Team Data-betics: David Kim, Jiyoon Moon, Rebekah Northrup, Sahithi Rampally

IDENTIFYING KEY RISK FACTORS FOR DIABETES AND DEVELOPING PREVENTION STRATEGIES

Introduction

Diabetes is a growing health crisis influenced by lifestyle and demographic factors. Our goal is to identify key risk factors and develop prevention strategies using insights from the Behavioral Risk Factor Surveillance System (BRFSS) dataset, focusing on the broader American population.

Data Overview

- Source: BRFSS 2015 (253,680 responses).
- Features: 22, including demographics, BMI, activity, general/mental health.
- Limitations: Self-reported data and underrepresentation of low-income and rural populations.

Methodology

Data Preparation

- Filtered dataset to focus on individuals with or without diabetes (excluded pre-diabetics).
- Created a binary target variable, Diabetes_bin.
- Identified top 11 features most correlated with diabetes (e.g., General Health, Blood Pressure, BMI).

Data Exploration

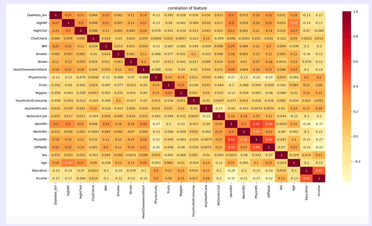
- Conducted correlation analysis to identify the most predictive features.
- Key findings: Variables like General Health (GenHlth), High Blood Pressure (HighBP), and Difficulty Walking (DiffWalk) are strongly associated with diabetes.

Our Model

Model Building

- Model: Decision Tree Classifier (max depth = 3 for interpretability and to avoid overfitting).
- Top 11 Features Used:

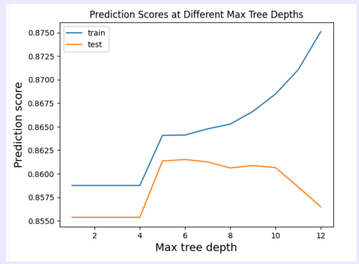
Diabetes_bin	1.000000
GenHlth	0.300347
HighBP	0.269319
DiffWalk	0.223991
BMI	0.222353
HighChol	0.205684
Age	0.181727
HeartDiseaseorAttack	0.181258
PhysHlth	0.175754
Income	0.168651
Education	0.128149
PhysActivity	0.121028



- Trained with 80% of data; tested on 20%

Model Performance Insights

- Performance plateaued at tree depths of 5-6 (Figure below).
- Simplified depth of 3 chosen to prevent overfitting.



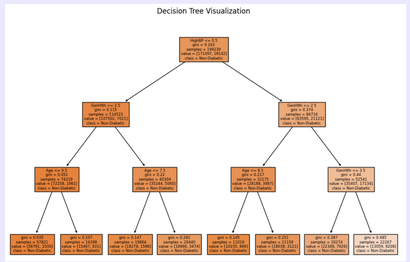
Prediction Scores at Different Tree Depths: Accuracy plateaued at depths of 5-6 (Figure 1).

Model Evaluation

- Accuracy:
  - Training: 85.88%
  - Testing: 85.54%
- Class Imbalance Challenge: With ~85% of individuals being non-diabetic, the model could achieve high accuracy simply by predicting "non-diabetic" for all cases.
- Key Insight: HighBP consistently emerged as the most significant predictor across multiple iterations.

	Num_Features	Root_Feature	Frequency
0	11	HighBP	30
1	8	HighBP	20
2	5	GenHlth	12
3	5	HighBP	10
4	8	GenHlth	7
5	5	DiffWalk	4
6	5	HighChol	3
7	8	DiffWalk	3
8	5	BMI	1

Random Feature Subsets: HighBP was the most frequent root feature (Figure 2).



Decision Tree Visualization: Key features and their hierarchical relationships (Figure 3).

Model Deployment & Challenges

Potential Applications:

- Screening for diabetes risk using simple patient inputs.
- Guiding targeted interventions for high-risk groups.

Challenges:

- Data bias from self-reporting and underrepresentation of low-income/rural populations.
- Ethical considerations and risks of misclassification due to class imbalance.
- Addressing overfitting and ensuring generalizability.

Results & Key Takeaways

- Top Predictors of Diabetes: HighBP, GenHlth, DiffWalk, BMI, and HighChol.
- HighBP was the most frequent root feature in random subsets of data.
- Strong correlation between poor health metrics (e.g., BMI, GenHlth) and diabetes risk.

Conclusion

Our model identifies critical risk factors for diabetes, with HighBP emerging as the most informative predictor. These findings provide actionable insights for prevention and early intervention strategies. Despite challenges like class imbalance and dataset bias, this approach provides a foundation for early diabetes detection and intervention planning.

References

Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System (BRFSS) 2015 Data. 2015, https://www.cdc.gov/brfss/annual_data/annual_2015.html.