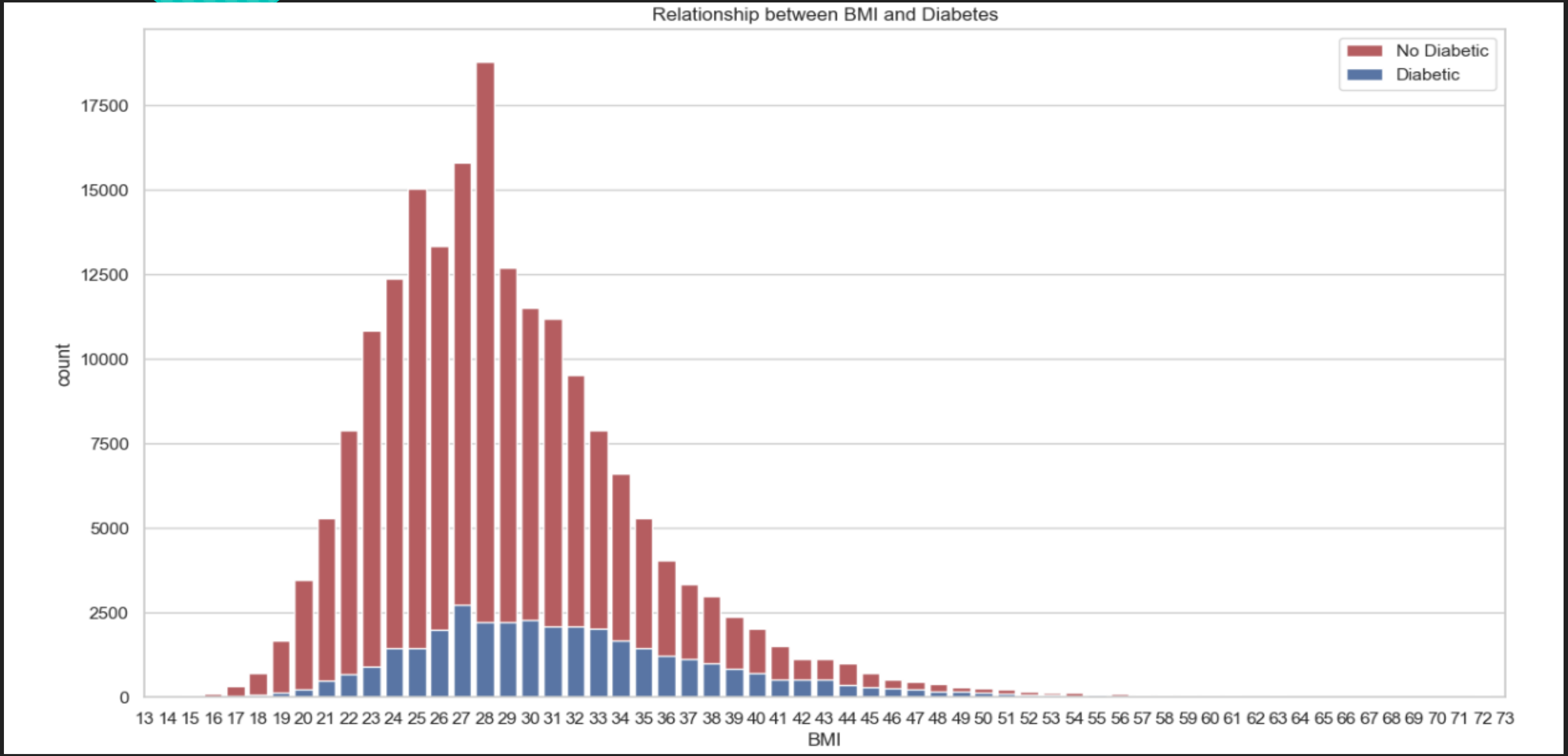
- Pandas
- NumPy
- Sklearn
- Keras
- Seaborn
- Matplotlib
- XGBoost

Results: People with diabetes are more likely to have High Cholesterol and High Blood Pressure

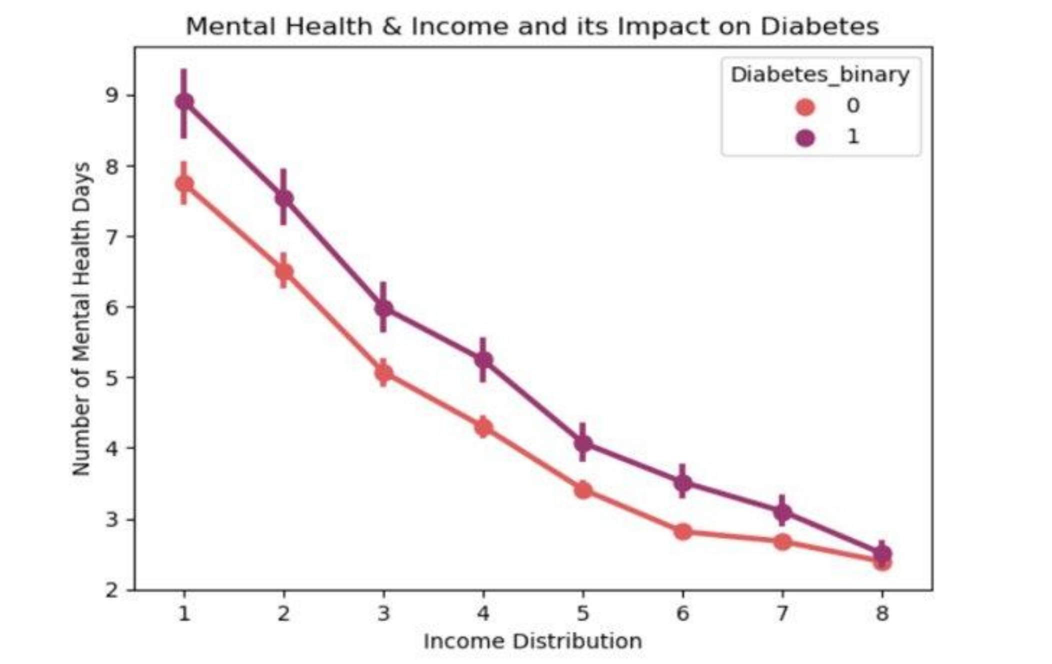




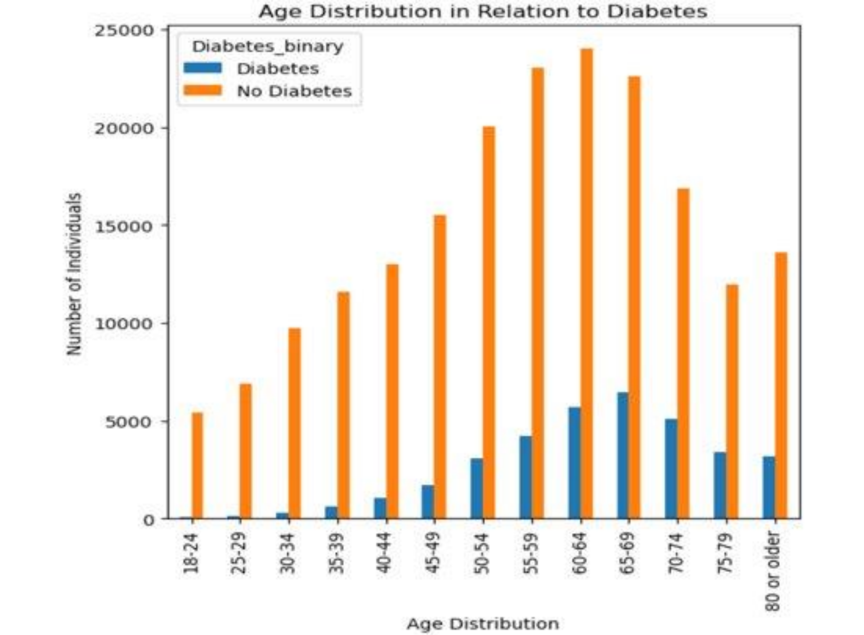
Results: Most people with diabetes have a BMI between 21-45

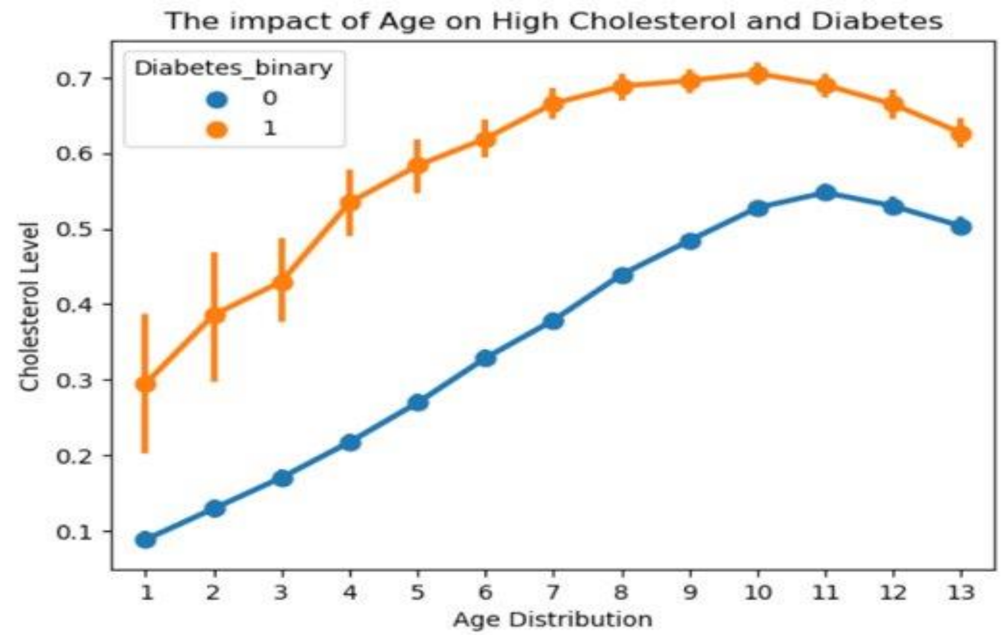


Results: People with diabetes with Lower Income have increased Mental Health Days

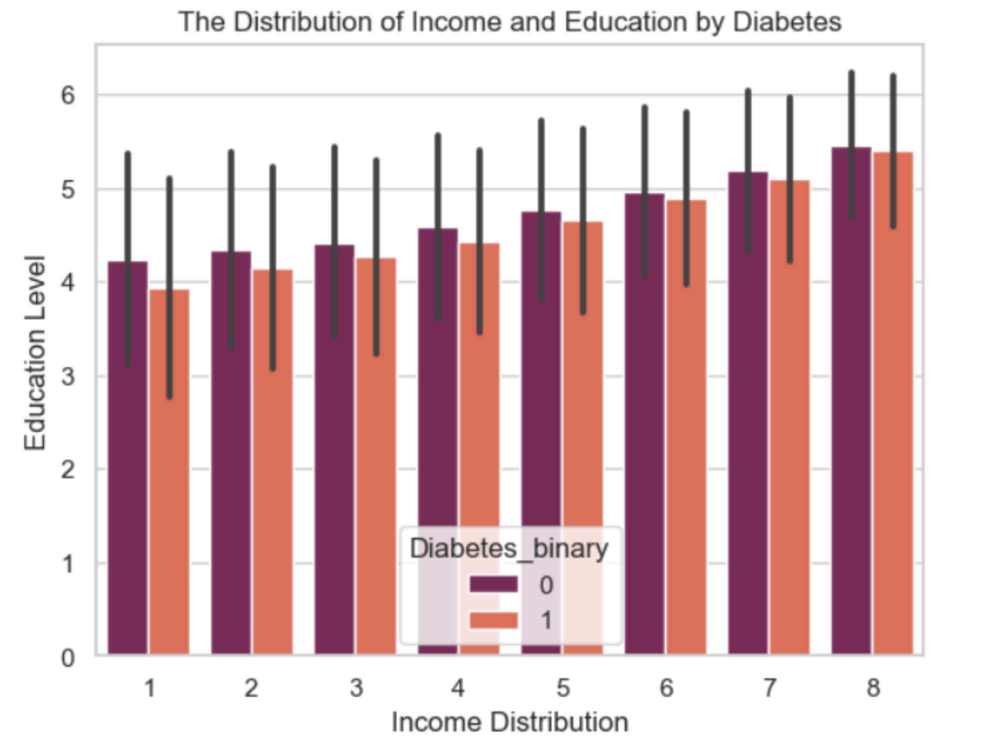


Results: People with diabetes are older in Age & have High Cholesterol

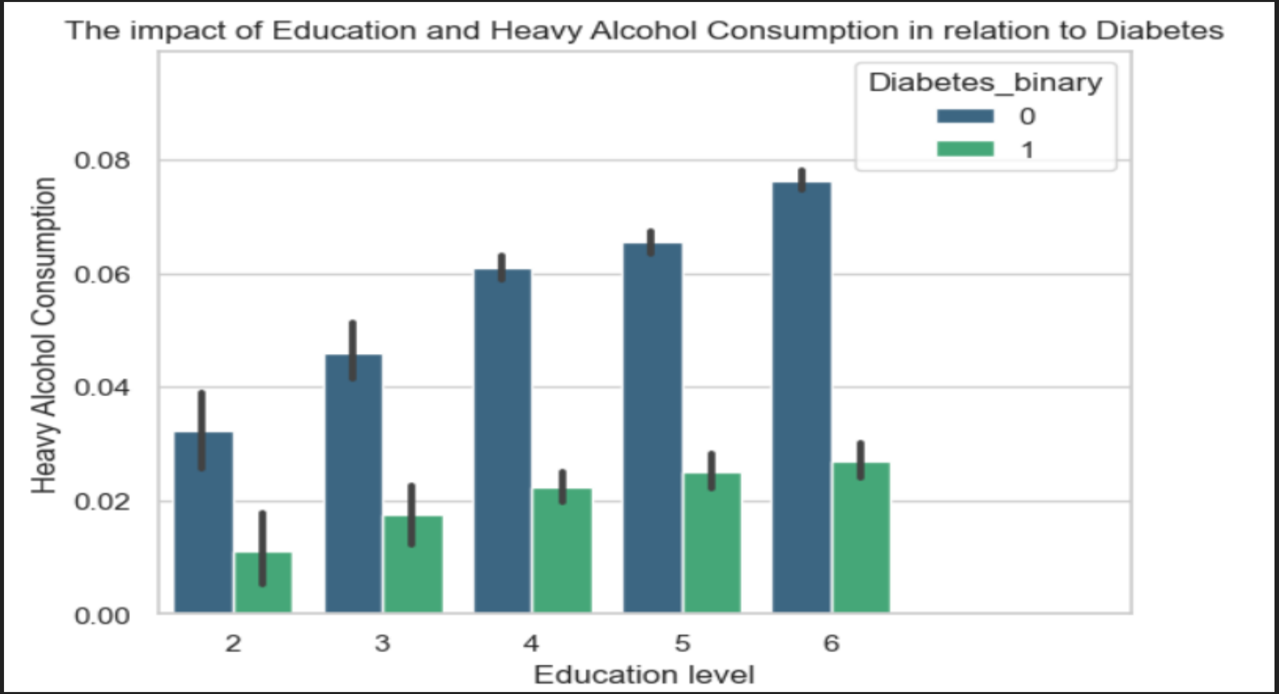




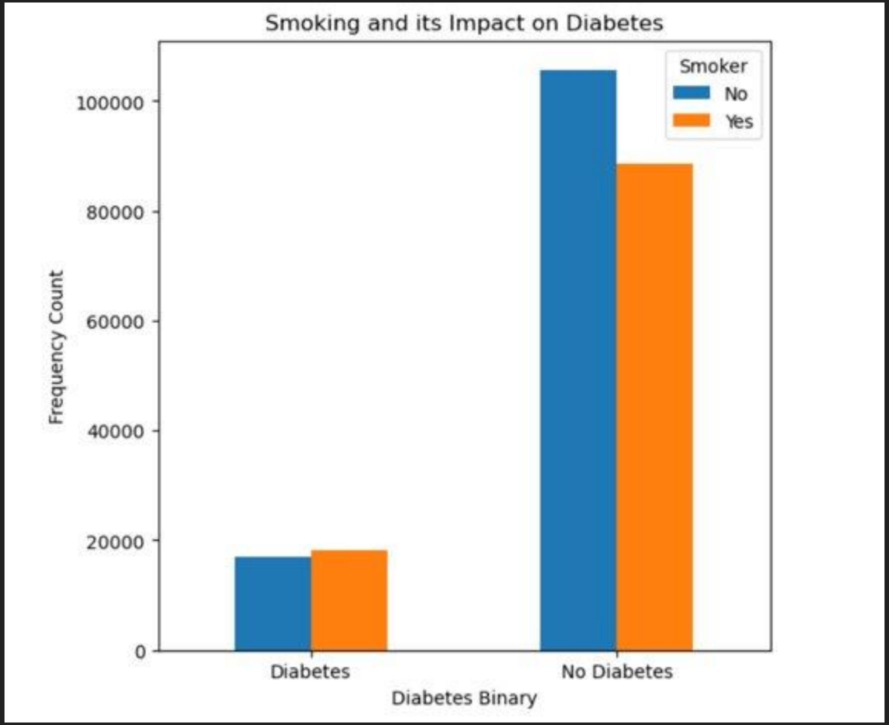
Results: People without diabetes have greater sources of income and are more educated



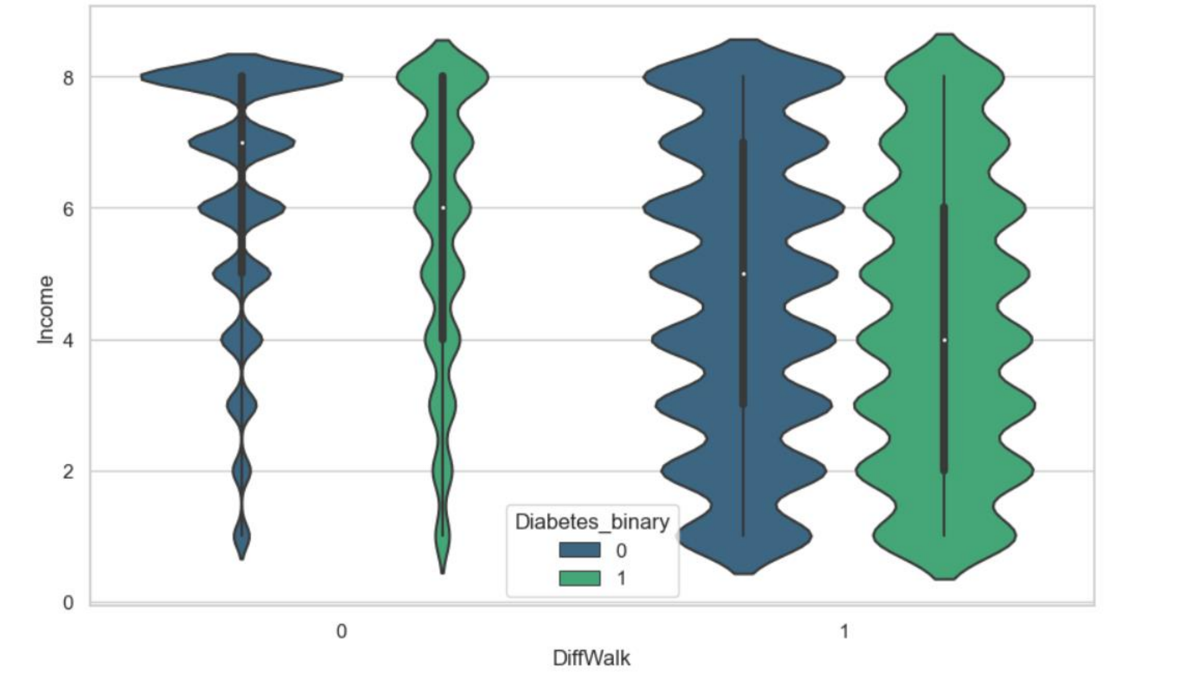
Results: Educated individuals consume more alcohol and are at risk of developing diabetes



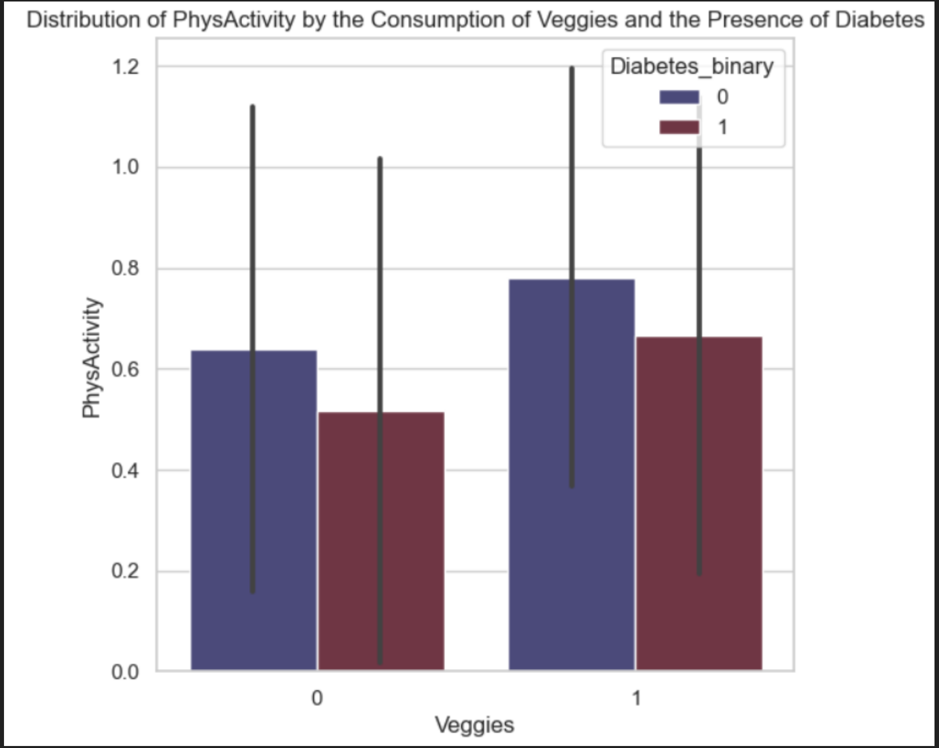
Results: Smokers are not any more likely to have diabetes than non-smokers



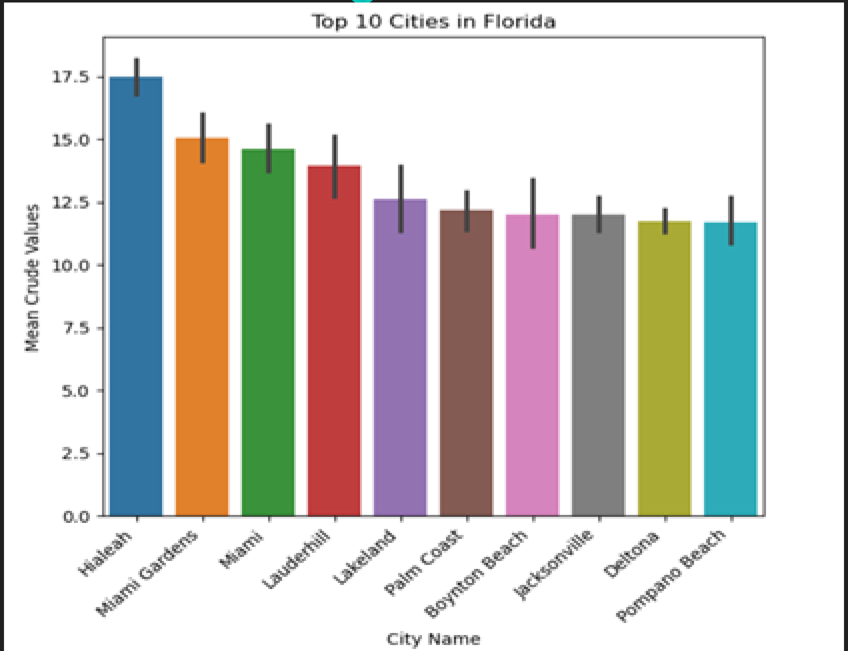
Results: People with diabetes that have difficulty walking have less income than those who don't have difficulty walking

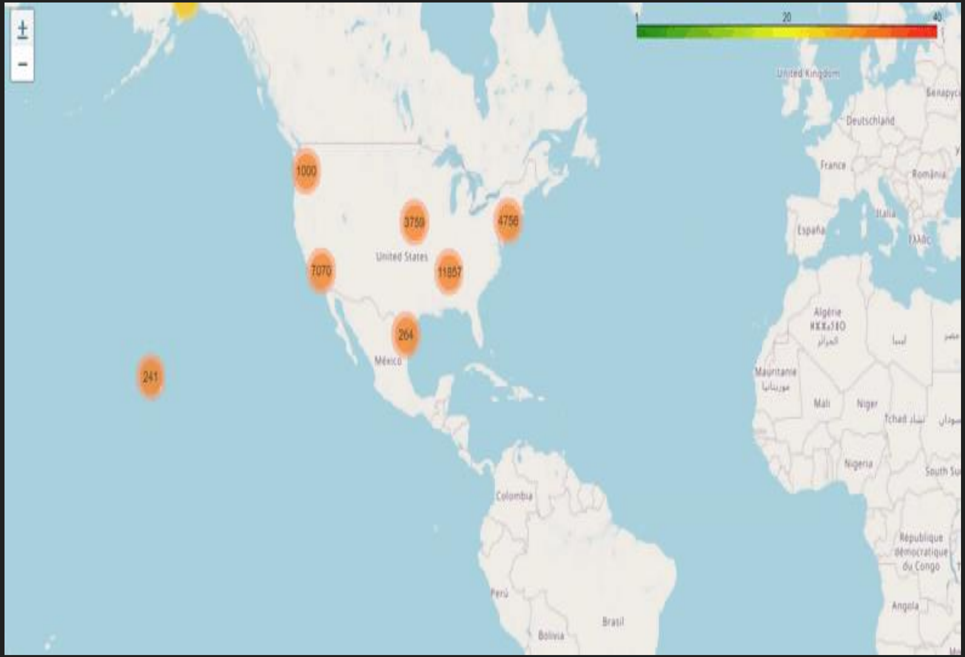


Results: People with diabetes who eat veggies tend to have higher physical activity

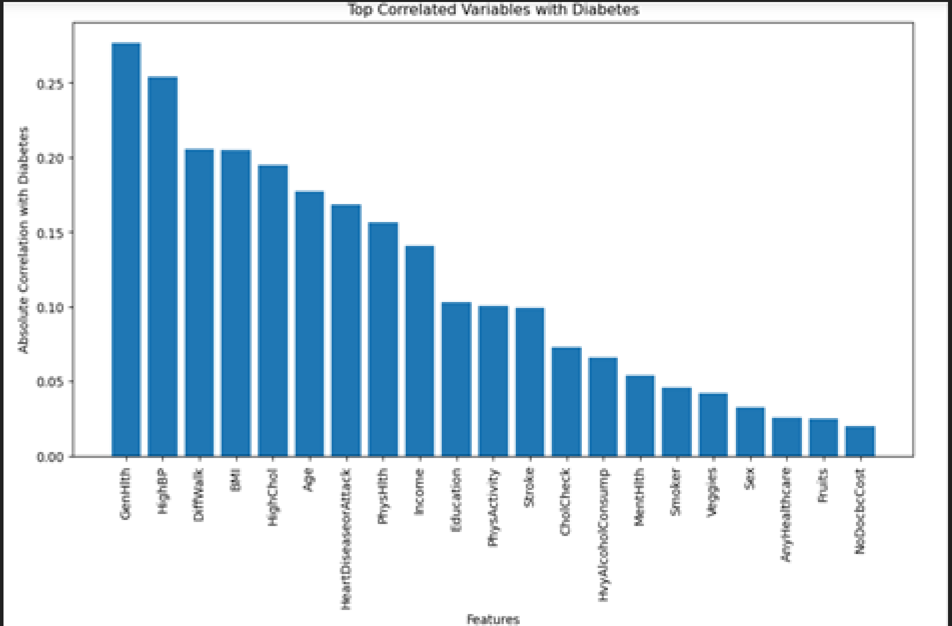


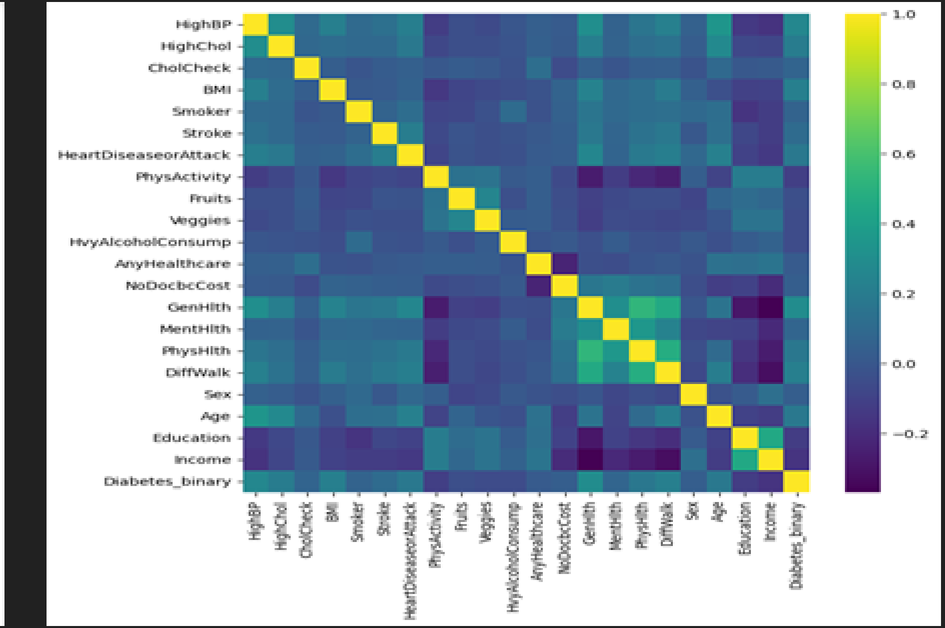
Results: City & State Analysis





Results: Correlation Analysis & Feature Selection





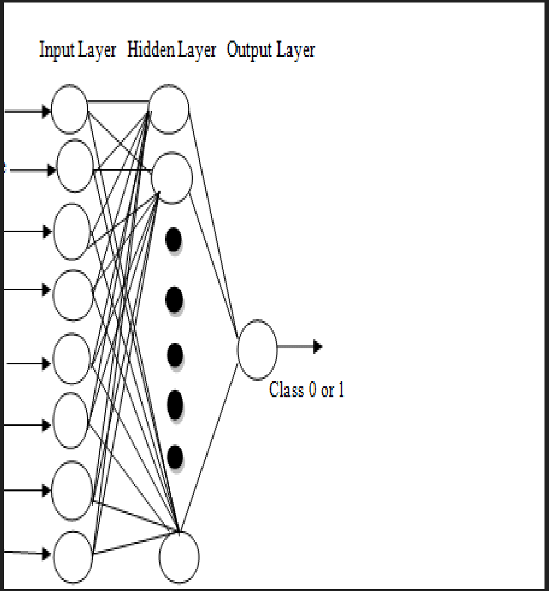
Results & Discussion: Classification Models

Predicting Diabetes/No Diabetes on a balanced Data Set

Classification Model	Accuracy
Decision Tree Classifier	80.6%
KNN	83.2%
Random Forest Classifier	86.4%
SVM	85.0%
XGBoost	85.2%
MLP Classifier	85.1%
ANN	86.5%
Ridge Classifier	84.7%
Logistic Regression	84.9%
Passive Aggressive Classifier	84.5%

Discussion: ANN Architecture

- Artificial neural network for binary classification
- Layer 1:
  - Neurons: 16
  - Activation Function: ReLU
  - Determined by the number of **features** in the input data
- Layer 2:
  - Neurons: 16
  - Activation Function: ReLU
  - Further refines learned features from the previous layer
- Output Layer:
  - Neurons: 1
  - Activation Function: Sigmoid
  - Classifies diabetes into 0 and 1

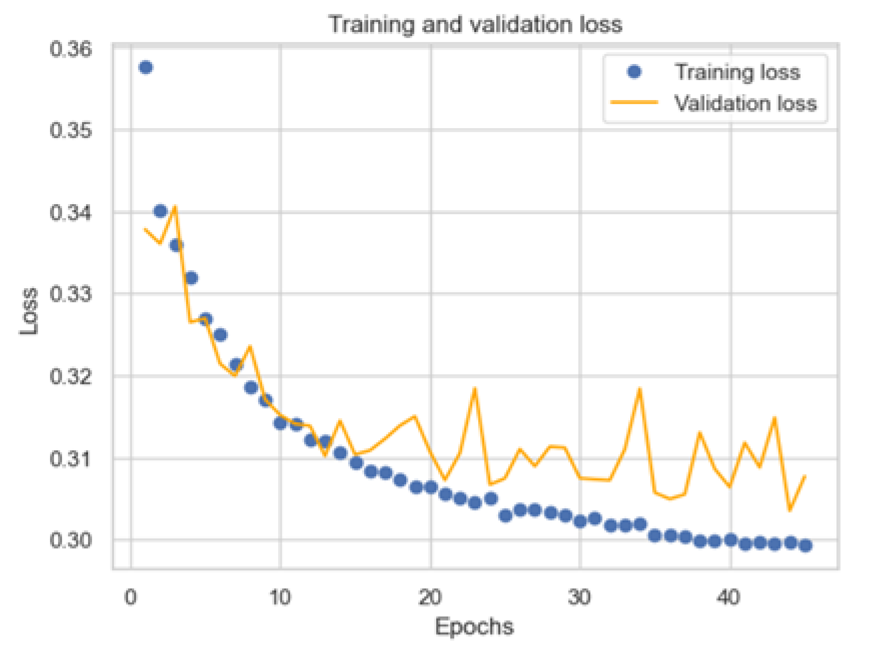


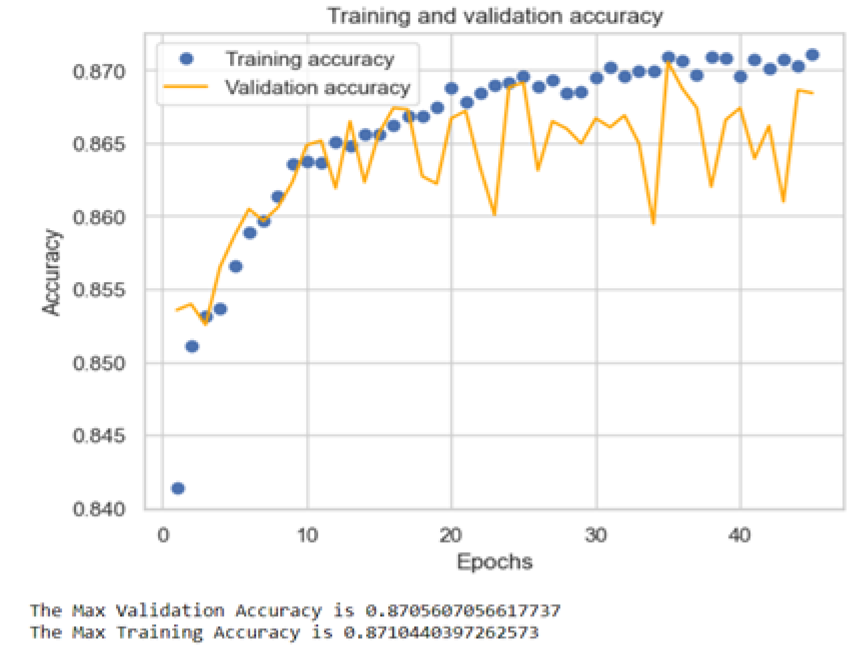
Discussion: ANN Fitting & Training

- Compiled using Adam Optimizer
- Implemented Early Stopping
- Batch Sizes of 10
- Validation split of 20%
- Monitoring Max Validation Accuracy

Discussion: ANN

Monitoring Training and Validation to Prevent Overfitting (Early Stopping)



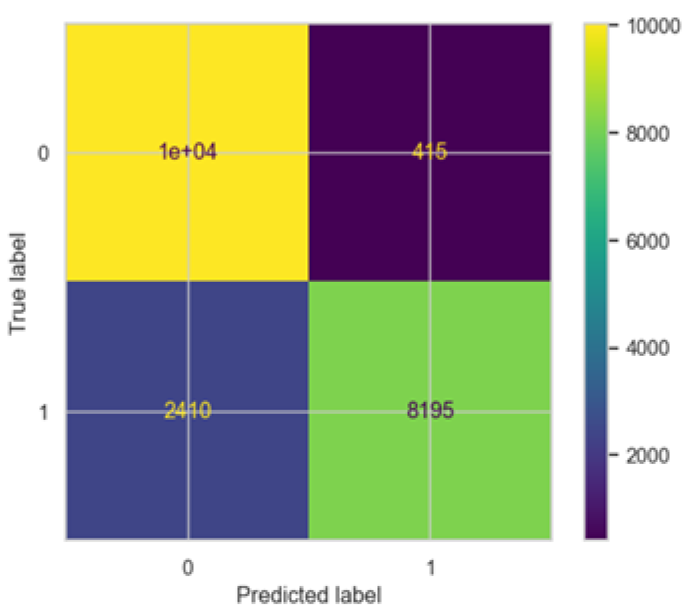


```
The Max Validation Accuracy is 0.8705607056617737
The Max Training Accuracy is 0.8710440397262573
```

ANN Evalutation

Confusion Matrix and Classification Report

	Precision	Recall	F1-score	Support
0	81%	96%	88%	10454
1	95%	77%	85%	10605
Accuracy			87%	21059
Weighted Avg	88%	87%	86%	21059



Discussion: Explainable AI Using SHAP Values

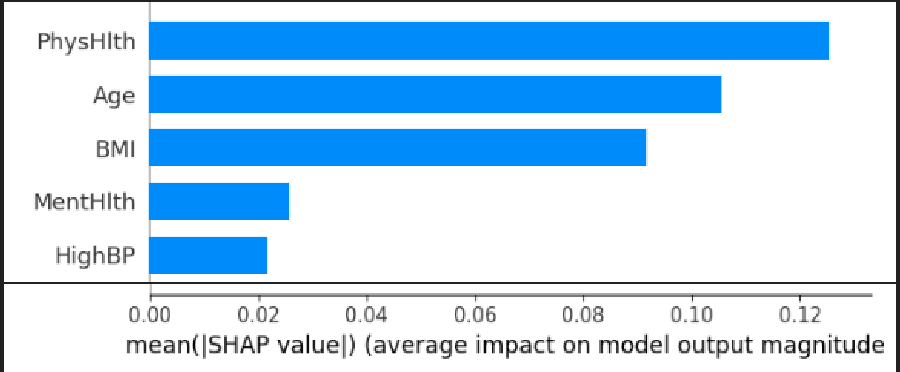
- "Shapley Additive Explanations"
- Originates from cooperative game theory
- How to distribute the "payout" (model prediction) to "players" (features)
  - Each feature's contribution to the model
- Global feature importance - mean absolute value SHAP values

SHAP Documentation

Discussion: ANN Feature Importance

Explainable AI Using Shap

Feature Name	Mean Abs Importance Score
PhysHlth	0.13
Age	0.11
BMI	0.09
MentHlth	0.03
HighBP	0.02



Recommendations & Conclusions

- **Most** important features:
  - Physical Health
  - Age
  - BMI
  - Mental Health
  - High Blood Pressure

- **Least** important features:
  - NoDocbcCost
  - Veggies
  - Fruits
  - CholCheck
  - AnyHealthcare

Recommendation and Conclusions

- Improving physical health (1) - Exercise daily & make health-conscious choices
- Increasing awareness of risk factors – higher age (2), mental health (4), & high blood pressure (5)
    - Screen often if in a high-risk group
- Maintain a healthy BMI (3) – via exercise, diet, & medical evaluations
- Public health initiatives such as National Diabetes Prevention Program | CDC
    - Focus on these five features
    - Locate events in high 'heat' areas from map tool

Contributions

- Genevieve Ferguson
  - Feature Selection & Importance
  - Classification Models & Evaluation
- Nitin Pagarani
  - Classification Models & Evaluation
  - Exploratory Analysis & Map
- Cristian Biondi
  - Exploratory Analysis
  - Correlations

Thank You! ☺

Questions?