MSDS 607 — Final Project Proposal: Predicting Onset of Diabetes (Pima Indians Dataset)

Taha Malik

2025-12-17

Contents

Executive one-paragraph proposal . . . 1
Motivation and research question . . . 2
Data sources (two different types) . . . 2
Planned data science workflow (OSEMN) . . . 2
Planned transformations (examples) . . . 2
Planned statistical analyses & graphics . . . 2
A novel feature beyond course coverage . . . 2
Initial reproducible code (skeleton + light EDA) . . . 3
Planned model training & evaluation approach . . . 4
Project checklist mapping . . . 5
Deliverables & schedule (suggested) . . . 5
Team / roles . . . 5
Reproducibility, code, and repo . . . 5
Risks and simplifying assumptions . . . 5
What I will demonstrate in the final presentation . . . 5
Appendix — web scraping / metadata (eval = FALSE) . . . 5
References . . . 6

Executive one-paragraph proposal

The goal of this project is to build, evaluate, and explain predictive models that determine whether an individual is likely to have diabetes (Outcome 0/1) using the Pima Indians Diabetes Dataset. I will (1) obtain the provided CSV and augment with supporting metadata scraped from OpenML (to satisfy the multi-source requirement), (2) perform principled cleaning and transformations (treat biologically implausible zeros as missing, impute, create derived features), (3) run exploratory data analysis, (4) train and compare models (logistic regression baseline, random forest, XGBoost) using cross-validated AUC and calibration, and (5) apply explainability methods (permutation importance and SHAP/LIME) to surface clinically relevant insights. Deliverables will include a reproducible RMarkdown report and a short presentation; all code and data will be published on GitHub for reproducibility.

Motivation and research question

- Motivation: Early detection of diabetes can improve patient outcomes and resource allocation; predicting onset from routine diagnostic measures is clinically useful.
- Primary question: Can routine diagnostic measures reliably predict diabetes onset in this cohort, and which features are most predictive?
- Secondary questions: How do different preprocessing choices (e.g., treating zeros as NA and the imputation method) affect model performance? Do complex models substantially outperform interpretable ones?

Data sources (two different types)

1. Local CSV: "diabetes.csv" — primary dataset (individual-level predictors and Outcome).
2. Web-scraped metadata / API: OpenML dataset page (https://www.openml.org/d/37) or other authoritative documentation for provenance and variable definitions.
    - Optional: public health API (e.g., CDC) for higher-level prevalence context if time permits.

Planned data science workflow (OSEMN)

- Obtain: read local CSV; scrape OpenML for metadata (stored locally for reproducibility).
- Scrub: replace implausible zeros with NA, impute (median/KNN/multivariate), create derived features (age groups, log-transform skewed predictors).
- Explore: summary stats, distributions, correlation matrix, pivoting to long format for small-multipanel plots.
- Model: logistic regression, random forest, XGBoost; hyperparameter tuning with repeated CV.
- Interpret: ROC/AUC, calibration, permutation importance, SHAP or LIME, partial dependence plots.

Planned transformations (examples)

- Replace zeros with NA for Glucose, BloodPressure, SkinThickness, Insulin, BMI.
- Median or KNN imputation; preserve a missingness indicator where informative.
- Age group factor (21–30, 31–40, 41–50, 51+).
- Log(Insulin + 1) to reduce skew.
- Pivot from wide to long for faceted visualizations.

Planned statistical analyses & graphics

- Univariate: histograms / density plots of predictors by Outcome.
- Bivariate: correlation heatmap; scatterplots for key predictor pairs.
- Modeling visuals: ROC curves, calibration plots, variable importance, SHAP summary/force plots, and partial dependence plots.
- Statistical tests: Wald tests for logistic regression coefficients; bootstrap or DeLong tests for comparing AUCs when appropriate.

A novel feature beyond course coverage

- Use SHAP values (via fastshap or DALEX/SHAP integration) to explain model predictions at both global and individual levels.
- Render presentation slides directly from R Markdown (xaringan or ioslides) for reproducible slide generation.

Initial reproducible code (skeleton + light EDA)

```
library(tidyverse)
library(knitr)

# Modeling & evaluation (loaded here, used in project)
library(caret)
library(pROC)
library(randomForest)
library(xgboost)
library(vip)
library(ggplot2)
# library(patchwork)
```

```
# Reads the provided CSV (ensure this file is in the repo/data folder)
diabetes <- readr::read_csv("diabetes.csv", show_col_types = FALSE)

glimpse(diabetes)
```

```
## Rows: 768
## Columns: 9
## $ Pregnancies              <dbl> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, ~
## $ Glucose                  <dbl> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125~
## $ BloodPressure            <dbl> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92, 74~
## $ SkinThickness            <dbl> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0, 0, ~
## $ Insulin                  <dbl> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0, 0, ~
## $ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.~
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.2~
## $ Age                      <dbl> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 3~
## $ Outcome                  <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, ~
```

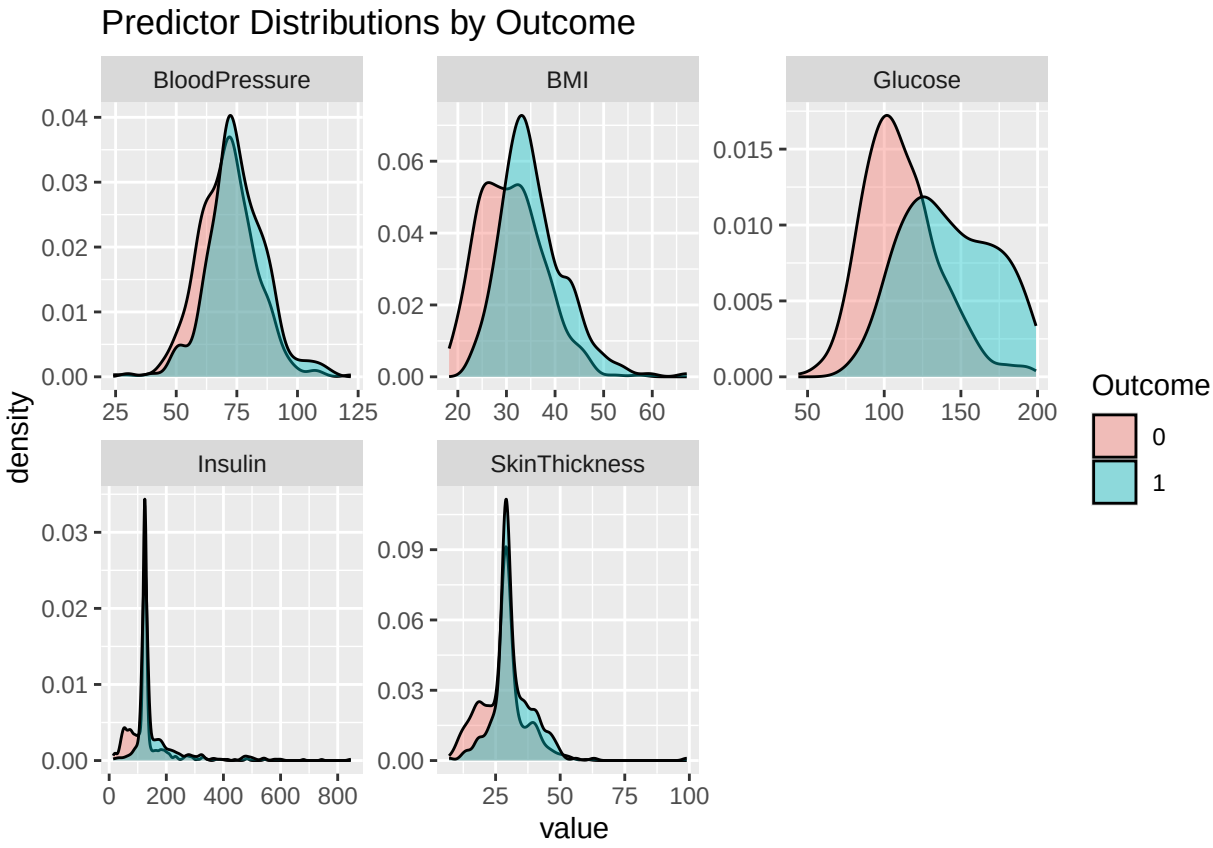
```
table(diabetes$Outcome)
```

```
##
##   0   1
## 500 268
```

```
zero_as_na_cols <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")

diabetes2 <- diabetes %>%
  mutate(across(all_of(zero_as_na_cols), ~ na_if(., 0)))

sapply(diabetes2, function(x) sum(is.na(x)))
```

```
##              Pregnancies                  Glucose            BloodPressure
##                        0                        5                       35
##            SkinThickness                  Insulin                      BMI
##                      227                      374                       11
## DiabetesPedigreeFunction                      Age                  Outcome
##                        0                        0                        0
```

```r
diabetes_imputed <- diabetes2 %>%
  mutate(across(
    all_of(zero_as_na_cols),
    ~ ifelse(is.na(.), median(., na.rm = TRUE), .)
  ))
diabetes_long <- diabetes_imputed %>%
  pivot_longer(
    cols = c(Glucose, BloodPressure, SkinThickness, Insulin, BMI),
    names_to = "measure",
    values_to = "value"
  )

ggplot(diabetes_long, aes(x = value, fill = factor(Outcome))) +
  geom_density(alpha = 0.4) +
  facet_wrap(~ measure, scales = "free") +
  labs(fill = "Outcome", title = "Predictor Distributions by Outcome")
```



Planned model training & evaluation approach

- Use repeated k-fold cross-validation (e.g., 5 repeats x 5 folds) for tuning and evaluation.
- Tune hyperparameters via caret or tidymodels workflows.
- Compare models with AUC as primary metric; also report sensitivity, specificity, PPV, NPV, and calibration.
- Prefer the most interpretable model when performance differences are small.

Project checklist mapping

This proposal addresses the project checklist items: - Motivation and data sources: described above. - Workflow: OSEMN documented. - Two types of data sources: local CSV + web-scraped metadata/API. - Transformations: zeros->NA, imputation, wide->long pivot, derived features. - Statistical analyses & graphics: logistic regression, ROC/AUC, calibration, importance plots. - Novel feature: SHAP/LIME and reproducible slides from R. - Reproducibility: plan to publish Rmd, data, and rendered HTML on GitHub; keep local copies of scraped metadata.

Deliverables & schedule (suggested)

- Proposal submitted: this document.
- Approval meeting: within one week of submission (15 minutes).
- Milestone 1 (data cleaning & EDA): 1 week after approval.
- Milestone 2 (modeling and evaluation): 2–3 weeks after Milestone 1.
- Milestone 3 (interpretation, visuals, slides): 1 week after Milestone 2.
- Final report (RMarkdown HTML) and GitHub repo: per syllabus due date.
- Presentation slides + delivery: per syllabus.

Team / roles

- Solo: malmal6565 — responsible for data cleaning, modeling, visualization, report, and slides.

Reproducibility, code, and repo

Planned repository structure: - data/diabetes.csv - R/ (helper scripts) - proposal.Rmd (this file) - final_report.Rmd - slides/ (Rmd or rendered slides) - README.md with reproduction instructions and package list I will commit the knitted HTML and any saved artifacts required to reproduce results without external web access.

Risks and simplifying assumptions

- Scope limited to 2–3 models rather than exhaustive model exploration.
- Use median or KNN imputation unless multiple imputation is shown to be necessary.
- Keep a local copy of any scraped metadata to avoid reproducibility issues when external pages change.

What I will demonstrate in the final presentation

- Motivation, data sources, and cleaning decisions.
- One or two code snippets showing a real challenge and solution.
- Model comparison (ROC/AUC) and chosen model.
- Model explanation (SHAP/LIME or permutation importance).
- Clear conclusions and clinical/operational recommendations.

Appendix — web scraping / metadata (eval = FALSE)

```r
# Obtain official dataset metadata from OpenML (dataset id = 37)
# Using the OpenML API is more robust and reproducible than HTML scraping

library(OpenML)

openml_ds <- getOMLDataSet(data.id = 37)

# Print the official dataset description / provenance
cat(openml_ds$description)
```

References

- Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus.
- OpenML dataset: Pima Indians Diabetes (dataset id 37) — https://www.openml.org/d/37
- R packages: tidyverse, caret, randomForest, xgboost, pROC, vip, pdp, lime / fastshap