Detection of Diabetes Among Patients With Various Risk Factors

Rohit Varanasy, Gayatri Chabra, Srinidhi Manikantan

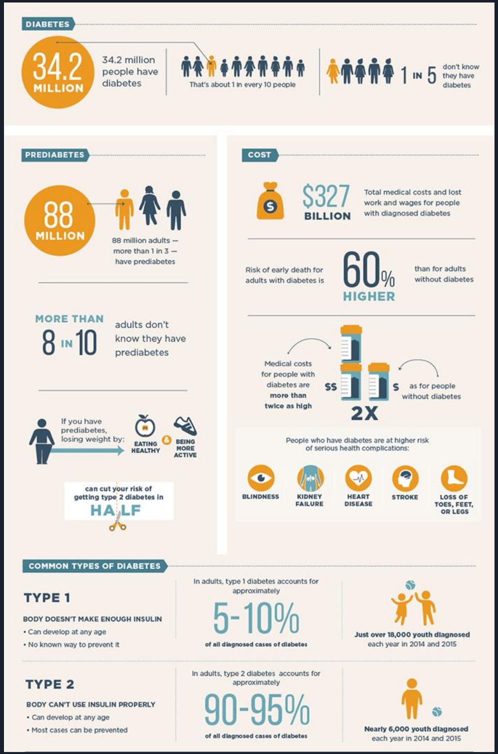
Diabetes in the States: A Dual Challenge of Health and Economics

- Diabetes, a widespread chronic disease in the U.S., poses significant health and economic challenges.

- Centers for Disease Control and Prevention (CDC) data from **2018** paints a concerning picture: **34.2 million** Americans diagnosed with diabetes, **88 million** with prediabetes.

- Furthermore, **1 in 5 diabetics** and roughly **8 in 10 pre-diabetics** are unaware of their risk.

- The economic burden is staggering, with diagnosed diabetes costs at **$327 billion** and total costs with undiagnosed diabetes and prediabetes nearing **$400 billion** annually.



Scientific Problem: Early Symptom Detection for Diabetes Prevention

The focus is on determining the likelihood of Diabetes through various risk factors.

Key Research Questions:

- **Machine Learning in Healthcare Efficacy:** How effective is machine learning in healthcare, particularly in early symptom detection for diabetes prevention?
- **Predictive Power of BRFSS Surveys:** Can survey questions from the BRFSS provide accurate predictions of whether an individual has diabetes?
- **Identifying Key Risk Factors:** What risk factors are most predictive of diabetes risk?
- **Optimizing Risk Factor Subset:** Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?

Project Aim: Developing Advanced Predictive Models for Diabetes Detection

- **Practical Significance:**
  Emphasizing the practical **impact** of our models, our aim extends beyond traditional methods. We strive for the early identification of high-risk individuals, contributing to a targeted approach for curbing the prevalence of Diabetes.

- **Methodological Perspective:**
  While existing models have relied heavily on **traditional approaches** like logistic regression, our focus is on advancing the field through the utilization of **more robust machine learning techniques.**

Information on the Dataset

Dataset

Data Source:

- The data is collected from the **Behavioral Risk Factor Surveillance Survey (BRFSS).**
- This is a **telephone survey** that is collected **annually** by the CDC, and collects health specific behaviour from about 400,000 Americans each year.
- It has been conducted every year since 1984. For the purpose of this project, we will be using the **2015 data** which is the most recent dataset that was available on Kaggle.

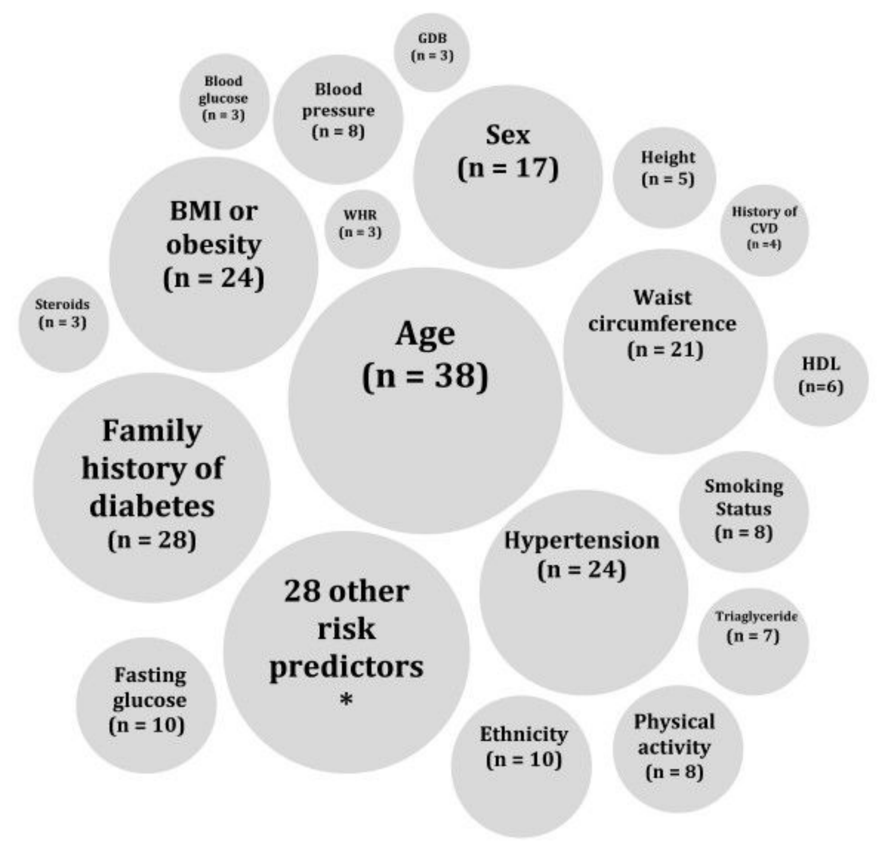
Data Description:

- The 2015 dataset we obtained from Kaggle contains **441k responses and 330 features.**
- These features are either questions directly asked of participants, or calculated variables based on individual participant responses.
- Since the dataset **collects generic health information (i.e. not specific to Diabetes),** not all 330 features will be relevant to our analysis.

Feature Selection

- Based on research and literature reviews we read online, we found the **<u>top 25 features</u>** that were important in influencing diabetes and only selected those factors for our analysis.

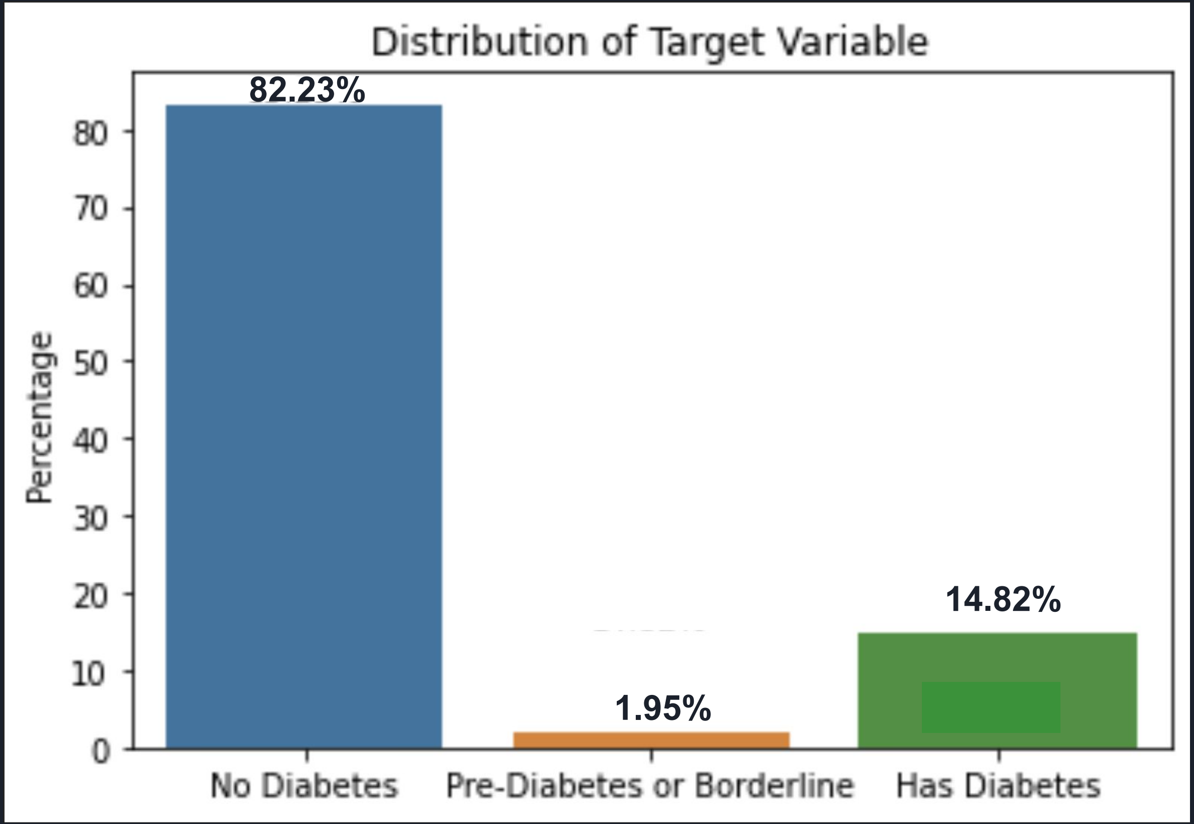
- We also analysed a systematic review of over 40 studies that looked into the prediction models for type 2 diabetes. The key variables used in their prediction models have also been considered and included in our analysis.



Chosen Predictor Variables

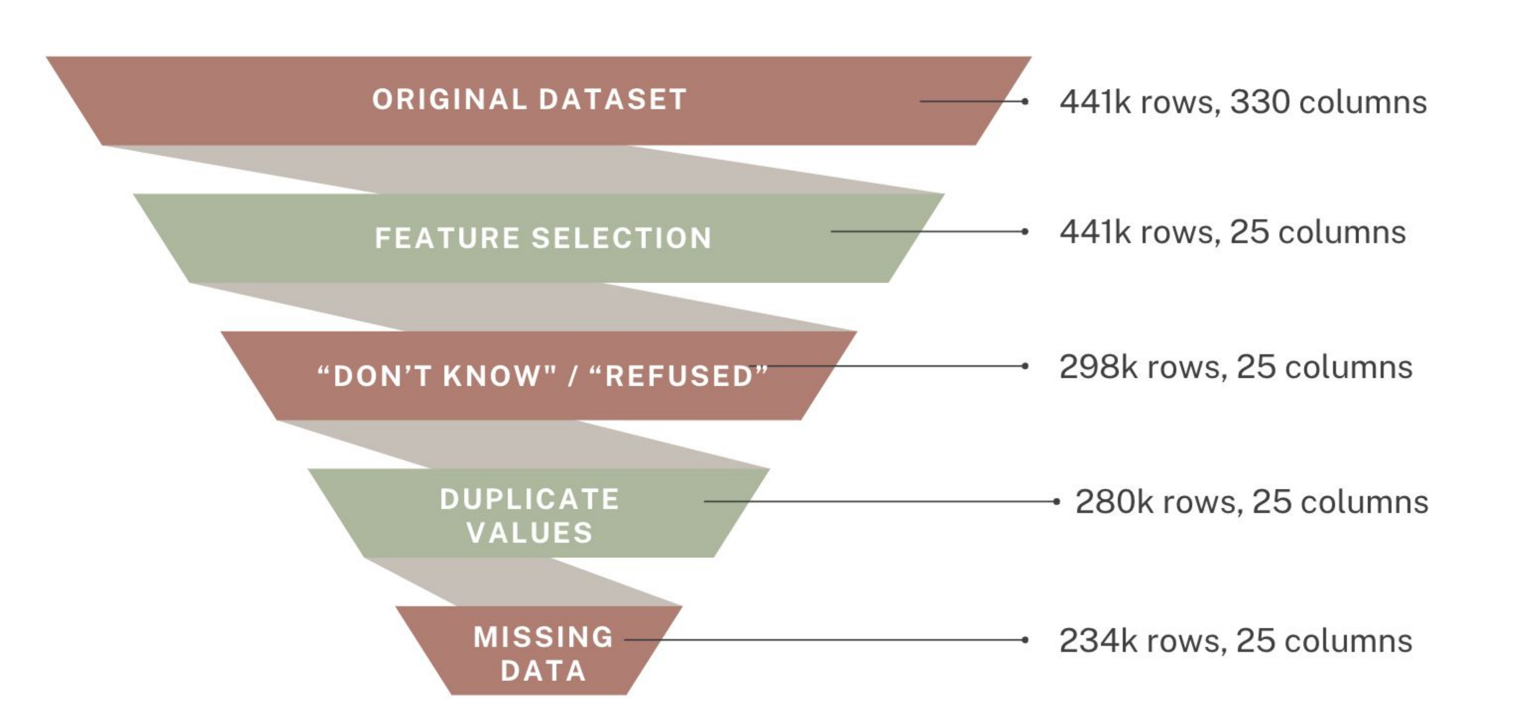
Demographics	Characteristics/ Social life	Healthcare Management	Other chronic diseases/medical conditions
• Age • Sex • Race • Education • Income • Marital Status • Employment Status	• Smoking • Physical Activity • Alcohol Consumption • Diet	• Frequency of Medical Check Up • Healthcare Coverage • Healthcare Costs • General Health • Physical Health • Mental Health • Depression	• High Blood Pressure • High Cholesterol • Cardiovascular Heart Disease • Stroke • Obesity/BMI

Target Variable



Data Curation and Processing

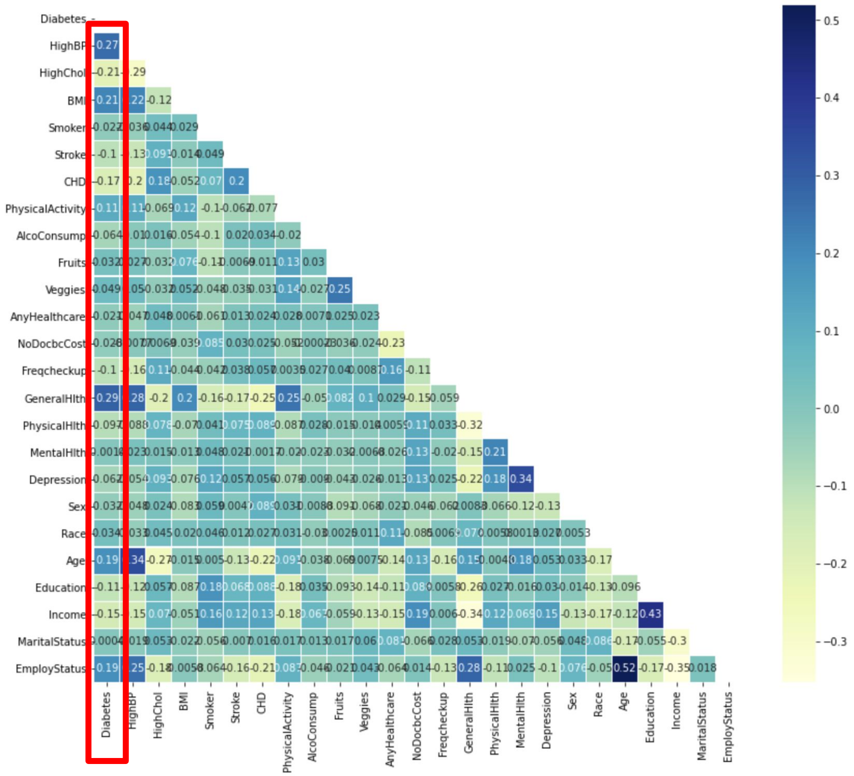
Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

In addition to that, we also checked for:

- Variables with **no unique values** (none found)
- **Outliers** (none found))
- **Erroneous Data (**none found)
- **Highly Correlated Variables** (none found)
- **Low Variance Variables** (none found)



```python
#Rename the columns to make them more readable
df_selected = df_selected.rename(columns = {'DIABETE3': 'Diabetes',
                                            '_RFHYPE5': 'HighBP',
                                            'TOLDHI2': 'HighChol',
                                            '_BMI5CAT':'BMI',
                                            '_SMOKER3':'Smoker',
                                            'CVDSTRK3':'Stroke',
                                            '_MICHD' : 'CHD',
                                            '_TOTINDA':'PhysicalActivity',
                                            '_RFDRHV5':'AlcoConsump',
                                            '_FRTLT1':'Fruits',
                                            '_VEGLT1':"Veggies",
                                            'HLTHPLN1':'AnyHealthcare',
                                            'MEDCOST':'NoDocbcCost',
                                            'CHECKUP1':'Freqcheckup',
                                            'GENHLTH': 'GeneralHlth',
                                            'PHYSHLTH':'PhysicalHlth',
                                            'MENTHLTH':'MentalHlth',
                                            'ADDEPEV2':'Depression',
                                            'SEX':'Sex',
                                            '_RACE':'Race',
                                            '_AGEG5YR':'Age',
                                            'EDUCA':'Education',
                                            'INCOME2':'Income',
                                            'MARITAL':'MaritalStatus',
                                            'EMPLOY1':'EmployStatus'})
```

Machine Learning Experiments

Simple Models

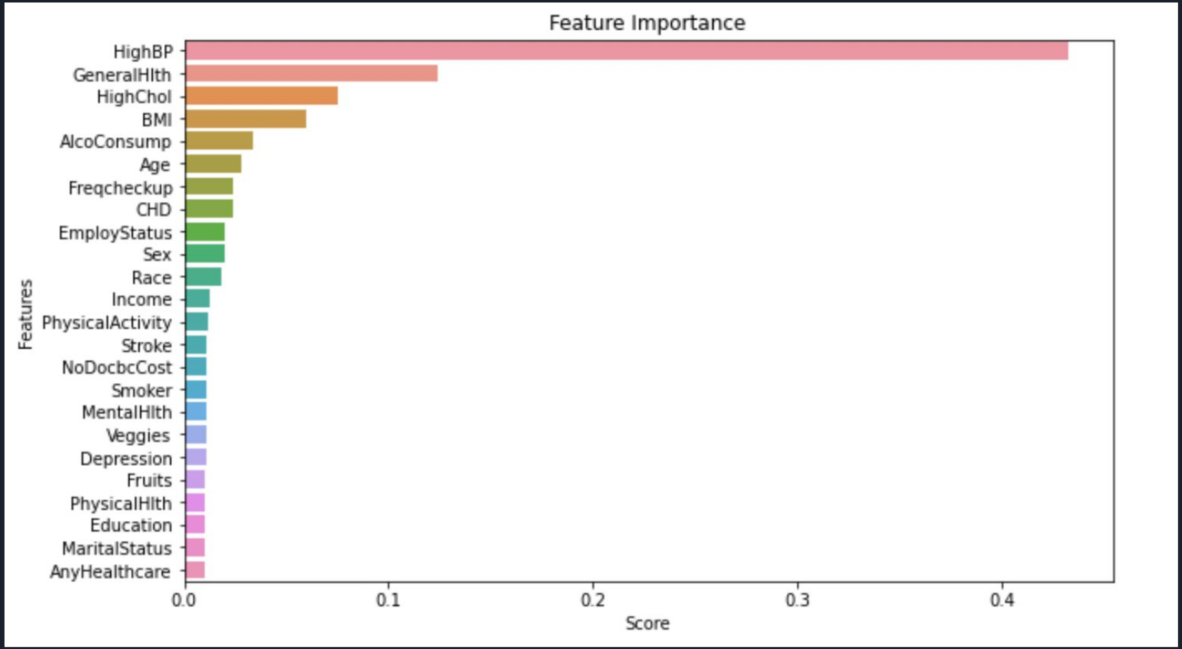
Model	**F1 score**	**Time**
Logistic Regression (no penalty)	0.794	30 seconds
Logistic Regression (L1 penalty)	0.797	3 minutes
Logistic Regression (L2 penalty)	0.797	4 minutes
Elastic Net	0.797	9 minutes
KNN	0.786	29 minutes

Ensemble Models

Model	**F1 score**	**Time**
Random Forest	0.794	3 hours
XGBoost	0.801	3 hours
ADABoost	0.797	24 minutes
LightGBM	0.800	28 minutes
Logistic Regression with L1 penalty as base, stacked with Logistic Regression with L2 as meta	0.800	22 minutes
XGBoost as base, stacked with LGBM as meta	0.797	13 minutes
Random Forest as base, stacked with XGBoost as meta	0.798	26 hours

Obtained Results

- Best F1 score is 0.801, from XGBoost
- High blood pressure is by far the most important feature when considering the diabetic state of a patient



Lessons Learned

- More complicated model might not always be better
- F1 score is more suited compared to accuracy if there is a class imbalance
- Dimensionality reduction won’t always help

Future Directions: Advancing Model Performance and Comprehensive Insights

- **Addressing Class Imbalance:**
  Integrate imbalanced learning techniques to mitigate the significant class imbalance present in the dataset, enhancing model performance.

- **Feature Selection for Precision:**
  Propose creating a concise set of questions derived from the BRFSS using feature selection to accurately predict diabetes or identify high-risk individuals.

- **Integration of Multimodal Data:**
  Explore the possibility of incorporating diverse data sources, such as clinical data, biomarkers, and lifestyle information, to augment the predictive power of our models and enhance their comprehensiveness.

Navigating Data Limitations: A Deeper Look

- **Causality Limitation:**
  Address the cross-sectional constraint in BRFSS data, noting the challenge of establishing causality.

- **Recall Bias Concerns:**
  Recognize potential recall bias in self-reported BRFSS data as a limitation influencing our predictive models.

- **Missing Predictor Variables:**
  Highlight the absence of sleep and family history variables as limitations, pointing to areas for refining predictive modeling insights.

References

1. *Behavioral Risk Factor Surveillance System 2015 Codebook Report Land-Line and Cell-Phone data*. (2016). https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf
2. Collins, G. S., Mallett, S., Omar, O., & Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine*, *9*(1). https://doi.org/10.1186/1741-7015-9-103
3. *Diabetes Health Indicators Dataset Notebook*. (n.d.). Kaggle.com. https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/notebook
4. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109


Thank You