Predicting Diabetes Status

Ryan Zomorrodi

Introduction

Diabetes is a significant global health challenge which affects millions of individuals and places immense strain on healthcare systems. In the United States alone, over 38.1 million adults—approximately 14.7% of adults—are estimated to have diabetes, with an additional 97.6 million adults (38%) having prediabetes (CDC, 2024a) , a condition that increases the risk of developing diabetes and related complications if left unmanaged. The burden of diabetes is profound, encompassing increased risks of cardiovascular disease, kidney failure, blindness, and amputations, alongside substantial economic costs. According to the American Diabetes Association, the total estimated cost of diagnosed diabetes in the U.S. in 2022 was $412.9 billion, a figure that continues to rise (Parker et al., 2024).

Accurate prediction tools are critical for the prevention and early management of diabetes. Timely identification of individuals at high risk enables healthcare providers and public health organizations to deploy targeted interventions, such as lifestyle modification programs, nutritional counseling, and regular glucose monitoring, which have been shown to delay or even prevent the onset of type 2 diabetes (An et al., 2024; CDC, 2024b; Shubrook et al., 2018). Others have pointed toward the need for "personal profiling" predictive tools utilizing clinical, anthropometric, and behavioral data to identify the right strategy for early intervention on an individual basis (Tuomilehto & Schwarz, 2016). These tools can be used to identify high-risk individuals who may not yet exhibit symptoms or have access to routine screenings, thereby bridging gaps in care. Additionally, predictive models can help optimize resource allocation by focusing preventive efforts on populations with the greatest need, ultimately reducing the overall burden of diabetes and its associated complications.

This study aims to develop and evaluate predictive models for diabetes status using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey responses. Predictors utilized can be gathered without the need for clinical staff or laboratory testing. This approach is especially advantageous for individuals with limited access to healthcare systems or those who experience barriers to regular medical check-ups. By using easily obtainable and non-invasive data, these predictive algorithms can empower individuals to assess their risk independently or through community-based programs. This accessibility not only facilitates early identification of high-risk individuals but also promotes health equity by extending the reach of preventive care to underserved and remote populations.

Methods

Data Source

The BRFSS is an ongoing, state-based, random digit–dialed telephone survey of non-institutionalized US adults aged 18 years or older. The BRFSS collects extensive data on health-related risk behaviors, chronic conditions, and the use of preventive services. For our efforts, we utilized a pre-cleaned dataset of 2015 BRFFS survey responses subset to include measures typically thought of as relevant to diabetes status.

The dataset utilized comprises 253,680 observations, with each record corresponding to an individual respondent's health survey data. The primary outcome variable is binary, indicating whether the individual has been diagnosed with diabetes or prediabetes or whether the individual has not. 35,346 (13.9%) individuals reported having diabetes or prediabetes. In addition to the outcome variable, the dataset includes 21 predictor variables capturing a diverse range of demographic, behavioral, and health-related characteristics.

Demographic information in the dataset includes variables such as age, gender, education level, and income. Behavioral factors encompass indicators of lifestyle choices, such as physical activity, frequency of fruit and vegetable consumption, and alcohol consumption patterns. Health-related factors include self-reported conditions such as high blood pressure, high cholesterol, history of stroke or heart disease, and Body Mass Index (BMI).

The dataset also includes variables that assess healthcare access and utilization, such as whether the respondent has access to any healthcare services and whether they have avoided seeking medical care due to cost constraints. Furthermore, indicators of general health status are included, such as self-reported overall health (measured on an ordinal scale) and the number of days in the past month that mental or physical health was considered poor. Functional limitations, such as difficulty walking or climbing stairs, were also captured.

Data Analysis

To facilitate exploratory data analysis, continuous features were binned into discrete categories based on clinically relevant thresholds. After preprocessing the data, unadjusted odds ratios (ORs) were calculated with a univariate logistic regression model for each feature to assess their individual association with diabetes status. A multivariate logistic regression model was then employed to compute adjusted odds ratios (aORs), controlling for all other features. These adjusted estimates provide insights into the independent effects of each predictor on diabetes risk while accounting for interactions among variables. The results of this analysis are summarized in Table 3.

The dataset was acquired pre-cleaned and no additional selection criteria on subjects or features was conducted. 75% of the data was allocated to the training set for model development and hyperparameter tuning, while the remaining 25% was reserved as the validation set to assess the tuned final models' performance on unseen data. In order to reduce the dimentionality of the data, the age and income features were collapsed (<= 29, 30-39, 40-49, 50-59, >= 60; < $25,000, $25,000 - $74,999, >= 75,000). Given the imbalanced nature of the outcome variable, where the prevalence of prediabetes and diabetes was substantially lower than the absence of diabetes, downsampling

was applied to match the number of prediabetic and diabetic observations to non-diabetic observations.

Five supervised machine learning algorithms were screened: elastic net logistic regression, random forest, k-nearest neighbors (KNN), naive Bayes, and LightGBM. For each model, 25 candidate parameter combinations were systematically generated to explore a range of hyperparameter settings. These combinations were evaluated using 5-fold cross-validation and a range of classification metrics including accuracy, sensitivity, specificity, receiver operating characteristic area under the curve, and mean log loss. For each model, the candidate parameter combination with the highest accuracy was chosen. Metrics for cross validated metrics are summarized in Table 1.

Final validation of selected models was conducted by training on the entire training dataset and testing on the validation set. Metrics for this final validation are summarized in Table 2.

All analyses were preformed using R 4.4.1. The `tidymodels` framework was used to preform hyperparameter tuning and model selection (Kuhn et al., 2024).

Results

Risk Factor Analysis

For most variables, unadjusted odds ratios suggest strong associations with diabetes status, but these associations are often attenuated after adjusting for confounding factors. High blood pressure, high cholesterol, and having had a cholesterol check were significantly associated with diabetes status, with adjusted ORs of 1.98, 1.69, and 3.39, respectively, indicating a substantial increase in diabetes risk for individuals with these conditions. Behavioral factors such as physical activity and consumption of fruits and vegetables look to be protective when unadjusted. However, the adjusted ORs approach 1, suggesting modest or no independent protective effects after adjustment.

General health and difficulty walking showed strong associations with diabetes status. Poor general health had an adjusted OR of 7.35, highlighting a sharp increase in diabetes risk as self-reported health declines. Difficulty walking was also significantly associated with diabetes status, albeit to a lesser extent (adjusted OR = 1.14). Increasing age was strongly associated with diabetes risk, with the adjusted OR rising progressively throughout the age groups, tailing off slightly among those 75+. Individuals within the 70-74 age group had an adjusted OR of 7.32 as the compared to the 18-24 reference group. Education and income levels demonstrated inverse relationships with diabetes status, where higher levels of education and income were associated with lower diabetes risk.

Body Mass Index (BMI) was a particularly strong predictor of diabetes status, with higher BMI categories showing progressively elevated odds. For example, individuals with Class III obesity had an adjusted OR of 6.95 compared to those with normal weight. Factors such as mental and physical health days, measured as days of poor health in the past month, also revealed significant associations. Interestingly, unadjusted odds ratios suggested that more bad mental/physical days were associated with higher odds of diabetes, but after adjustement, bad mental/physical days were weakly associated with lower odds of diabetes.

Predictive Models

After final validation, naive Bayes achieved the highest accuracy (0.821) and sensitivity (0.876) among the models, but its specificity (0.485) was the lowest, indicating a trade-off in correctly identifying non-diabetic cases. The decision tree model achieved the second highest accuracy (0.767), but achieved a considerably higher specificity (0.613). Logistic regression and LightGBM demonstrated balanced performance, with accuracies of 0.734 and 0.749, respectively, and ROC AUC values of 0.823 and 804. Random forest showed slightly lower accuracy (0.724) but performed comparably in sensitivity and specificity, with a ROC AUC of 0.820.

Table 1: Five-Fold Cross-Validated Metrics by Model

	Accuracy	Sensitivity	Specificity	J-Index	Mean Log Loss	ROC AUC
Naive Bayes	0.8201671	0.8737217	0.4890038	0.5088522	0.3627255	0.8102471
Decision Tree	0.7644802	0.7883731	0.6166803	0.5956696	0.4050534	0.7269096
LightGBM	0.7372570	0.7438161	0.6966257	0.6931471	0.4404419	0.7985818
Logistic Regression	0.7349547	0.7296412	0.7678344	0.5244064	0.4974756	0.8238299
Random Forest	0.7244376	0.7173705	0.7681321	0.5688123	0.4855026	0.8182843

Table 2: Final Validation Metrics by Model

	Accuracy	Sensitivity	Specificity	J-Index	Mean Log Loss	ROC AUC
Naive Bayes	0.8212315	0.8758763	0.4846176	0.3604938	0.4928560	0.8119004
Decision Tree	0.7674629	0.7924857	0.6133220	0.4058077	0.5925925	0.7281168
LightGBM	0.7485415	0.7588545	0.6850127	0.4438672	0.6931471	0.8038264
Logistic Regression	0.7335225	0.7288431	0.7623483	0.4911914	0.5226951	0.8233297
Random Forest	0.7244560	0.7165178	0.7733559	0.4898737	0.5695660	0.8196836

Discussion

The observation that adjusted ORs for many covariates in Table 3 are lower than their corresponding unadjusted ORs suggests that some of the apparent effects of these variables on diabetes status are partially explained by associations with other variables in the model. In other words, the unadjusted ORs may reflect not only the direct association between a given covariate and diabetes but also indirect effects mediated through other covariates. This is to be expected as prior research has demonstrated associations between variables such as hypertension and income (Anstey et al., 2019), physical activity and heart disease (National Heart Lung and Blood Institute, 2022), or educational attainment and obesity (Cohen et al., 2013). Furthermore, there is light correlation between the continuous covariates (BMI, number of bad physical health days within last 30, and number of bad mental health days within last 30) as can be seen in Figure 1. This finding is par-

ticularly relevant to the use of the naive Bayes model due to it's independence assumption, and casts doubt on it's application within this setting.

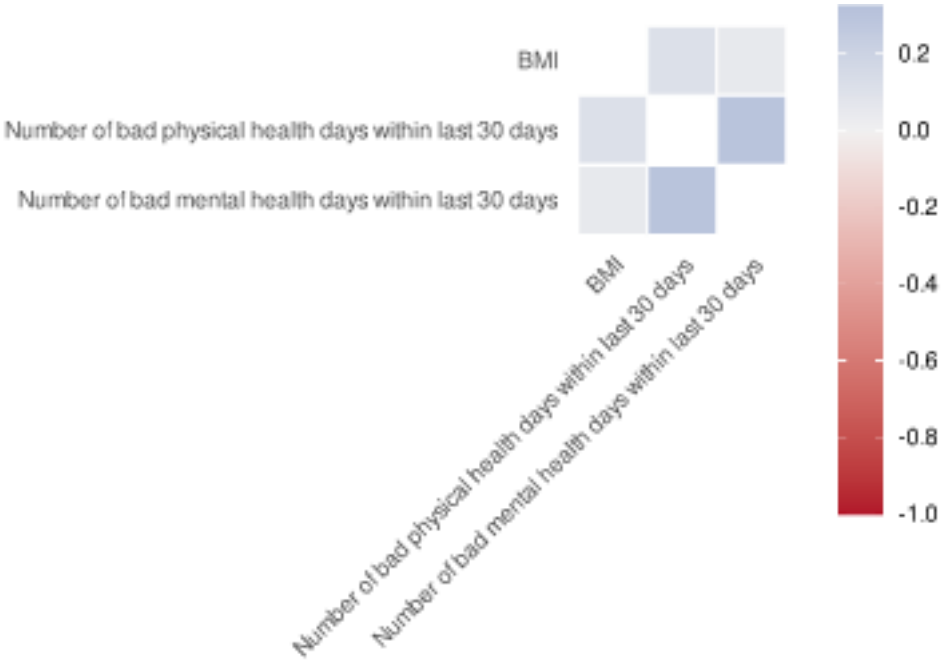

Figure 1: Association Between Continuous Covariates

Although the naive Bayes model yielded the highest accuracy of all models screened, it performed especially poor at predicting non-diabetes outcomes. This low sensitivity can be directly attributed to the model's inability to account for associations between the covariates, which causes it to overpredict prediabetes and diabetes cases.

In contrast, the decision tree model had a substantially lower accuracy, but a considerably better specificity (although its specificity was still second lowest of all models). The decision tree model demonstrated the lowest ROC AUC (area under the receiver operating characteristic curve) among the models tested, indicating its limited ability to distinguish between positive and negative cases across different classification thresholds. The decision tree bases its predictions on just six variables, In order of importance: high blood pressure, general health, age, high cholesterol, BMI, and income (See Figure 2).

Light GBM had an accuracy of 0.749, but a specificity of only 0.685. While the model is substantially better at predicting prediabetes and diabetes cases among those who have prediabetes or diabetes than predicting non-diabetes cases among those who do not have diabetes, it's sensitivity is still higher than that of the naive Bayes or Decision Tree models. Like the decision tree model, high blood pressure and general health were the two most important variables. BMI and high cholesterol are still fairly importance within the model, but age is not especially important in Light GBM (See Figure 4).

Logistic regression had an accuracy near that of the Light GBM model, but a lower sensitivity and higher specificity. Logistic regression had the highest ROC AUC of any of the models screened. Like the decision tree model, BMI, general health, and high blood pressure, and age had the largest effect on prediction outcome (See Figure 5). Random Forest displayed a balanced performance with relatively high specificity (0.7734) but did not excel in any specific metric, showing consistent but unremarkable results. General health, high blood pressure, BMI, high cholesterol, and age are the most important variables within the model (See Figure 3). These variables mirror those included in the decision tree.

Overall, all models indicated that high blood pressure, general health, high cholesterol and BMI were important predictors of diabetes status. Interestingly, demographic data like income and sex were not particularly strong predictors of diabetes status. Additionally, healthcare access seem to have very little predictive power. In fact, the coefficients for both healthcare coverage and inability to see a doctor due to cost were reduced to 0 within the logistic regression model. Likewise, smoking and alcohol status were not among the top predictors for any of the models.

The fact that high cholesterol and high blood pressure were important predictors seems to indicate that lab testing results may be good predictors diabetes status. While this study explicitly used data that does not require clinical staff or laboratory testing to collect, the results indicate that a model utilizing laboratory testing data may be promising. Further studies should be explored to consider the potential extra predictive power that routine laboratory testing may yield.

This study utilized survey data to predict diabetes status. While survey data offers the advantage of capturing a wide array of demographic, behavioral, and health-related factors, it also comes with inherent limitations. One major limitation is the cross-sectional nature of the data, which prevents the establishment of temporal or causal relationships between predictors and diabetes status. As such, the analysis can identify associations but cannot determine whether certain factors contribute to the development of diabetes or are consequences of the condition. Additionally, the reliance on self-reported information introduces potential biases, which may affect the accuracy of responses. Individuals with low access to healthcare resources are particularly of concern because indviduals with low access may go undiagnosed at higher rates, and thus, models maybe worse at predicting diabetes status among this population. These limitations highlight the need for caution when interpreting the results and underscore the importance of supplementing survey-based analyses with longitudinal or clinically verified data in future research.

Conclusion

The results indicate that high blood pressure, general health, high cholesterol and BMI were good predictors of diabetes status. Additionally, although the naive Bayes model yielded the highest accuracy, it's unrealistic independence assumptions led to low specificity, which makes it unsuitable as a model for this data. LightGBM may be a good choice for this specific data because it yielded moderately accuracy as compared to the others screened, but fairly balanced prediction accuracy. Logistic regression must also be considered for its increased interpretability when compared to LightGBM and relatively similar performance.

Future studies should consider including individual-level laboratory testing data from a representative sample of the population. Likewise, further exploration of non-testing data is necessary to further evaluate whether variable omitted from this dataset like race, hours of sleep, family history, geographic location, etc. may be additionally helpful in predicting diabetes status. It may even be advantageous to utilize a multiclass outcome to distinguish between diabetes, pre-diabetes, and no diabetes, but future studies would have to evaluate such a question.

Appendix

Figures and Tables

Calculating odds ratios

Table 3: Association Between Covariates and Diabetes Status

	Odds Ratio (95% Confidence Interval)	
	Unadjusted	Adjusted
High Blood Pressure		
No		1 (Reference)
Yes	5.04 (4.91-5.17)	1.98 (1.92-2.04)
High Cholesterol		
No		1 (Reference)
Yes	3.25 (3.18-3.33)	1.69 (1.65-1.74)
Cholesterol Check		
No		1 (Reference)
Yes	6.43 (5.65-7.31)	3.39 (2.96-3.87)
Smoker		
No		1 (Reference)
Yes	1.42 (1.39-1.45)	0.97 (0.94-0.99)
Stroke		
No		1 (Reference)
Yes	3.06 (2.94-3.20)	1.19 (1.13-1.25)
Heart Disease or Attack		
No		1 (Reference)
Yes	3.62 (3.52-3.73)	1.29 (1.24-1.33)
Physically Active		
No		1 (Reference)

	Odds Ratio (95% Confidence Interval)	
	Unadjusted	Adjusted
Yes	0.49 (0.48-0.50)	0.95 (0.93-0.98)
Consumes Fruit		
No		1 (Reference)
Yes	0.79 (0.77-0.81)	0.98 (0.95-1.01)
Consumes Vegetables		
No		1 (Reference)
Yes	0.68 (0.66-0.70)	0.97 (0.94-1.00)
Heavy Alcohol Consumption		
No		1 (Reference)
Yes	0.37 (0.34-0.40)	0.47 (0.44-0.51)
Healthcare Coverage		
No		1 (Reference)
Yes	1.27 (1.20-1.34)	1.07 (1.00-1.14)
Inable to see doctor due to cost		
No		1 (Reference)
Yes	1.35 (1.30-1.40)	1.00 (0.96-1.05)
General Health		
Excellent		1 (Reference)
Very good	2.99 (2.80-3.19)	1.98 (1.85-2.12)
Good	8.38 (7.88-8.92)	3.92 (3.68-4.18)
Fair	17.41 (16.34-18.55)	6.08 (5.67-6.52)
Poor	23.63 (22.05-25.33)	7.35 (6.77-7.99)
Difficulty to walk		
No		1 (Reference)
Yes	3.77 (3.68-3.87)	1.14 (1.10-1.18)
Sex		
Female		1 (Reference)
Male	1.20 (1.17-1.23)	1.29 (1.25-1.32)
Age		
18 to 24		1 (Reference)

	Odds Ratio (95% Confidence Interval)	
	Unadjusted	Adjusted
25 to 29	1.35 (1.02-1.79)	1.12 (0.85-1.49)
30 to 34	2.09 (1.63-2.69)	1.53 (1.19-1.98)
35 to 39	3.42 (2.70-4.33)	2.22 (1.75-2.83)
40 to 44	5.01 (3.98-6.32)	2.87 (2.26-3.63)
45 to 49	6.95 (5.53-8.73)	3.47 (2.75-4.38)
50 to 54	9.58 (7.64-12.02)	4.34 (3.45-5.47)
55 to 59	11.56 (9.23-14.49)	4.75 (3.77-5.97)
60 to 64	15.02 (11.99-18.81)	5.86 (4.66-7.38)
65 to 69	18.44 (14.72-23.09)	6.94 (5.52-8.73)
70 to 74	20.15 (16.08-25.24)	7.32 (5.82-9.22)
75 to 79	19.50 (15.55-24.46)	6.83 (5.41-8.61)
80 to 99	16.34 (13.03-20.50)	5.86 (4.65-7.39)
Education		
No school		1 (Reference)
Elementary	1.12 (0.79-1.57)	0.97 (0.66-1.42)
Some high school	0.86 (0.62-1.21)	0.86 (0.59-1.25)
High school graduate	0.58 (0.41-0.81)	0.81 (0.55-1.18)
Some college or technical school	0.47 (0.34-0.66)	0.84 (0.58-1.23)
College graduate	0.29 (0.21-0.41)	0.78 (0.54-1.14)
Income		
Less than $10,000		1 (Reference)
Less than $15,000	1.11 (1.04-1.18)	0.98 (0.92-1.05)
Less than $20,000	0.90 (0.84-0.95)	0.95 (0.89-1.02)
Less than $25,000	0.79 (0.74-0.83)	0.93 (0.87-0.99)
Less than $35,000	0.66 (0.62-0.69)	0.85 (0.80-0.91)
Less than $50,000	0.53 (0.50-0.56)	0.78 (0.73-0.83)
Less than $75,000	0.43 (0.41-0.46)	0.76 (0.71-0.81)
$75,000 or more	0.27 (0.26-0.28)	0.66 (0.62-0.71)
Bad mental health days within last 30 class		

	Odds Ratio (95% Confidence Interval)	
	Unadjusted	Adjusted
Less than 5		1 (Reference)
Between 5 and 9	1.07 (1.02-1.13)	0.92 (0.87-0.97)
Between 10 and 19	1.42 (1.35-1.48)	0.93 (0.88-0.98)
Greater than or equal to 20	1.89 (1.82-1.96)	0.94 (0.90-0.99)
Bad physical health days within last 30 class		
Less than 5		1 (Reference)
Between 5 and 9	1.71 (1.63-1.79)	0.96 (0.91-1.01)
Between 10 and 19	2.45 (2.36-2.56)	0.99 (0.95-1.04)
Greater than or equal to 20	3.30 (3.20-3.40)	0.92 (0.88-0.96)
BMI class		
Underweight		1 (Reference)
Normal weight	1.06 (0.90-1.24)	1.49 (1.26-1.76)
Overweight	2.25 (1.93-2.63)	2.45 (2.08-2.89)
Obesity (Class I)	4.17 (3.56-4.87)	3.82 (3.24-4.50)
Obesity (Class II)	6.62 (5.65-7.75)	5.43 (4.60-6.41)
Obesity (Class III)	8.83 (7.53-10.35)	6.95 (5.88-8.23)

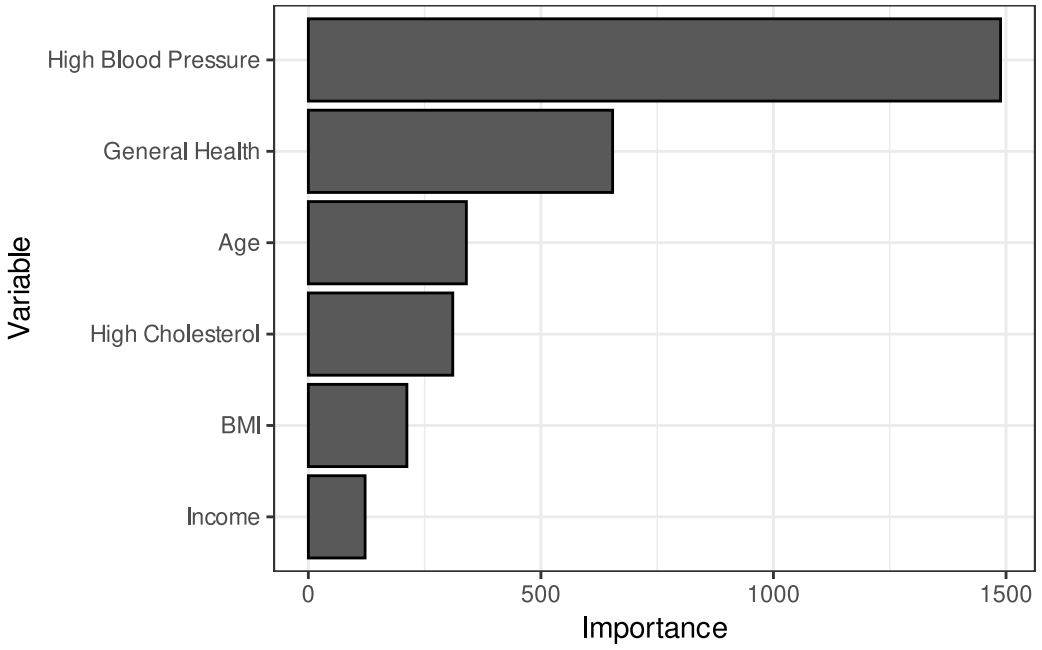

Figure 2: Decision Tree Variable Importance

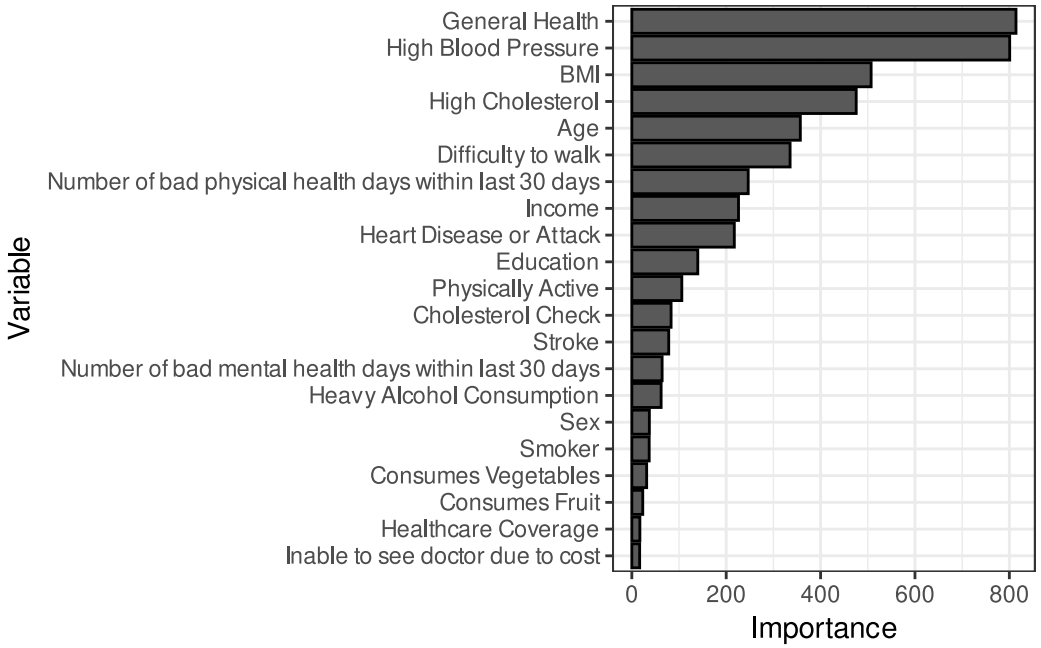

Figure 3: Random Forest Variable Importance

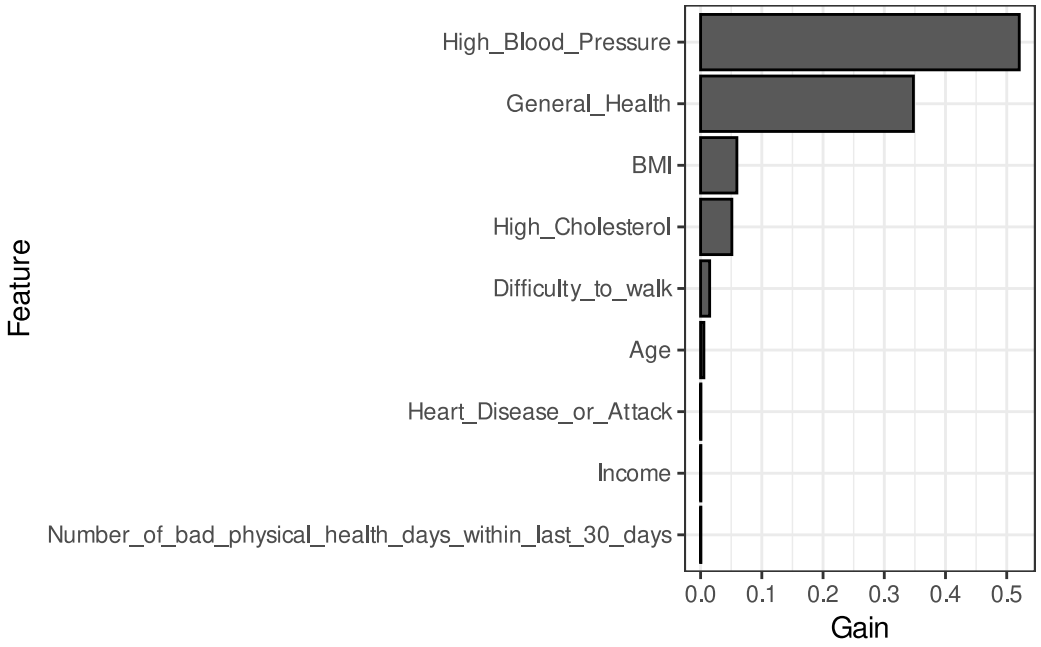

Figure 4: Light GBM Variable Importance

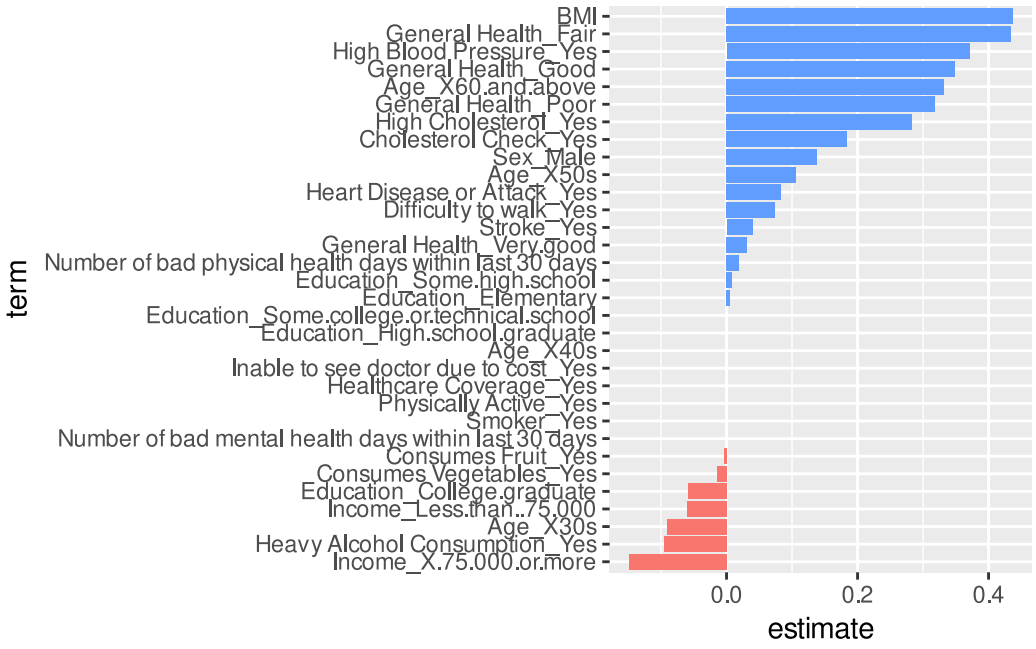

Figure 5: Logistic Regression Normalized Variable Coefficients

R code

```
library(tidyverse)

# modeling
library(tidymodels)
library(finetune) # tune_race_anova
library(future) # multithreading
library(themis) # downsampling
library(discrim) # naive bayes
library(bonsai) # light gbm

# for odds ratio table
library(gt)
library(broom.helpers)
```

Processing data

```
diabetes <- read_csv("../data/Diabetes_BRFSS2015.csv")

diabetes_proc <- diabetes %>%
  mutate(across(
    c(where(~ all(unique(.x) %in% 0:1)), -Sex),
    ~ fct_reorder(case_match(.x, 0 ~ "No", 1 ~ "Yes"), .x)
  )) %>%
  mutate(GenHlth = fct_reorder(case_match(
```

```
    GenHlth,
    1 ~ "Excellent",
    2 ~ "Very good",
    3 ~ "Good",
    4 ~ "Fair",
    5 ~ "Poor"
  ), GenHlth)) %>%
  mutate(Sex = fct_reorder(case_match(Sex, 0 ~ "Female", 1 ~ "Male"), Sex)) %>%
  mutate(Age = fct_reorder(case_when(
    Age == 1 ~ "18 to 24",
    Age < 13 ~ paste(15 + 5 * Age, "to", 19 + 5 * Age),
    Age == 13 ~ "80 to 99"
  ), Age)) %>%
  mutate(Education = fct_reorder(case_match(
    Education,
    1 ~ "No school",
    2 ~ "Elementary",
    3 ~ "Some high school",
    4 ~ "High school graduate",
    5 ~ "Some college or technical school",
    6 ~ "College graduate"
  ), Education)) %>%
  mutate(Income = fct_reorder(case_when(
    Income < 5 ~ paste0("Less than $", Income * 5 + 5, ",000"),
    Income == 5 ~ "Less than $35,000",
    Income == 6 ~ "Less than $50,000",
    Income == 7 ~ "Less than $75,000",
    Income == 8 ~ "$75,000 or more",
  ), Income)) %>%
  mutate(`Bad mental health days within last 30 class` = fct_reorder(case_when(
    MentHlth < 5 ~ "Less than 5",
    MentHlth < 10 ~ "Between 5 and 9",
    MentHlth < 20 ~ "Between 10 and 19",
    MentHlth >= 20 ~ "Greater than or equal to 20"
  ), MentHlth)) %>%
  mutate(`Bad physical health days within last 30 class` = fct_reorder(case_when(
    PhysHlth < 5 ~ "Less than 5",
    PhysHlth < 10 ~ "Between 5 and 9",
    PhysHlth < 20 ~ "Between 10 and 19",
    PhysHlth >= 20 ~ "Greater than or equal to 20"
  ), PhysHlth)) %>%
  mutate(`BMI class` = fct_reorder(case_when(
    BMI < 18.5 ~ "Underweight",
    BMI < 24.9 ~ "Normal weight",
    BMI < 29.9 ~ "Overweight",
    BMI < 34.9 ~ "Obesity (Class I)",
    BMI < 39.9 ~ "Obesity (Class II)",
    BMI >= 40 ~ "Obesity (Class III)"
```

```
  ), BMI)) %>%
  rename(
    Diabetes = Diabetes_binary,
    `High Blood Pressure` = HighBP,
    `High Cholesterol` = HighChol,
    `Cholesterol Check` = CholCheck,
    `Heart Disease or Attack` = HeartDiseaseorAttack,
    `Physically Active` = PhysActivity,
    `Consumes Fruit` = Fruits,
    `Consumes Vegetables` = Veggies,
    `Heavy Alcohol Consumption` = HvyAlcoholConsump,
    `Healthcare Coverage` = AnyHealthcare,
    `Inable to see doctor due to cost` = NoDocbcCost,
    `General Health` = GenHlth,
    `Number of bad mental health days within last 30 days` = MentHlth,
    `Number of bad physical health days within last 30 days` = PhysHlth,
    `Difficulty to walk` = DiffWalk
  )
```

Calculating odds ratios

```
diabetes_proc_factor <- diabetes_proc %>%
  select(where(is.factor))

# unadjusted
formulas <- str_c("Diabetes ~ `", colnames(diabetes_proc_factor)[-1], "`") %>%
  map(as.formula)
or_tibble <- map(formulas, function(formula) {
  model <- glm(formula, family = binomial, diabetes_proc_factor)

  ci <- exp(confint.default(model))
  colnames(ci) <- c("lower_ci", "upper_ci")

  model %>%
    tidy_and_attach(exponentiate = TRUE, conf.int = FALSE) %>%
    bind_cols(ci) %>%
    tidy_add_reference_rows() %>%
    tidy_add_estimate_to_reference_rows() %>%
    tidy_add_term_labels() %>%
    tidy_remove_intercept() %>%
    select(variable, label, estimate, lower_ci, upper_ci)
}) %>%
  bind_rows()

# adjusted
aor_model <- diabetes_proc_factor %>%
  glm(formula = Diabetes ~ ., family = binomial)
```

```r
aor_ci <- exp(confint.default(aor_model))
colnames(aor_ci) <- c("lower_ci", "upper_ci")

aor_tibble <- aor_model %>%
  tidy_and_attach(exponentiate = TRUE, conf.int = FALSE) %>%
  bind_cols(aor_ci) %>%
  tidy_add_reference_rows() %>%
  tidy_add_estimate_to_reference_rows() %>%
  tidy_add_term_labels() %>%
  tidy_remove_intercept() %>%
  dplyr::select(variable, label, estimate, lower_ci, upper_ci)

# combined
or_tibble_comb <- left_join(
  or_tibble,
  aor_tibble,
  by = join_by(variable, label),
  suffix = c("_or", "_aor")
)


gt(
  or_tibble_comb,
  rowname_col = "label",
  groupname_col = "variable"
) |>
  fmt_number(
    columns = c(ends_with("or"), ends_with("aor"))
  ) |>
  cols_merge(
    columns = c(estimate_or, lower_ci_or, upper_ci_or),
    pattern = "<<{1} ({2}-{3})>>"
  ) |>
  cols_merge(
    columns = c(estimate_aor, lower_ci_aor, upper_ci_aor),
    pattern = "{1}<< ({2}-{3})>>"
  ) |>
  cols_label(
    estimate_or = "Unadjusted",
    estimate_aor = "Adjusted"
  ) |>
  tab_spanner(
    label = "Odds Ratio (95% Confidence Interval)",
    columns = starts_with("estimate")
  ) |>
  sub_values(
    values = 1,
    replacement = "1 (Reference)"
```

```
  ) %>%
  tab_options(table.font.size = 14, table.font.names = "Libertinus Serif")
```

Calculating Continuous Correlation

```
library(corrr)

diabetes_proc %>%
    select(where(is.numeric)) %>%
    correlate(method = "spearman", quiet = TRUE) %>%
    autoplot(triangular = "full")
```

Reducing Dimensionality

```
diabetes_proc <- diabetes_proc %>%
  mutate(Age = fct_collapse(Age,
    `29 and below` = c("18 to 24", "25 to 29"),
    `30s` = c("30 to 34", "35 to 39"),
    `40s` = c("40 to 44", "45 to 49"),
    `50s` = c("50 to 54", "55 to 59"),
    `60 and above` = c("60 to 64", "65 to 69", "70 to 74", "75 to 79", "80 to
99")
  )) %>%
  mutate(Income = fct_collapse(Income,
    `Less than $25,000` = c("Less than $10,000", "Less than $15,000", "Less than
$20,000", "Less than $25,000"),
    `Less than $75,000` = c("Less than $35,000", "Less than $50,000", "Less than
$75,000"),
    `$75,000 or more` = "$75,000 or more"
  ))
```

Modeling

Split data

```
diabetes_split <- diabetes_proc %>%
  select(-ends_with(" class")) %>%
  initial_split()
diabetes_train <- training(diabetes_split)
diabetes_test <- testing(diabetes_split)

diabetes_folds <- vfold_cv(diabetes_train, repeats = 5)
```

Specify Preprocessing

```
basic_recipe <- recipe(Diabetes ~ ., data = diabetes_train) %>%
  step_downsample(Diabetes)
```

```
normalized_recipe <- basic_recipe %>%
  step_dummy(all_factor_predictors()) %>%
  step_normalize(all_predictors())
```

Model Specifications

```
glmnet_spec <- logistic_reg(penalty = tune(), mixture = tune()) %>%
  set_engine("glmnet") %>%
  set_mode("classification")

rf_spec <- rand_forest(mtry = tune(), min_n = tune(), trees = 500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

tree_spec <- decision_tree(
  cost_complexity = tune(),
  tree_depth = tune(),
  min_n = tune()
) %>%
  set_engine("rpart") %>%
  set_mode("classification")

naiveBayes_spec <- naive_Bayes(smoothness = tune(), Laplace = tune()) %>%
  set_engine("naivebayes") %>%
  set_mode("classification")

lgbm_spec <- boost_tree(
  tree_depth = tune(),
  min_n = tune(),
  loss_reduction = tune(),
  sample_size = tune(),
  mtry = tune(),
  learn_rate = tune(),
  trees = 500
) %>%
  set_engine("lightgbm") %>%
  set_mode("classification")
```

Workflows

```
basic_wkfl <- workflow_set(
  preproc = list(basic = basic_recipe),
  models = list(naiveBayes_spec, rf_spec, tree_spec, lgbm_spec)
)

normal_wkfl <- workflow_set(
```

```
  preproc = list(norm = normalized_recipe),
  models = list(glmnet_spec)
)

comb_wkfl <- bind_rows(basic_wkfl, normal_wkfl)
```

Tuning Model

```
plan(multisession)

grid_results <- comb_wkfl %>%
  workflow_map(
    "tune_race_anova",
    seed = 1503,
    resamples = diabetes_folds,
        metrics  =  metric_set(accuracy,  sensitivity,  specificity,  j_index,
mn_log_loss, roc_auc),
    verbose = TRUE,
    grid = 25,
    control = control_race(parallel_over = "everything")
  )
```

Create Cross Validation Table

```
table <- grid_results %>%
    rank_results("accuracy") %>%
     pivot_wider(id_cols = wflow_id, names_from = .metric, values_from = mean)
%>%
    arrange(desc(accuracy))

table <- table %>%
    select(
        Model = wflow_id,
        Accuracy = accuracy,
        Sensitivity = sensitivity,
        Specificity = specificity,
        `J-Index` = j_index,
        `Mean Log Loss` = mn_log_loss,
        `ROC AUC` = roc_auc
    ) %>%
    mutate(Model = case_match(Model,
        "basic_naive_Bayes" ~ "Naive Bayes",
        "basic_decision_tree" ~ "Decision Tree",
        "basic_boost_tree" ~ "LightGBM",
        "norm_logistic_reg" ~ "Logistic Regression",
        "basic_rand_forest" ~ "Random Forest"
    ))
```

```
gt(table, rowname_col = "Model") %>%
  tab_options(table.font.size = 14, table.font.names = "Libertinus Serif")
```

Final Fits

```
final_results <- map(set_names(grid_results$wflow_id), function(wflow_id) {
    best_results <- extract_workflow_set_result(grid_results, wflow_id) %>%
        select_best(metric = "accuracy")

    grid_results %>%
        extract_workflow(wflow_id) %>%
        finalize_workflow(best_results) %>%
        last_fit(
            split = diabetes_split,
             metrics = metric_set(accuracy, sensitivity, specificity, j_index,
mn_log_loss, roc_auc)
        ) %>%
        collect_metrics()
})
```

Create Final Fit Table

```
table2 <- final_results %>%
    bind_rows(.id = "wflow_id") %>%
   pivot_wider(id_cols = wflow_id, names_from = .metric, values_from = .estimate)
%>%
    arrange(desc(accuracy))

table2 <- table2 %>%
    select(
        Model = wflow_id,
        Accuracy = accuracy,
        Sensitivity = sensitivity,
        Specificity = specificity,
        `J-Index` = j_index,
        `Mean Log Loss` = mn_log_loss,
        `ROC AUC` = roc_auc
    ) %>%
    mutate(Model = case_match(Model,
        "basic_naive_Bayes" ~ "Naive Bayes",
        "basic_decision_tree" ~ "Decision Tree",
        "basic_boost_tree" ~ "LightGBM",
        "norm_logistic_reg" ~ "Logistic Regression",
        "basic_rand_forest" ~ "Random Forest"
    ))
```

```
gt(table2, rowname_col = "Model") %>%
  tab_options(table.font.size = 14, table.font.names = "Libertinus Serif")
```

Additional Figures

Decision Tree Variable Importance

```
best_tree <- extract_workflow_set_result(grid_results, "basic_decision_tree")
%>%
        select_best(metric = "accuracy")

tree_fit <- grid_results %>%
    extract_workflow("basic_decision_tree") %>%
    finalize_workflow(best_tree) %>%
    last_fit(split = diabetes_split) %>%
    extract_fit_engine()

tree_imp <- tree_fit$variable.importance %>%
    enframe() %>%
    arrange(value) %>%
    mutate(name = fct_inorder(name)) %>%
    rename(Variable = name, Importance = value)

ggplot(tree_imp) +
    geom_col(aes(y = Variable, x = Importance), col = "black", show.legend = F)
+
    scale_fill_grey() +
    theme_bw()
```

Random Forest Variable Importance

```
library(vip)

best_rf <- extract_workflow_set_result(grid_results, "basic_rand_forest") %>%
    select_best(metric = "accuracy")

rf_wflow <- grid_results %>%
    extract_workflow("basic_rand_forest") %>%
    finalize_workflow(best_rf)

rf_spec <- rf_wflow %>%
    extract_spec_parsnip() %>%
    set_args(importance = "impurity")

rf_fit <- rf_wflow %>%
    update_model(rf_spec) %>%
    last_fit(split = diabetes_split) %>%
```

```
    extract_fit_engine()

rf_imp <- vip(rf_fit, num_features = 40)
rf_imp
```

Light GBM Variable Importance

```
best_lgbm <- extract_workflow_set_result(grid_results, "basic_boost_tree") %>%
        select_best(metric = "accuracy")

lgbm_fit <- grid_results %>%
    extract_workflow("basic_boost_tree") %>%
    finalize_workflow(best_lgbm) %>%
    last_fit(split = diabetes_split) %>%
    extract_fit_engine()

lgbm_imp <- lgb.importance(lgbm_fit) %>%
    arrange(Gain) %>%
    mutate(Feature = fct_inorder(Feature))

ggplot(lgbm_imp) +
    geom_col(aes(y = Feature, x = Gain), col = "black", show.legend = F) +
    scale_fill_grey() +
    theme_bw()
```

Logistic Regression Coefficients

```
best_glmnet <- extract_workflow_set_result(grid_results, "norm_logistic_reg")
%>%
        select_best(metric = "accuracy")

glmnet_fit <- grid_results %>%
    extract_workflow("norm_logistic_reg") %>%
    finalize_workflow(best_glmnet) %>%
    last_fit(split = diabetes_split) %>%
    extract_fit_parsnip()

glmnet_imp <- tidy(glmnet_fit) %>%
    filter(term != "(Intercept)") %>%
    arrange(estimate) %>%
    mutate(term = fct_inorder(term))

ggplot(glmnet_imp) +
    geom_col(
        aes(y = term, x = estimate, fill = as.factor(sign(estimate))),
        show.legend = F,
```

```
        position = "identity"
    )
```

Bibliography

An, X., Zhang, Y., Sun, W., Kang, X., Ji, H., Sun, Y., Jiang, L., Zhao, X., Gao, Q., Lian, F., & Tong, X. (2024). Early effective intervention can significantly reduce all-cause mortality in prediabetic patients: a systematic review and meta-analysis based on high-quality clinical studies. *Frontiers in Endocrinology*, *15*, 1294819–1294820. https://doi.org/10.3389/fendo.2024.1294819

Anstey, D. E., Christian, J., & Shimbo, D. (2019). Income inequality and hypertension control. *Journal of the American Heart Association*, *8*(15), e13636. https://doi.org/10.1161/JAHA.119.013636

CDC. (2024b, May). *Health and economic benefits of diabetes interventions*. https://www.cdc.gov/nccdphp/priorities/diabetes-interventions.html

CDC. (2024a, July). *National diabetes statistics report*. https://www.cdc.gov/diabetes/php/data-research/index.html

Cohen, A. K., Rai, M., Rehkopf, D. H., & Abrams, B. (2013). Educational attainment and obesity: a systematic review. *Obesity Reviews*, *14*(12), 989–1005. https://doi.org/10.1111/obr.12062

Kuhn, M., Wickham, H., & Hvitfeldt, E. (2024). *recipes: Preprocessing and Feature Engineering Steps for Modeling*. https://github.com/tidymodels/recipes

National Heart Lung and Blood Institute. (2022, March). *Physical activity and your heart - benefits*. https://www.nhlbi.nih.gov/health/heart/physical-activity/benefits

Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., & Bannuru, R. R. (2024). Economic costs of diabetes in the u. S. In 2022. *Diabetes Care*, *47*(1), 26–43. https://doi.org/10.2337/dci23-0085

Shubrook, J. H., Chen, W., & Lim, A. (2018). Evidence for the prevention of type 2 diabetes mellitus. *Journal of Osteopathic Medicine*, *118*(11), 730–737. https://doi.org/10.7556/jaoa.2018.158

Tuomilehto, J., & Schwarz, P. E. (2016). Preventing diabetes: early versus late preventive interventions. *Diabetes Care*, *39*(Supplement_2), S115–S120. https://doi.org/10.2337/dcS15-3000