AI Colab Group 1 - Clinical Data Science & Modeling

Machine Learning-Based Prediction of Type 2 Diabetes from Kidney Function and β-cell Dysfunction

Yara Yaghi, Naod Dawit, Kareem Aly, Indira Kuppa

Proposed Research Question

How does kidney function, as measured by creatinine clearance (CCR) and blood urea nitrogen (BUN), relate to β-cell dysfunction and type 2 diabetes risk, and does this relationship differ by smoking status and alcohol consumption among U.S. adults?

Potential Hypothesis

Reduced kidney function, indicated by lower creatinine clearance and higher BUN, is associated with impaired β-cell function and increased risk of type 2 diabetes, especially among current smokers and drinkers.

BACKGROUND AND INTRODUCTION

- Type 2 Diabetes (T2D) affects over 36 million U.S. adults and is a leading cause of cardiovascular disease, kidney failure, and premature mortality.
- Early identification of individuals at risk is crucial to enable timely intervention and prevent complications.
- Kidney function markers such as Creatinine Clearance (CCR) and Blood Urea Nitrogen (BUN) have shown associations with metabolic dysfunction and may serve as early predictors of T2D.
- β-cell dysfunction, which impairs insulin secretion, plays a central role in the development of T2D and can be measured using HOMA-β, derived from fasting glucose and insulin.
- Lifestyle factors like smoking and alcohol consumption may further influence the relationship between biological markers and diabetes risk.

Literature Review

Metabolic & Behavioral factors:

- Individuals with high-normal fasting glucose (91–99 mg/dL) have a greater risk of developing diabetes compared to those with lower normal levels (Brambilla et al., 2011)
- Triglycerides are an independent and early predictor of type 2 diabetes (Zhao et al., 2019)
- High BMI amplifies risk across all metabolic risk markers (Zhao et al., 2019)
- Smoking and alcohol use worsen metabolic regulation (Akhuemonkhan & Lazo, 2017)

Literature Review

Beta Cell Dysfunction

- Beta cells located in the pancreas produce and secrete insulin (Dludla et al., 2023).
- Beta cell dysfunction indicates impaired insulin secretion, contributing to T2DM (Dludla et al., 2023).
- The Homeostatic Model Assessment of Beta-cell Function (HOMA-B) and the Insulinogenic Index can be used to indicate beta cell function (Kim et al., 2024; Sung et al., 2009)
    - These indicators can be calculated with fasting and post-load glucose and insulin values.

Literature Review

Type II Diabetes:

- Type II diabetes, also known as adult-onset diabetes, occurs when the body is not able to utilize insulin correctly and sugar builds up in the blood (Mayo Clinic, 2025).
    - Type II diabetes is more common in older adults (hence adult-onset), however, more and more children are being diagnosed with the rise of childhood obesity (Mayo Clinic, 2025).
- As of 2024, more than 38 million Americans have diabetes, with close to 95% of diagnoses being for Type II diabetes (CDC, 2024)
    - Mostly in adults over 45 years old, but more and more children are getting diagnosed.

OBJECTIVES

- Primary Objective
  ➢ To evaluate the relationship between kidney function (measured by Creatinine Clearance and BUN) and β-cell dysfunction (via HOMA-β) in predicting the risk of Type 2 Diabetes among U.S. adults.
- Secondary Objectives
  ▸ To assess whether smoking status modifies the association between kidney/β-cell function and diabetes risk.
  ▸ To determine if alcohol consumption influences these relationships.
  ▸ To build a predictive model for Type 2 Diabetes using clinical and lifestyle variables from NHANES.

DATA SOURCE

- Dataset: National Health and Nutrition Examination Survey (NHANES)
- Years Covered: 1999–2020 (Multiple 2-year cycles combined)
- Population: U.S. adults aged 30 and above
- Source Website: https://wwwn.cdc.gov/nchs/nhanes/

NHANES Data Modules Used (2007–2017):

- Demographics Module
- Laboratory Module
  - Fasting Glucose & Insulin (LBXGLU, LBXIN)
  - Kidney Function Biomarkers: BUN & Creatinine (LBXSBU, LBXSCR)
  - Urine Albumin-Creatinine Ratio & Components (URDACT, URXUMA, URXUCR)
- Diabetes Questionnaire Module
- Smoking Questionnaire Module
- Alcohol Use Questionnaire Module

INCLUSION AND EXCLUSION CRITERIA

Included:

- Participants from NHANES cycles 1999-2020
- Age ≥ 30
  - To minimize inclusion of early-onset or Type 1 diabetes
- Has Type 2 diabetes (self-reported)
- Available fasting glucose and fasting insulin values

Excluded:

- Missing key health variables
- Missing demographic & behavioral variables
- eGFR < 30 mL/min/1.73m^2
  - Indicates severe chronic kidney disease (Stage 4+)
- Participants without diabetes (self-reported)

DATA DICTIONARY BEFORE CLEANING

	Feature Name	Data Type	Missing Values	Unique Values	Description
0	SEQN	float64	0	27706	Respondent sequence number (unique ID for each...
1	LBXGLU	float64	1489	1332	Fasting glucose (mg/dL)
2	LBXIN	float64	2015	4210	Fasting insulin (μU/mL)
3	LBXSBU	float64	1801	75	Blood urea nitrogen (BUN) (mg/dL), marker of k...
4	LBXSCR	float64	1800	317	Serum creatinine (mg/dL), used to assess kidne...
5	LBXSATSI	float64	1826	210	Serum sodium concentration (mmol/L)
6	SMQ020	float64	140	4	Ever smoked at least 100 cigarettes in life (1...
7	SMQ040	float64	15294	3	Current smoking status (1 = Every day, 2 = Som...
8	EverDrank	float64	17061	3	No description available
9	DrinkFrequency	float64	6055	82	No description available
10	AvgDrinksPerDay	float64	11018	30	No description available
11	RIDAGEYR	float64	0	68	Age in years at time of screening
12	RIAGENDR	float64	0	2	Gender (1 = Male, 2 = Female)
13	DMDEDUC2	float64	628	7	Education level (1 = Less than 9th grade to 5 ...
14	INDFMPIR	float64	2668	501	Ratio of family income to poverty level (highe...
15	DIQ010	float64	0	4	Doctor told you have diabetes (1 = Yes, 2 = No)
16	DID040	float64	24323	85	Age when first told you had diabetes
17	DIQ050	float64	4369	3	Currently taking insulin (1 = Yes, 2 = No)
18	DIQ070	float64	22596	3	Currently taking pills to lower blood sugar (1...
19	SurveyCycle	object	0	10	NHANES survey cycle years
20	URDACT	float64	641	12090	Urine albumin-to-creatinine ratio (mg/g), mark...

STEP 1: DATA MERGE

- Downloaded relevant NHANES datasets from 1999-2020 by cycle (e.g., GLU, BIOPRO, ALQ, etc.)
- Selected only features required for analysis (e.g., glucose, insulin, smoking, alcohol, blood urea nitrogen, etc.)
- Merged tables within each year on SEQN (patient ID's) to build year-specific datasets
- Performed feature engineering to standardize column names across years
- Calculated missing features from older datasets to standardize all columns (e.g., using Creatine and Albumin to calculate ACR (urine albumin-to-creatinine ratio)
- Added 'SurveyCycle' column to track year range
- Merged all yearly datasets into one master dataframe

STEP 2: DATA PREPARATION

- Checking for duplicate entry IDs in the all years datasets
- Selected individuals with diabetes
- Flagged likely Type 1 Diabetes based on:
  - Diagnosed before age 30
  - Started insulin within one year
  - Not currently on diabetes pills or missing pill info
- Removed flagged Type 1 cases to retain likely Type 2 Diabetes patients only
- Handled missing values using median, mode, and logical imputation (e.g., setting alcohol variables to 0 for non-drinkers).
- Feature Engineer:
  - HOMA-B (β-cell function): (20 x Insulin) / (Glucose - 3.5)
  - Excluded participants with eGFR < 30 mL/min/1.73m^2 (indicating severe kidney disease)

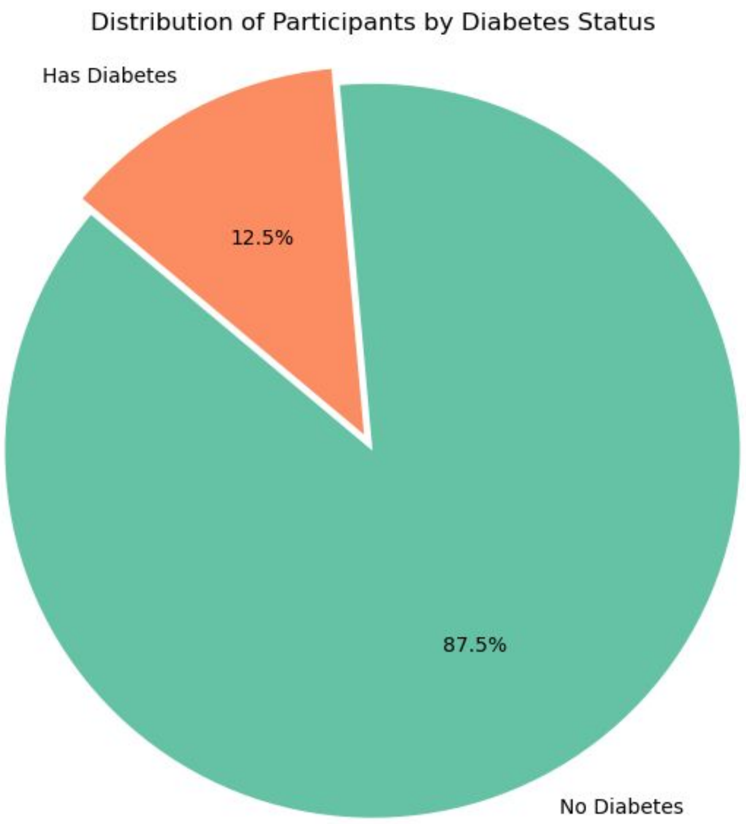
DATA DICTIONARY AFTER CLEANING

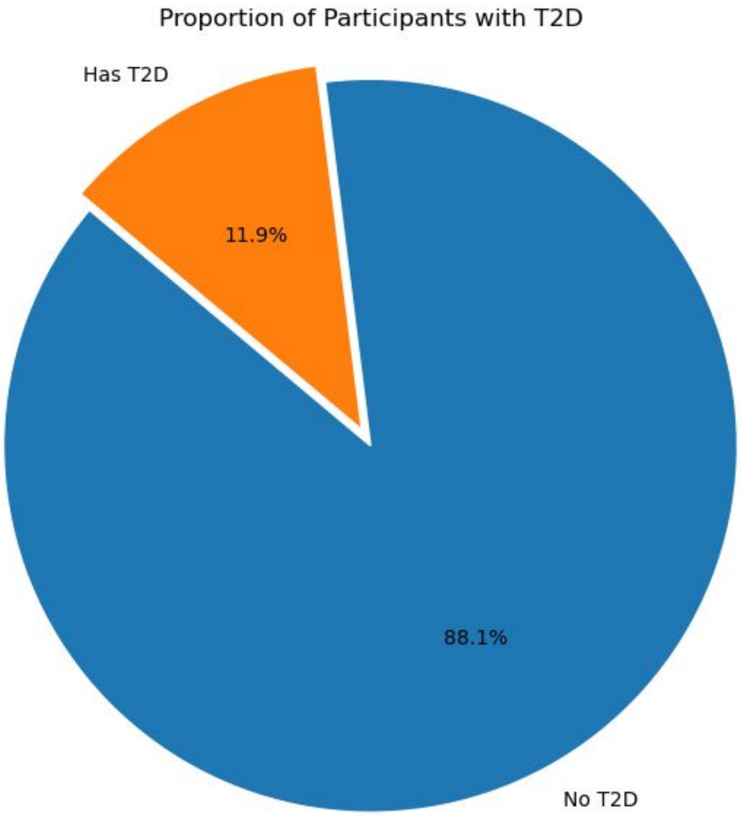
	Feature Name	Data Type	Missing Values	Unique Values	Description
0	SEQN	float64	0	3113	Respondent sequence number (unique ID for each...
1	FastingGlucose	float64	0	733	Fasting glucose level (mg/dL)
2	FastingInsulin	float64	0	1948	Fasting insulin level (µU/mL)
3	BUN	float64	0	51	Blood urea nitrogen (mg/dL), marker of kidney ...
4	SerumCreatinine	float64	0	167	Serum creatinine (mg/dL), used to estimate kid...
5	LBXSATSI	float64	0	107	Serum sodium concentration (mmol/L)
6	EverSmoked100	float64	0	4	Ever smoked at least 100 cigarettes in life (1...
7	EverDrank	float64	0	3	Ever had at least one alcoholic drink (1 = Yes...
8	DrinkFrequency	float64	0	28	Drinking frequency over past 12 months (0 = Ra...
9	AvgDrinksPerDay	float64	0	18	Average number of alcoholic drinks per day ove...
10	Age	float64	0	68	Age in years at time of screening
11	Gender	float64	0	2	Gender (1 = Male, 2 = Female)
12	Education	float64	0	7	Education level (1 = Less than 9th grade to 5 ...
13	IncomeToPovertyRatio	float64	0	464	Ratio of family income to poverty level
14	HasDiabetes	float64	0	1	Has doctor-diagnosed diabetes (1 = Yes, 2 = No)
15	DIQ050	float64	0	3	Currently taking insulin (1 = Yes, 2 = No)
16	DIQ070	float64	0	3	Currently taking diabetes pills (1 = Yes, 2 = No)
17	ACR	float64	0	2538	Urine albumin-to-creatinine ratio (mg/g), mark...
18	likely_type1	bool	0	1	Flag for likely type 1 diabetes (True/False)
19	T2D	int64	0	1	Type 2 diabetes classification (1 = Yes, 0 = No)
20	IncomeMissing	int64	0	2	Flag if income data is missing (True/False)
21	HOMA_B	float64	0	2920	Homeostatic Model Assessment of Beta-cell func...
22	CurrentSmoker_2.0	int64	0	2	Current smoking status: some days (dummy varia...
23	CurrentSmoker_3.0	int64	0	2	Current smoking status: not at all (dummy vari...
24	CurrentSmoker_Missing	int64	0	2	Current smoking status missing (dummy variable)
25	CurrentSmoker_Not at all	int64	0	2	Current smoking status labeled 'Not at all'
26	SurveyCycle_2001-2002	int64	0	2	Survey cycle dummy for 2001–2002
27	SurveyCycle_2003-2004	int64	0	2	Survey cycle dummy for 2003–2004
28	SurveyCycle_2005-2006	int64	0	2	Survey cycle dummy for 2005–2006
29	SurveyCycle_2007-2008	int64	0	2	Survey cycle dummy for 2007–2008
30	SurveyCycle_2009-2010	int64	0	2	Survey cycle dummy for 2009–2010
31	SurveyCycle_2011-2012	int64	0	2	Survey cycle dummy for 2011–2012
32	SurveyCycle_2013-2014	int64	0	2	Survey cycle dummy for 2013–2014
33	SurveyCycle_2015-2016	int64	0	2	Survey cycle dummy for 2015–2016
34	SurveyCycle_2017-2020	int64	0	2	Survey cycle dummy for 2017–2020
35	EverDrank_Label_Drinks Alcohol	int64	0	2	Label indicating whether participant drinks al...
36	eGFR	float64	0	2233	Estimated glomerular filtration rate (mL/min/1...

Features: 37
Rows (# of patients): 3,113
No missing values.
No duplicates.

After some further analysis, many features will be dropped.

EXPLORATORY ANALYSIS AND VISUALIZATIONS



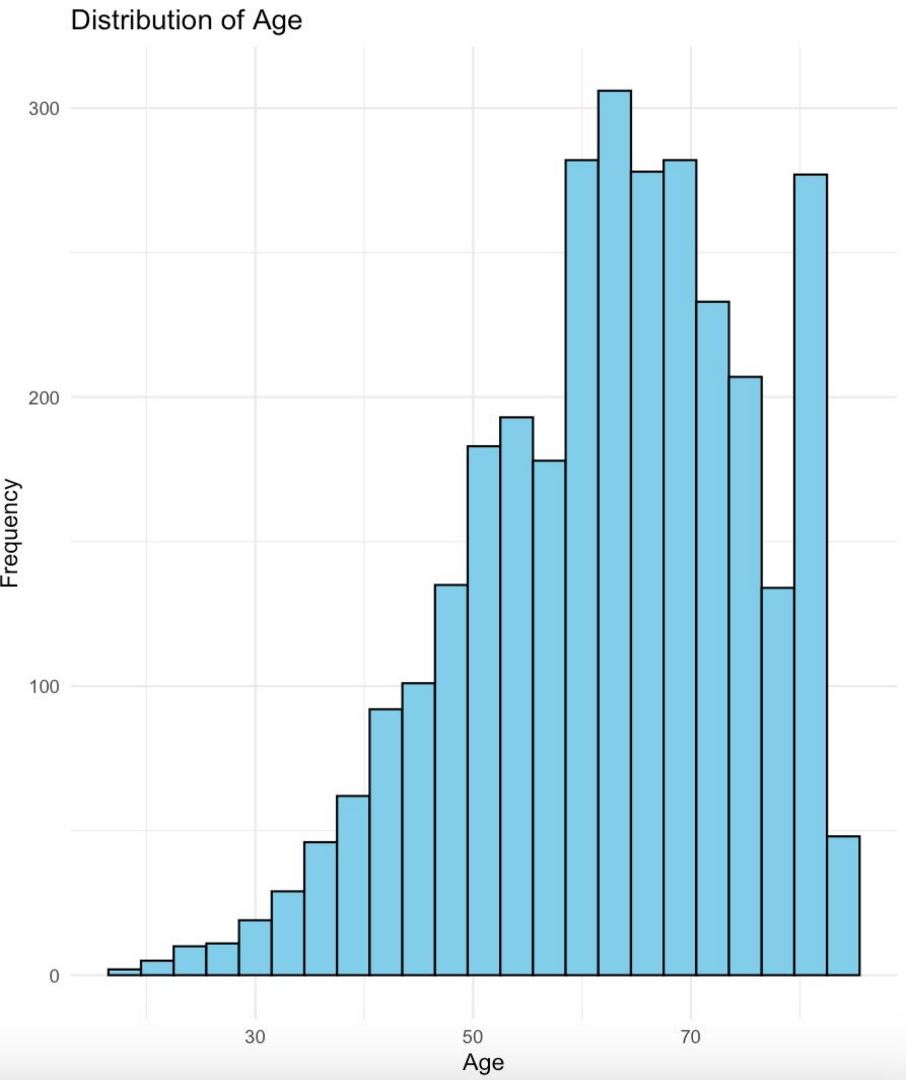


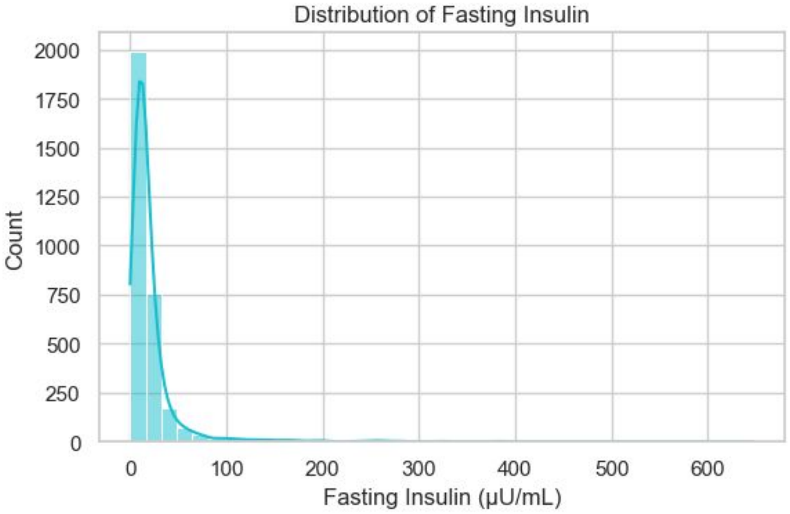
Pie chart represents the proportion of participants with and without diabetes using the HasDiabetes variable (DIQ010) from the NHANES dataset. DIQ010 is a self-reported survey question that asks: “Has a doctor ever told you that you have diabetes?”

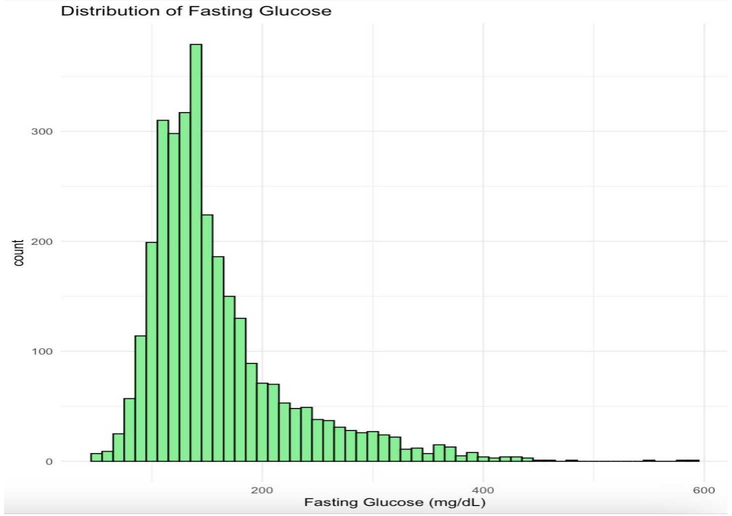
Pie chart includes participants after filtering out likely Type 1 diabetes, so it shows

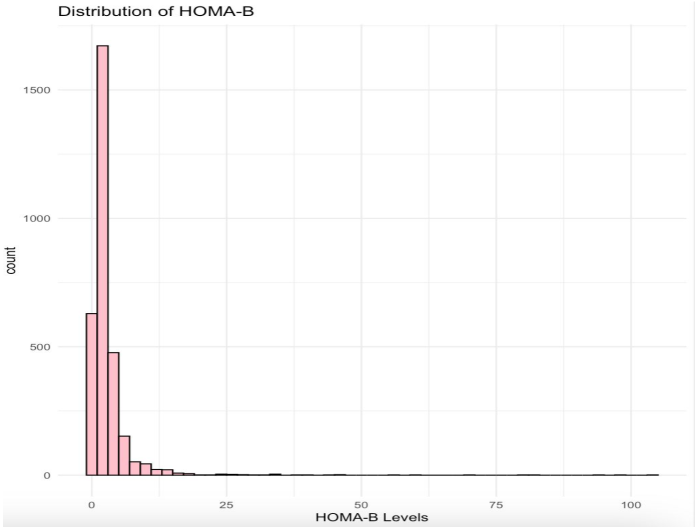
- T2D = 1 → Confirmed Type 2 Diabetes
- T2D = 0 → Everyone else (Non-Diabetic only)

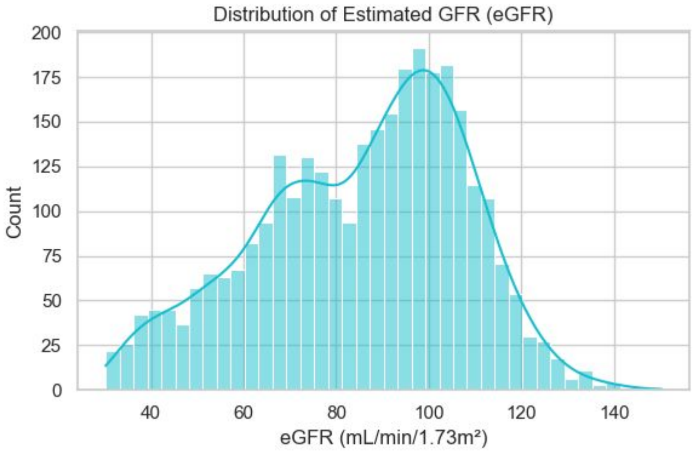
Reflects a cleaned T2D cohort (e.g., 3113 out of ~27,000)

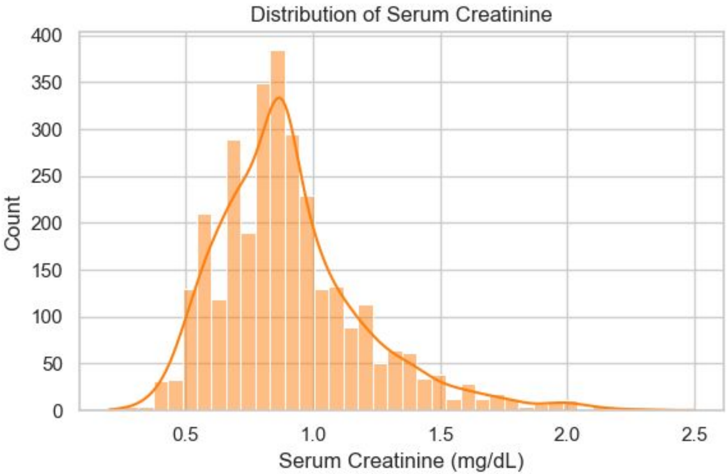

Distribution of Age
Frequency
Age
0
100
200
300
30
50
70

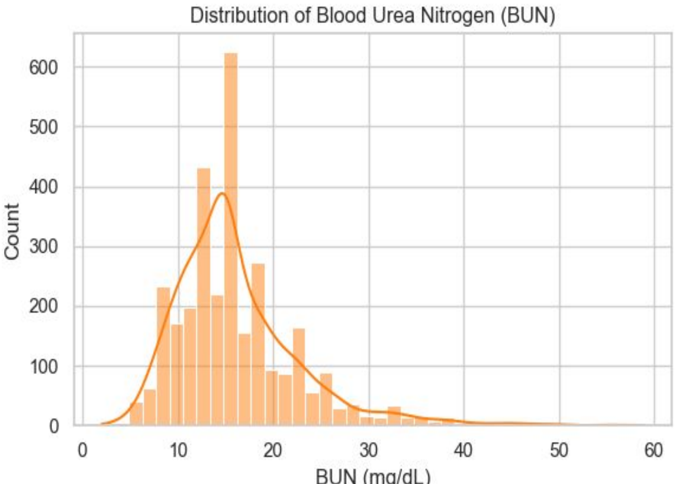

Distribution of Fasting Insulin
Count
Fasting Insulin (µU/mL)
0
250
500
750
1000
1250
1500
1750
2000
0
100
200
300
400
500
600

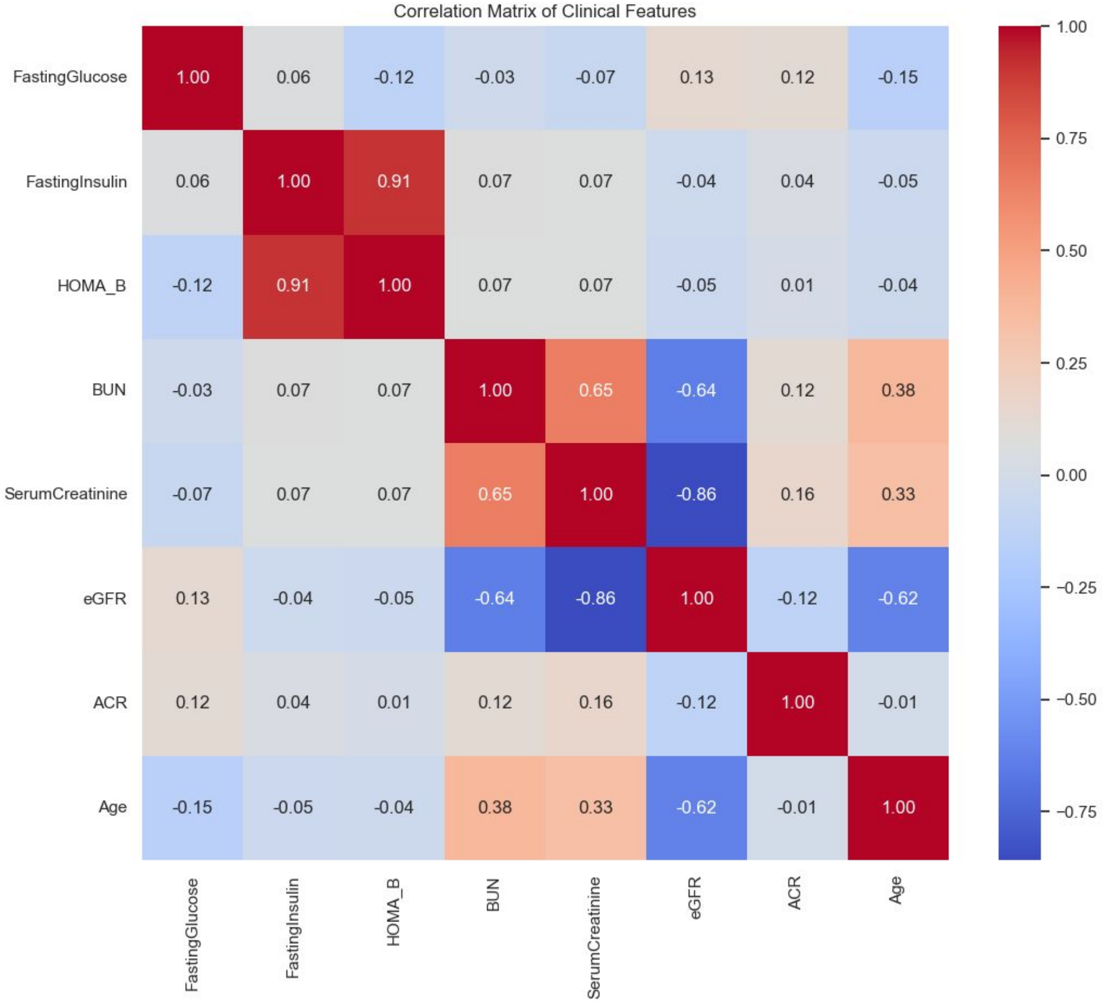

Distribution of Fasting Glucose
count
Fasting Glucose (mg/dL)
0
100
200
300
200
400
600


Distribution of HOMA-B
count
1500
1000
500
0
0
25
50
75
100
HOMA-B Levels


Distribution of Estimated GFR (eGFR)
Count
200
175
150
125
100
75
50
25
0
40
60
80
100
120
140
eGFR (mL/min/1.73m²)


Distribution of Serum Creatinine
Count
400
350
300
250
200
150
100
50
0
0.5
1.0
1.5
2.0
2.5
Serum Creatinine (mg/dL)


Distribution of Blood Urea Nitrogen (BUN)
Count
600
500
400
300
200
100
0
0
10
20
30
40
50
60
BUN (mg/dL)



Correlation Matrix of Clinical Features

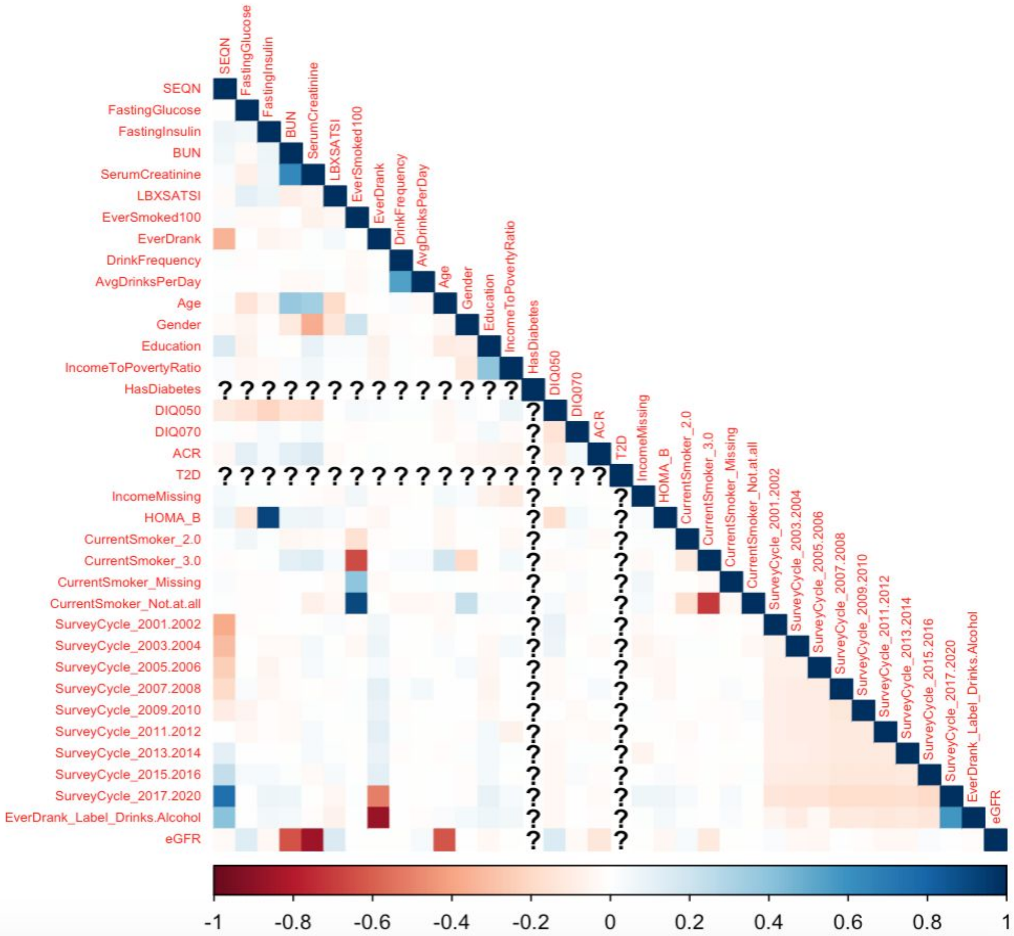
HOMA-B and Fasting Insulin show a very strong positive correlation (0.91), confirming their close mathematical and physiological relationship.

Serum Creatinine and eGFR display a very strong inverse relationship (–0.86), which is expected since eGFR is calculated from serum creatinine and age.

BUN is moderately correlated with both eGFR (–0.64) and Serum Creatinine (0.65), reflecting its role as a renal function marker.

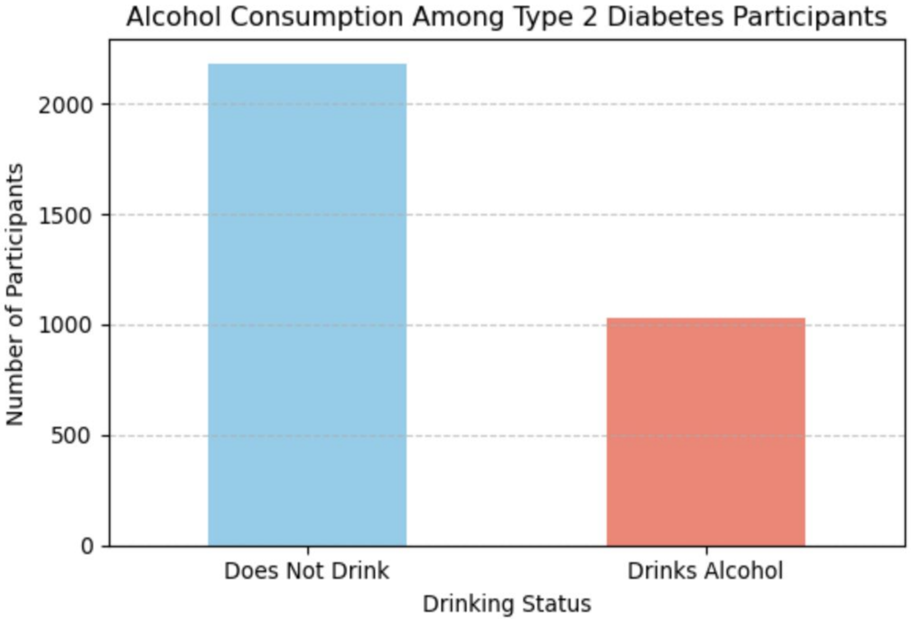
Age shows a moderate negative correlation with eGFR (–0.62), indicating declining kidney function with aging.

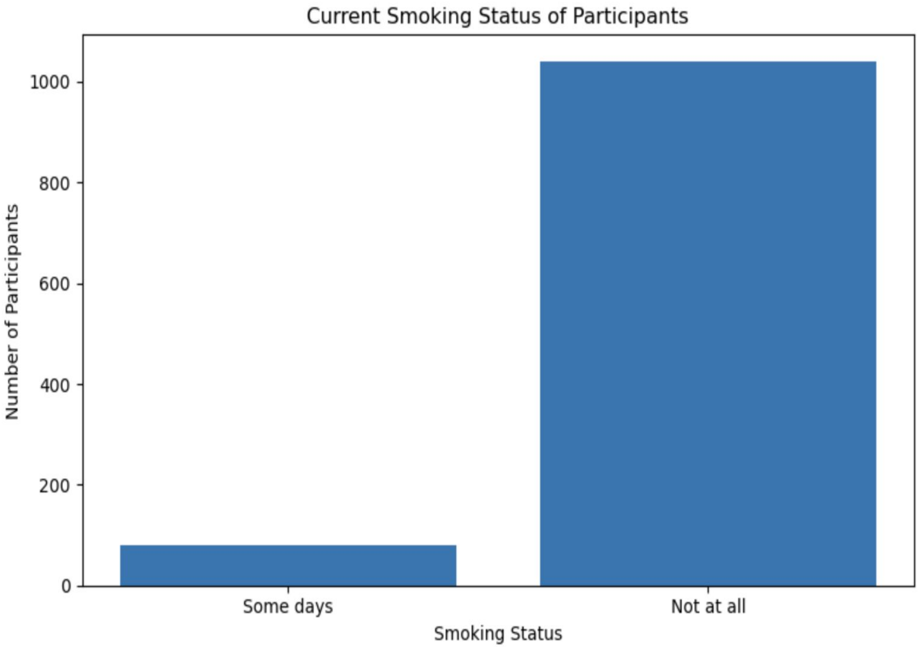
ACR shows weak correlations with all other variables, suggesting it captures unique information (e.g., microalbuminuria) that complements other markers.



Strong Correlations:

- <u>Fasting Insulin & HOMA-B:</u> HOMA- B is essentially proportional to fasting insulin (numerator of HOMA equation). As fasting insulin increases, so does HOMA-B levels.
- <u>Serum Creatinine & eGFR:</u> eGFR calculates the amount of blood filtered by the kidneys per minute, while Serum Creatine is a waste product filtered out by the kidneys.
- <u>BUN & Serum Creatinine:</u> when kidney function worsens and waste isn't filtered out effectively, both BUN and Serum Creatine rise.





May need SMOTE to balance dataset.

Methodology

Completed:

- **Data Collection & Integration:** Merged NHANES datasets from 1999–2020 across multiple survey cycles.
- **Feature Selection & Engineering:** Created derived features like HOMA-B and eGFR; harmonized alcohol and smoking variables across years.
- **Data Cleaning & Filtering:** Addressed missing values using mode/median/logical rules, removed biologically implausible entries, and filtered for Type 2 diabetes patients with valid eGFR.
- **One-Hot Encoding & Preprocessing:** Prepared categorical variables and normalized features for analysis.
- **Exploratory Data Analysis (EDA):** Visualize trends and feature distributions across subgroups (e.g., race, age, income).

Future Work:

- **Feature Selection & Multicollinearity Handling:** Investigate and mitigate redundancy among features (e.g., insulin, glucose, HOMA-B).
- **Model Development:** Train classification models to predict outcomes of interest (e.g., poor β-cell function).
- **Model Evaluation & Tuning:** Use metrics like AUC, precision, and recall with cross-validation to assess performance.
- **Interpretability & Insights:** Identify top predictors and draw health-related conclusions.

References

Akhuemonkhan, E. & Lazo, M. (2017). Association between family history of diabetes and cardiovascular disease and lifestyle risk factors in the U.S. population (NHANES 2009–2012). *Preventive Medicine, 96*, 129–134. https://doi.org/10.1016/j.ypmed.2016.12.015

Baliunas, A., Taylor, B., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). Alcohol as a Risk Factor for Type 2 Diabetes. *Diabetes Care, 32*(11), 2123-2132. https://doi.org/10.2337/dc09-0227

Brambilla, P., La Valle, E., Falbo, R., Limonta, G., Signorini, S., Cappellini, F., & Mocarelli, P. (2011). Normal fasting plasma glucose and risk of type 2 diabetes. *Diabetes Care, 34*(6), 1372–1374. https://doi.org/10.2337/dc10-2263

CDC. (2024, May 15). About Type 2 Diabetes. Diabetes. https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html

Dludla, P. V., Mabhida, S. E., Ziqubu, K., Nkambule, B. B., Mazibuko-Mbeje, S. E., Hanser, S., Basson, A. K., Pheiffer, C., & Kengne, A. P. (2023). Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World Journal of Diabetes, 14*(3), 130–146. https://doi.org/10.4239/wjd.v14.i3.130

Kim, J. Y., Lee, J., Kim, S. G., & Kim, N. H. (2024). Recent Glycemia Is a Major Determinant of β-Cell Function in Type 2 Diabetes Mellitus. *Diabetes & Metabolism Journal*, *48*(6), 1135-1146. https://doi.org/10.4093/dmj.2023.0359

Mayo Clinic. (2025, February 27). Type 2 diabetes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Sung, K. C., Reaven, G., & Kim, S. (2009). Utility of Homeostasis Model Assessment of β-Cell Function in Predicting Diabetes in 12,924 Healthy Koreans. *Diabetes Care, 33*(1), 200-202. https://doi.org/10.2337/dc09-1070

Zhao, J., Zhang, Y., Wei, F., Song, J., Li, W.-D., Chen, C., Zhang, K., & Feng, S. (2019). Triglyceride is an independent predictor of type 2 diabetes among middle-aged and older adults: A prospective study with 8-year follow-ups in two cohorts. *Journal of Translational Medicine, 17*, 354. https://doi.org/10.1186/s12967-019-02156-3

Literature Review

Family History:

- Individuals with a first degree relative with diabetes have significantly higher odds of developing type 2 diabetes, independent of lifestyle (Duschek et al., 2023)
    - FH+ individuals show elevated BMI, triglycerides, ALT, and impaired fasting glucose, indicators of increased risk for type 2 diabetes (Akhuemonkhan & Lazo, 2017)
- Individuals with FH are more likely to have lifestyle risks
    - In a 2017 study, participants with a family history of diabetes were more likely to be smokers and/or overweight (Akhuemonkhan & Lazo, 2017)

Literature Review

Healthcare Access Barriers:

- Barriers such as lack of insurance, high medication costs, and transportation issues are linked to worse diabetes and hypertension control.
  - For instance, patients with transportation barriers had higher HbA1c and blood pressure levels over three years (Berkowitz et al., 2024).
- Social needs like housing insecurity and unemployment predict cardiometabolic risk (Drake et al., 2021)
- EHR models that include social determinants improve risk prediction for vulnerable patients (Howell et al., 2025).

SES group

Literature Review

- Income and education levels change how behaviors and barriers translate into disease.
  - Lower-SES individuals face compounded effects of stress, access issues, and less healthy environments (Liu et al., 2023)
  - For example, smoking in a high-income, educated person may carry a lower cardiometabolic risk (due to better care access) than in a low-income person with the same behavior.
- Low-SES groups face “double jeopardy”: more barriers and more stress (Liu et al., 2023)
- High-stress latent profiles were disproportionately non-Hispanic Black and Hispanic. These groups had 2–4 times higher odds of having unmet needs than whites, contributing to elevated cardiometabolic burden (Fernandez et al., 2022)

2. Data Source: Kaggle Diabetes Dataset

Source:
Kaggle Dataset: Diabetes Dataset with 18 Features

- **Dataset Size: 100,000 records**
- **Features: 18 clinical and behavioral variables related to diabetes**

Key Variables Included:

- **Metabolic Indicators: Glucose, Blood Pressure, Insulin, BMI, Skin Thickness**
- **Demographic Factors: Age, Gender**
- **Behavioral Factors: Smoking, Alcohol Consumption**
- **Outcome: Type 2 Diabetes (Diagnosed or Not)**

Link to the Dataset -
https://www.kaggle.com/datasets/pkdarabi/diabetes-dataset-with-18-features/data

Literature Review

Physical Activity and Psychosocial Factors

- Among both adolescents and adults, physical activity is linked to reduce depressive symptoms and lower psychosocial distress (White et al., 2024).
  - This is contingent on how severe the depressive symptoms and mental health conditions are.
- A peer-reviewed cross-sectional study conducted with data from the CDC's Behavioral Risk Factors Surveillance System survey found that, on average, individuals have 3.4 poor mental health days per month.
  - However, those who exercise regularly had their average poor mental health days reduced by 40% (UCLA Health, 2018).

References

Berkowitz, S. A., Ochoa, A., LaPoint, M. L., Kuhn, M. L., Dankovchik, J., Donovan, J. M., … Gold, R. (2024). Transportation barriers and diabetes outcomes: A longitudinal analysis. *Journal of Primary Care & Community Health*. https://doi.org/10.1177/21501319231203650

Drake, C. P., Webb Hooper, M., Tseng, C.-H., & Tsai, J. (2021). Evaluating the association of social needs assessment data with cardiometabolic health status in a federally qualified community health center patient population. *BMC Cardiovascular Disorders, 21*(1), 410. https://doi.org/10.1186/s12872-021-02149-5

Fernandez, J. R., Montiel Ishino, F. A., Montiel Ishino, A. Y., & Tseng, C.-H. (2022). Hypertension and diabetes status by patterns of stress in older adults from the US Health and Retirement Study: A latent class analysis. *Journal of the American Heart Association, 11*(12), e024594. https://doi.org/10.1161/JAHA.121.024594

Liu, C., He, L., Li, X., Yang, K., Zhang, Y., & Luo, H. (2023). Diabetes risk among US adults with different socioeconomic status and behavioral lifestyles: Evidence from NHANES. *Nutrients, 15*(8), 3990. https://doi.org/10.3390/nu15081990

Howell, C. R., Tanaka, S., Zhang, L., & Cherrington, A. L. (2025). Adding social determinants of health to the equation: Development of a cardiometabolic disease staging model to predict type 2 diabetes. *Diabetes, Obesity and Metabolism, 27*(5), 2454–2462. https://doi.org/10.1111/dom.16241

DATA FILES AND SOURCES

https://www.cdc.gov/brfss/annual_data/annual_2015.html

https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system/data

https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data