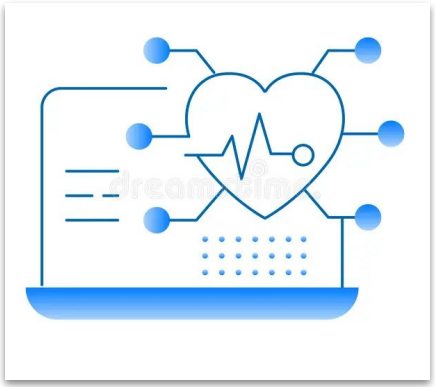
AI Colab Group 1 - Clinical Data Science & Modeling

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

Among U.S. adults with type 2 diabetes, how is kidney function, as measured by estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN), associated with β-cell dysfunction, and how do smoking and alcohol behaviors influence this relationship?

Potential Hypothesis

Among U.S. adults with type 2 diabetes, reduced kidney function — indicated by lower eGFR and higher BUN levels — is associated with decreased β-cell function (as measured by HOMA-B). This association is stronger among individuals who currently smoke or consume alcohol more frequently.

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)

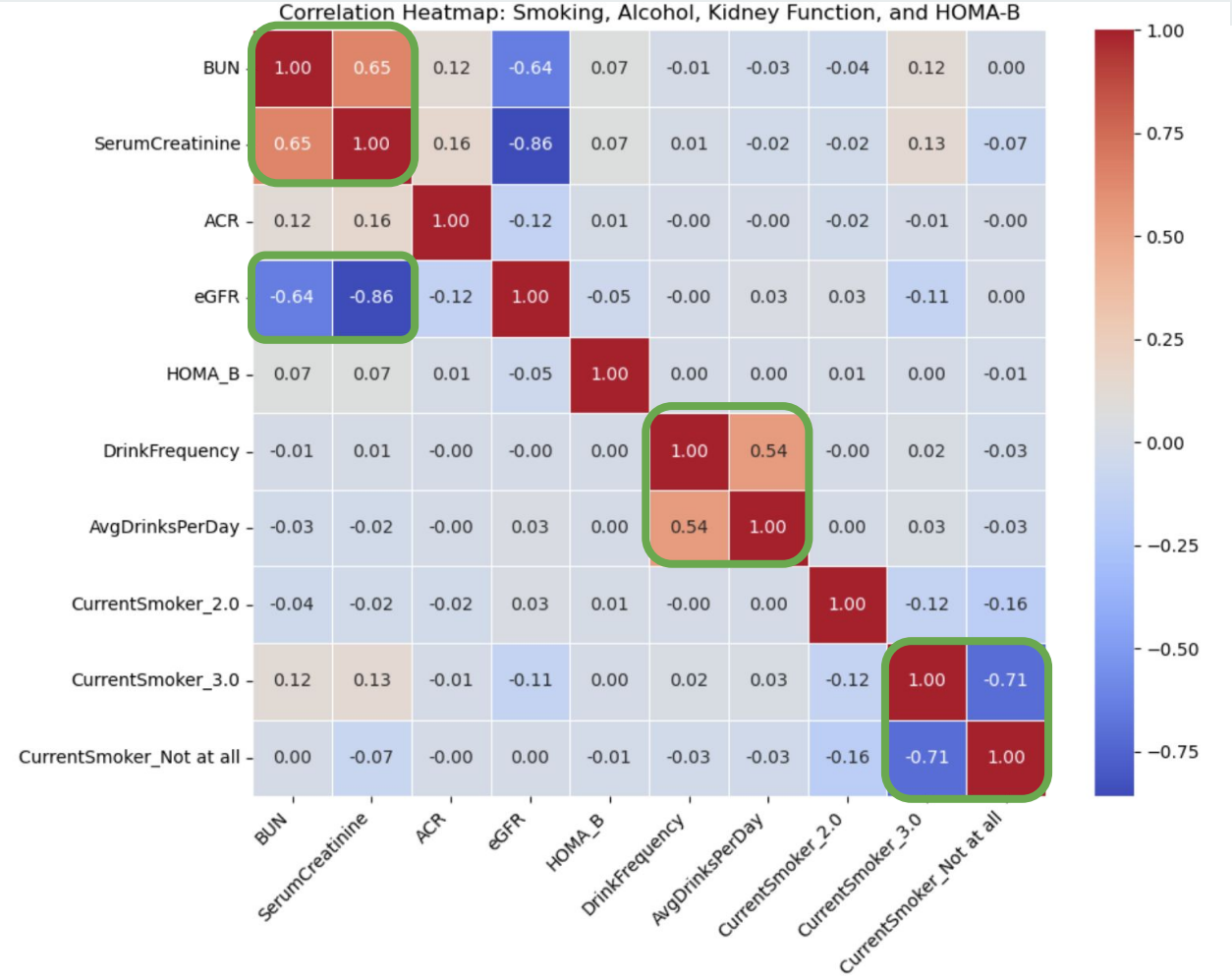
DATA DICTIONARY AFTER CLEANING

	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.

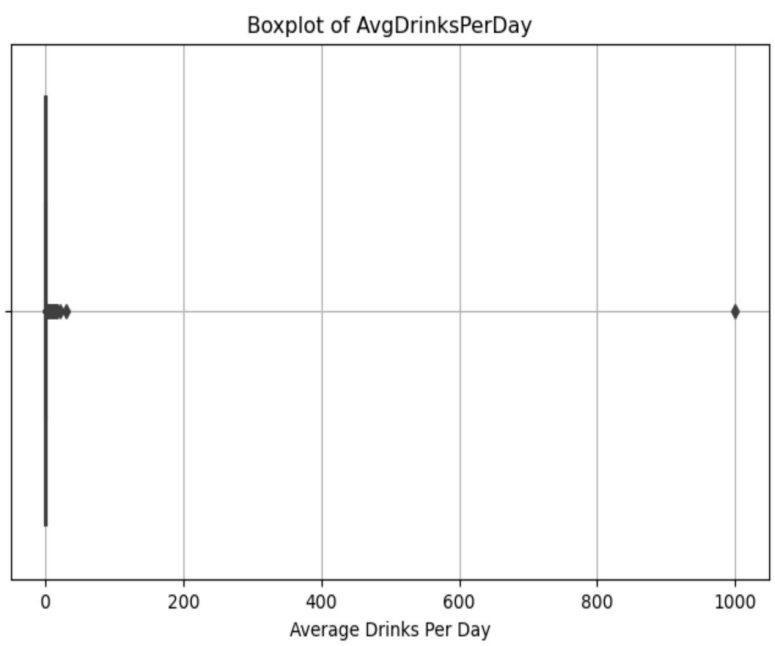
After some further analysis, many features will be dropped.

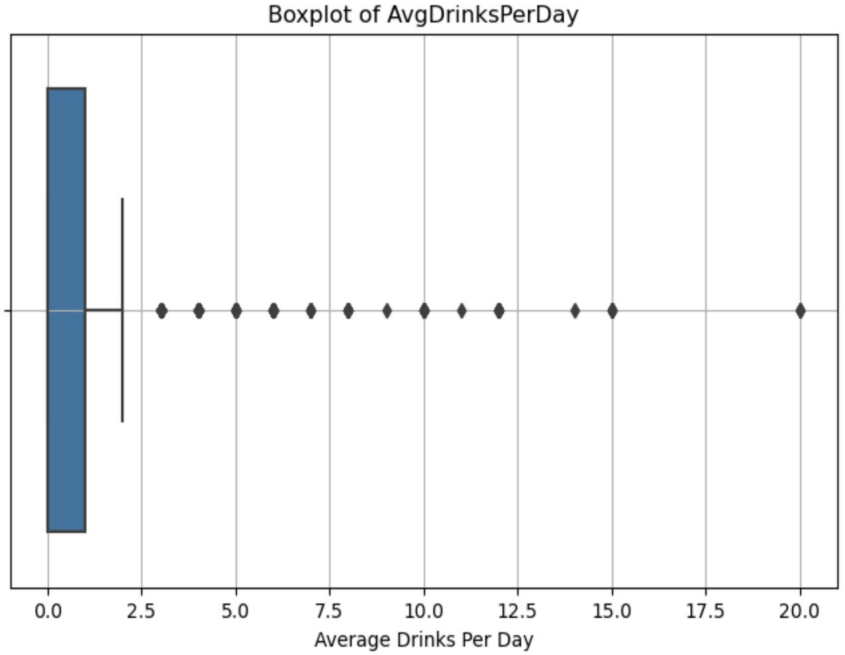
DATA PREP FOR MODELING



Dropped Features:
'SerumCreatinine'
'CurrentSmoker_Not at_all'
'DrinkFrequency'

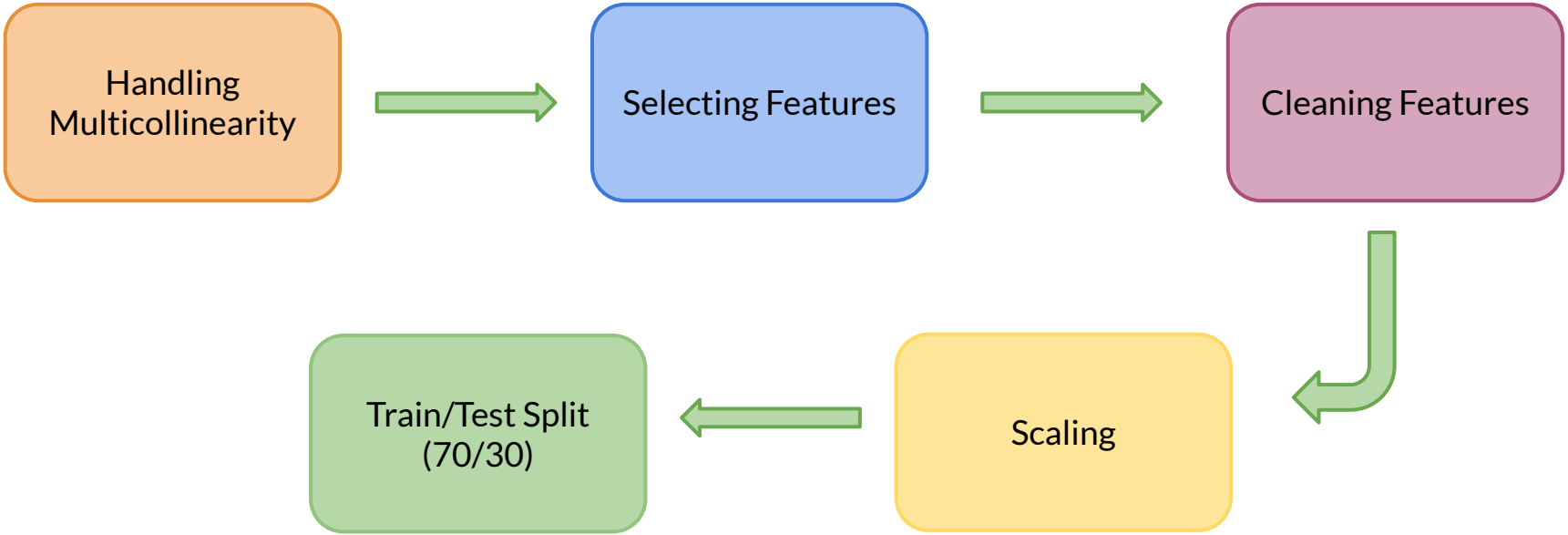
DATA PREP FOR MODELING





Capped Average drinks at 20

DATA PREP FOR MODELING



DATA DICTIONARY BEFORE AFTER

	Feature Name	Data Type	Missing Values	Unique Values	Description
0	BUN	float64	0	51	Blood urea nitrogen (mg/dL), marker of kidney ...
1	AvgDrinksPerDay	float64	0	16	Average number of alcoholic drinks per day ove...
2	Age	float64	0	68	Age in years at time of screening
3	Gender	float64	0	2	Gender (1 = Male, 2 = Female)
4	Education	float64	0	7	Education level (1 = Less than 9th grade to 5 ...
5	IncomeToPovertyRatio	float64	0	464	Ratio of family income to poverty level
6	HOMA_B	float64	0	2920	No description available
7	CurrentSmoker_2.0	int64	0	2	Current smoking status: some days (dummy varia...
8	CurrentSmoker_3.0	int64	0	2	Current smoking status: not at all (dummy vari...
9	SurveyCycle_2009-2010	int64	0	2	No description available
10	EverDrank_Label_Drinks Alcohol	int64	0	2	Label indicating whether participant drinks al...
11	eGFR	float64	0	2233	Estimated glomerular filtration rate (mL/min/1...
12	ACR_log	float64	0	2534	Log of ACR to calculate urine albumin-to-creat...

MODEL VISUALIZATIONS

Linear Model

```r
model <- lm(HOMA_B ~ BUN + CurrentSmoker_2.0 + eGFR +
              CurrentSmoker_3.0 + EverDrank_Label_Drinks.Alcohol +
              Age + Gender + Education + IncomeToPovertyRatio,
            data = df)
summary(model)

predicted <- predict(model, newdata = df)

# Calculate MSE (Mean Squared Error)
mse_value <- mean((df$HOMA_B - predicted)^2)
cat("MSE:", mse_value, "\n")
```

Result

```
Multiple R-squared:  0.01518,   Adjusted R-squared:  0.01233
```

```
MSE: 27.53435
```

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.295165   1.287933   4.888 1.07e-06 ***
BUN                            0.049987   0.018806   2.658   0.0079 **
CurrentSmoker_2.0              0.374973   0.600312   0.625   0.5323
eGFR                          -0.017621   0.006586  -2.676   0.0075 **
CurrentSmoker_3.0              0.136747   0.208495   0.656   0.5120
EverDrank_Label_Drinks.Alcohol 0.502239   0.203238   2.471   0.0135 *
Age                           -0.046862   0.009691  -4.836 1.39e-06 ***
Gender                         0.059841   0.196443   0.305   0.7607
Education                      0.042367   0.076222   0.556   0.5784
IncomeToPovertyRatio          -0.080492   0.072203  -1.115   0.2650
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

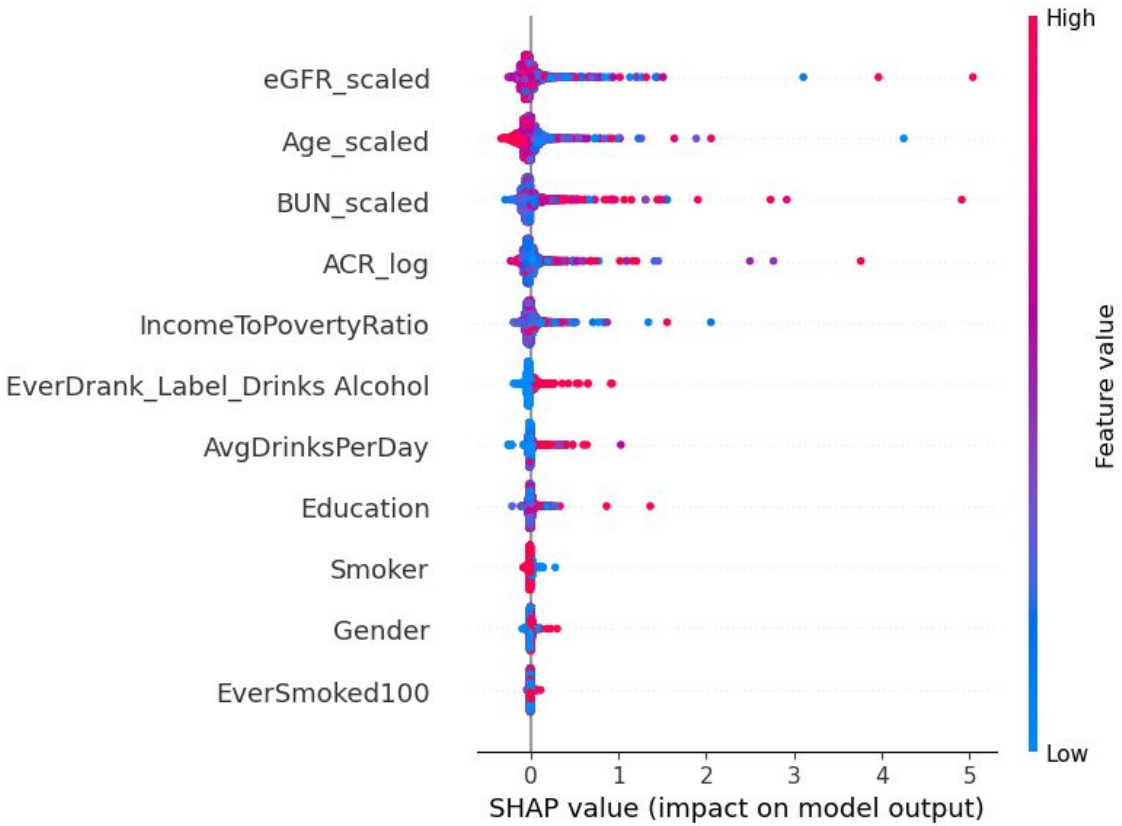
Random Forest Model

```r
install.packages("randomForest")
library(randomForest)

predictors <- c("BUN", "CurrentSmoker_2.0", "CurrentSmoker_3.0","eGFR",
                "AvgDrinksPerDay", "Gender", "Age", "Education", "IncomeToPovertyRatio")
outcome <- "HOMA_B"

df_rf <- df[, c(outcome, predictors)]
df_rf <- na.omit(df_rf)

set.seed(123)
rf_model <- randomForest(HOMA_B ~ ., data = df_rf, importance = TRUE, ntree = 500)

print(rf_model)

predicted_rf <- predict(rf_model, newdata = df_rf)

mse_rf <- mean((df_rf$HOMA_B - predicted_rf)^2)
r2_rf <- 1 - (sum((df_rf$HOMA_B - predicted_rf)^2) / sum((df_rf$HOMA_B - mean(df_rf$HOMA_B))^2))

cat("Random Forest MSE:", mse_rf, "\n")
cat("Random Forest R²:", r2_rf, "\n")
```

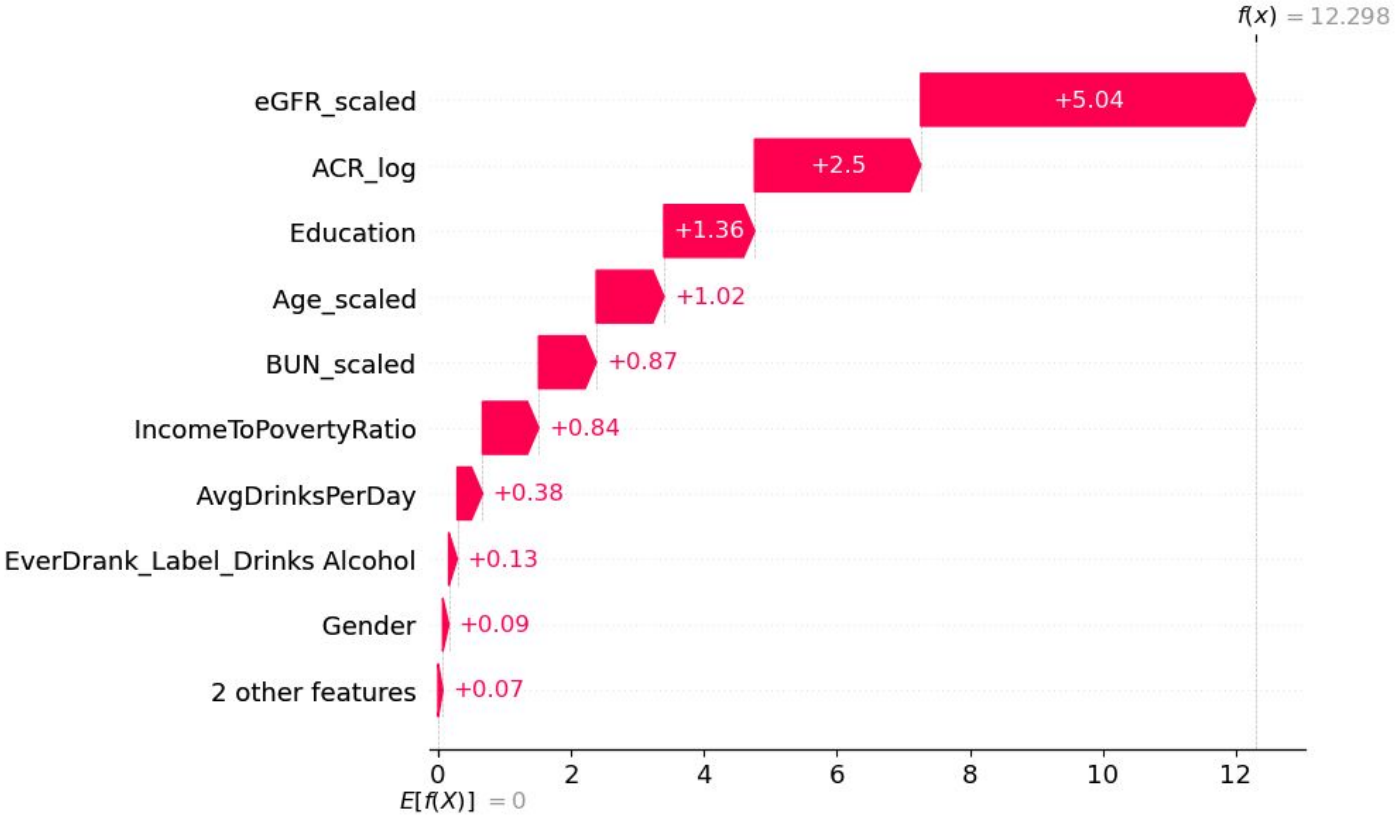
Results

```
> r2_rf <- 1 - (sum((df_rf$HOMA_B - predicted_rf)^2) / sum((df_rf$HOMA_B - mean(df_rf$HOMA_B))^2))
> cat("Random Forest MSE:", mse_rf, "\n")
Random Forest MSE: 8.454876
> cat("Random Forest R²:", r2_rf, "\n")
Random Forest R²: 0.6975955
```

SHAP-Based Model Interpretation



SHAP Waterfall Plot: Individual Prediction Breakdown



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