Predicting Chronic Conditions Using Machine Learning (PCC)

CSCI 4146/6409 – Process of Data Science

Winter 2025

Submitted by:

Shubhpreet (B01026120)

Jay Mewada (B00989979)

Prabuddha Deore (B01006790)

Instructor:

Dr. Evangelos Milios

Dalhousie University

Table Of Content

1. Abstract

This project presents Predicting Chronic Conditions Using Machine Learning (PCC)—a screening-oriented data science pipeline designed to predict three critical chronic conditions: high blood pressure, diabetes, and cardiovascular disease. Using the Canadian Community Health Survey (CCHS) 2019–2020 dataset, we apply a modular and interpretable machine learning approach across progressively dependent targets. Preprocessing involved rigorous value cleaning, categorical transformation, and bivariate statistical analysis. Multiple models were evaluated, including logistic regression, decision trees, random forests, XGBoost, and neural networks, using both SMOTE and undersampling to address class imbalance. Logistic regression, optimized for recall using Youden's J statistic, emerged as the most interpretable and reliable model. The final deployment includes a Streamlit-based health screening tool enabling progressive risk prediction. This system supports early intervention and personalized screening in non-clinical settings, demonstrating the potential of interpretable machine learning in public health applications.

2. Introduction

Chronic diseases such as high blood pressure, diabetes, and cardiovascular conditions account for a significant share of healthcare burden globally. Early identification of at-risk individuals can improve outcomes, yet large-scale clinical screenings are often limited by resource constraints, especially in non-clinical or underserved populations.

This project aims to develop a modular, interpretable machine learning (ML) pipeline to screen for these conditions using data from the Canadian Community Health Survey (CCHS) 2019–2020. The CCHS provides comprehensive, population-level survey data on demographics, lifestyle, and health indicators—making it a viable foundation for non-clinical risk prediction.

We designed a sequential prediction framework that reflects clinical progression: high blood pressure → diabetes → cardiovascular disease. By structuring the models this way, the system supports both independent and progressive use, allowing risk estimation even when only partial data is available.

Our goal is to deliver interpretable, recall-optimized models suitable for deployment in public health settings, enabling early intervention through non-invasive, survey-based screening tools.

3. Literature Review

Recent studies have demonstrated the growing role of machine learning (ML) in healthcare, particularly for early detection of chronic conditions. ML models have been used to predict outcomes such as diabetes, cardiovascular disease, and hypertension using both clinical and non-clinical data. For instance, Panahiazar et al. (2015) applied ensemble methods to electronic health records (EHRs) to predict diabetes with high accuracy. However, these approaches often require sensitive clinical data, which limits scalability in population-level screening.

Survey-based datasets like the Canadian Community Health Survey (CCHS) offer an alternative. Although they lack the depth of EHRs, they provide rich, self-reported information on lifestyle, socioeconomic status, and general health—making them valuable for population-level modeling and social determinants of health analysis. Despite this potential, few ML applications using CCHS data have implemented full modeling pipelines optimized for screening purposes.

Interpretability is particularly important in public health applications where end users may not be data scientists. Ribeiro et al. (2016) emphasized the need for explainable models like logistic regression or decision trees to gain trust in non-clinical domains. In alignment with this, our project selects models that offer transparency and actionable insight into predictors of chronic conditions.

To handle class imbalance—a common issue in chronic condition prediction—past studies (Chawla et al., 2002) have proposed techniques like SMOTE (Synthetic Minority Over-sampling Technique) and undersampling. Additionally, Youden's J statistic has been recommended in medical literature (Fluss et al., 2005) to optimize decision thresholds for improving sensitivity (recall), especially in screening contexts.

Our work builds on these foundations by:

- Using survey-based, non-clinical data (CCHS),
- Designing a progressive multi-target pipeline (BP → Diabetes → Cardio),
- Applying interpretable models with recall-focused threshold tuning,

- And enabling real-world usability through a Streamlit screening app.

4. Methodology

This project followed the CRISP-DM methodology to build a screening-oriented machine learning pipeline that predicts three chronic conditions—high blood pressure, diabetes, and cardiovascular disease—using population-level survey data from the Canadian Community Health Survey (CCHS) 2019–2020.

4.1 Data Understanding and Decoding

This project uses the **Canadian Community Health Survey (CCHS) 2019–2020**, a national-level public health dataset released by Statistics Canada. The CCHS includes data on demographic, behavioral, and health-related factors collected from thousands of individuals across Canada.

Each **instance** in the dataset represents one respondent's survey record, encompassing their responses to various health-related questions. The raw dataset contains approximately **690 variables**, of which many are optional modules specific to provinces or age groups.

To ensure **generalizability and consistency**, only **core content variables**—those asked to all respondents—were retained. The feature selection process involved:

- Removing metadata and irrelevant administrative fields (e.g., sequential IDs)
- Excluding short-term behavioral variables (e.g., alcohol use in past 7 days)
- Prioritizing known correlates of chronic disease (e.g., smoking, BMI, perceived health)

After decoding and filtering, we finalized a set of **32 features**, including both behavioral and clinical indicators.

Analytic Base Table (ABT)

No.	Feature	Domain	Description
1	Age Group	Demographics	Age category (e.g., 18–34, 35–49)
….	….	….	….
4	Suicide – Lifetime	Mental Health	Ever considered suicide

….	….	….	….
32	Cardiovascular Condition	Diagnosed Conditions	Target: Heart disease or stroke

**See Appendix A for the full Analytic Base Table (ABT) summarizing the selected features.*

4.2 Data Cleaning and Preprocessing

The data cleaning and preprocessing approach is structured into the following planned steps:

Step 1: Age Group Review

Review the distribution of target variables specifically for the **12–17 age group**, as this group is medically known to have very low prevalence of the chronic conditions targeted. Based on this review, plan whether to include or exclude this age group to ensure accurate and reliable modeling.

Step 2: Iterative Descriptive Statistics and Visualization

Perform iterative rounds of descriptive statistics and visual analyses on all remaining features. Develop detailed transformation plans after each round and implement these transformations. This iterative approach ensures stable and interpretable feature distributions.

Step 3: Bivariate Analysis for Feature Selection

Conduct bivariate analyses, including Chi-square tests of independence and normalized stacked bar plots, to identify features strongly associated with target variables. Use these analyses to finalize the selection of features included in modeling.

4.3 Dataset Preparation

We plan to create three distinct analytic datasets, each aligned specifically with one target condition. To maintain clinical logic and prevent data leakage, each dataset is structured with specific feature exclusions:

- **Dataset for High Blood Pressure:** Exclude diabetes status, cardiovascular condition status, and blood pressure medication use.

- **Dataset for Diabetes:** Exclude cardiovascular condition status; include high blood pressure status.
- **Dataset for Cardiovascular Condition:** Include all available features, reflecting its position as the last stage in clinical progression.

These target-specific datasets provide clear, medically valid inputs for modeling.

4.4 Modeling Strategy and Pipeline Design

Our modeling strategy involves carefully structured pipelines designed to predict each chronic condition accurately and interpretably. Each pipeline is planned as follows:

Step 1: One-Hot Encoding

Categorical features will be encoded into numeric form via **one-hot encoding**, ensuring compatibility with all selected machine learning algorithms.

Step 2: Stratified Train-Test Splitting

Each dataset will be split into training (80%) and testing (20%) subsets, using **stratified sampling** to ensure balanced class distributions across training and testing datasets.

Step 3: Class Imbalance Handling

Given significant class imbalance in the chronic condition data, we will apply the following methods:

- **Undersampling**: Reducing majority-class observations to balance classes.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: Increasing minority-class observations synthetically.

Step 4: Model Pipelines

We will implement three separate modeling pipelines, each applied to all three target conditions:

- **Pipeline 1 (Traditional Models)**: Logistic Regression (for interpretability), Decision Tree (for explicit decision rules) and Random Forest (for ensemble robustness)
- **Pipeline 2 (XGBoost)**: Gradient-boosted decision trees to capture nonlinear patterns.

- **Pipeline 3 (Neural Network)**: Multi-layer Perceptron (MLP) for capturing complex relationships.

4.5 Model Evaluation Approach

Our evaluation approach involves a comprehensive, multi-metric assessment:

- **Cross-validation**: Perform 5-fold cross-validation on training data to reliably estimate model performance.
- **Metrics**: Evaluate each model using: Accuracy, Precision, Recall (primary metric given clinical screening context), F1-score and ROC AUC.

These metrics will guide the assessment of overall predictive quality, model robustness, and recall-oriented performance critical in medical screening applications.

4.6 Threshold Tuning and Final Model Selection

Since default classification thresholds (0.5) are often suboptimal for clinical risk screening, we will tune thresholds specifically to maximize recall using **Youden's J statistic**. We plan to:

- Analyze precision-recall trade-offs across a range of thresholds.
- Select thresholds optimized specifically for recall.

4.7 Deployment Strategy

For practical usability, the finalized model pipeline will be deployed as a user-friendly health screening application using **Streamlit**, with the following features:

- Progressive prediction logic following clinical progression (High BP → Diabetes → Cardiovascular).
- Simple, dropdown-based user interface.
- Predictive flexibility, allowing partial user-provided inputs for each stage.

This deployment approach ensures real-world applicability, interpretability, and ease of use in non-clinical settings.

5. Experiments

The following experiments were conducted to systematically evaluate our modeling approach:

5.1 Data Preparation Experiments

- **Age Filtering**: We analyzed target distributions for the age group 12–17. Due to extremely low prevalence (BP: 40, Diabetes: 12, Cardio: 14 Positive Cases respectively), this age group was removed from the dataset to ensure reliable modeling.
- **Iterative Descriptive Statistics and Transformation**: Two rounds of descriptive statistics were performed, and transformations were implemented accordingly. Ambiguous responses ("Refusal," "Not stated," "Valid skip") were consolidated into "Unknown" or logically replaced ("Valid skip" → "No" for certain features). Sparse categories were merged to ensure feature stability and interpretability.

* *Details of transformation plans are summarized in Appendix B.*

5.2 Bivariate Feature Selection

We conducted **bivariate analysis** using **Chi-square tests** and **visual plots** for all three target conditions—**High Blood Pressure**, **Diabetes**, and **Cardiovascular Condition**. The results showed that **almost all features demonstrated strong statistical associations** with the targets (p-values close to 0). For **High Blood Pressure**, three features (**Sex at Birth**, **Mood disorder**, and **Anxiety disorder**) had relatively weaker p-values (all greater than 0.01), but were retained due to their **clinical relevance** and **interpretability**. As a result, **no features were dropped** at this stage, and **all variables were carried forward for multivariate modeling**.

5.3 Modeling Experiments

To evaluate our prediction framework, we implemented and benchmarked three modeling pipelines across all three target conditions. Each pipeline was designed to test different model families and sampling techniques, and all experiments were executed using the target-specific datasets described in earlier sections.

5.3.1 Pre-processing & Data Splits

- **One-hot encoding** of every categorical field via `OneHotEncoder(handle_unknown="ignore", sparse_output=False)` inside a `ColumnTransformer`
- Each table was split **stratified 80 % / 20 %** (train / test, *random_state = 42*) before any resampling or encoding

5.3.2 Class‑Imbalance Handling

Pipeline	Technique(s)	Library	Note
Traditional (LogReg, DT, RF)	**RandomUnderSampler** & **SMOTE**	`imblearn`	Both samplers evaluated for every model
XGBoost	RandomUnderSampler & SMOTE	`imblearn`	Same grid & CV as traditional models
MLP	**SMOTE** only	`imblearn`	Undersampling dropped to keep training size

5.3.3 Model Configurations & Hyper‑parameter Search

Model	Key Grid (5‑fold CV)	Defaults Used
Logistic Regression	*C ∈ {0.1, 1, 10}*	solver=*lbfgs*, *max_iter=500*
Decision Tree	*max_depth ∈ {3, 5, 10, None}*	criterion=*gini*
Random Forest	*n_estimators ∈ {50, 100}*, *max_depth ∈ {5, 10, None}*	—
XGBoost	*max_depth ∈ {3, 5, 10}*, *learning_rate ∈ {0.01, 0.1, 0.2}*, *n_estimators ∈ {50, 100, 200}*	*eval_metric="logloss"*

MLP (Sequential)	Dense 64 → Drop 0.2 → Dense 32 → Drop 0.2 → Sigmoid,
	*epochs = 30*, *batch_size = 64*,
	EarlyStopping(patience = 5)
		optimiser=*adam*,
		loss=*binary_crossentropy*

5.3.4 Cross‑validation & Metrics

- **5‑fold stratified CV** (`GridSearchCV`) with multi‑metric scoring: *F1‑Yes*, *Recall‑Yes*, *ROC‑AUC*

- MLP used a **manual** `StratifiedKFold` **5× loop** to collect fold-wise F1, Recall and AUC
- We report **mean ± SD** and select hyper-parameters that maximise **F1-Yes**.

6. Results

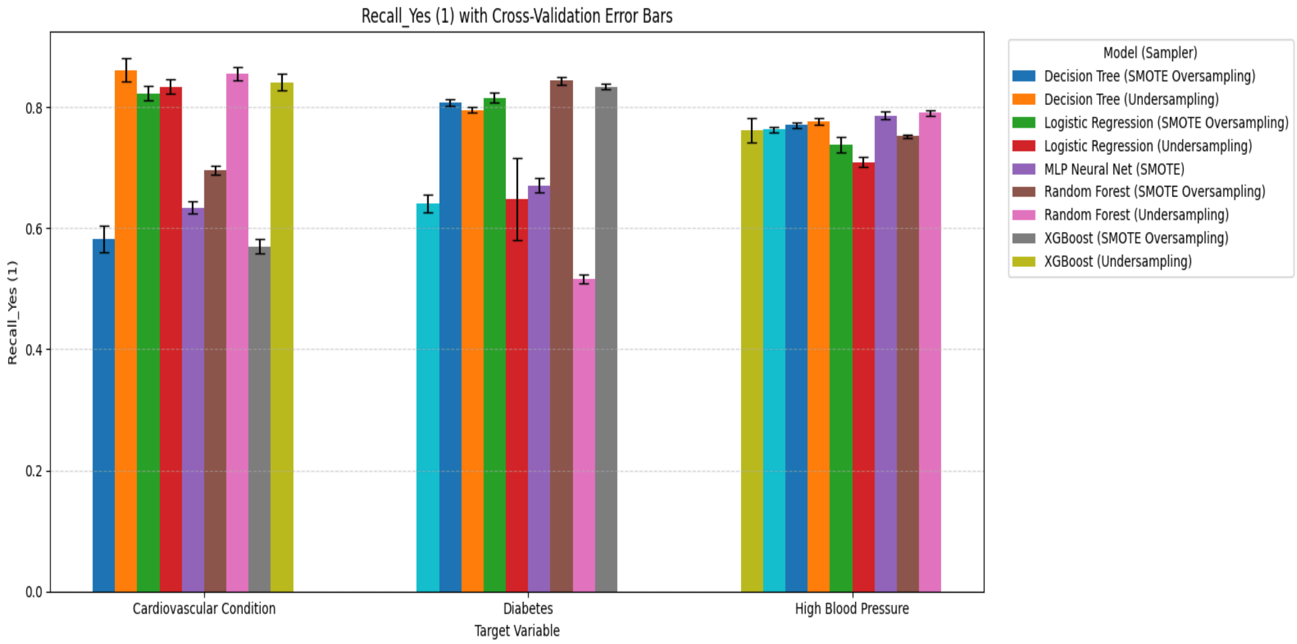
6.1 Evaluation Metrics

Each model was evaluated on both cross-validation (mean ± std) and test sets using the following metrics: Accuracy, Precision, Recall *(primary metric),* F1-score and ROC AUC

Top 5 Models – High Blood Pressure:

Model	AUC ± SD	Recall ± SD	F1 ± SD
XGBoost (Undersampling)	0.810 ± 0.002	0.797 ± 0.004	0.614 ± 0.002
Logistic Regression (Undersampling)	0.810 ± 0.002	0.784 ± 0.005	0.615 ± 0.002
Random Forest (Undersampling)	0.808 ± 0.002	0.787 ± 0.007	0.615 ± 0.001
Logistic Regression (SMOTE Oversampling)	0.807 ± 0.002	0.782 ± 0.004	0.616 ± 0.001
Random Forest (SMOTE Oversampling)	0.807 ± 0.002	0.708 ± 0.008	0.606 ± 0.004

**See Appendix C.1 For Diabetes & Cardiovascular Condition*



**See Appendix C.2 For F1_Yes and ROC AUC Cross-validation error bars*

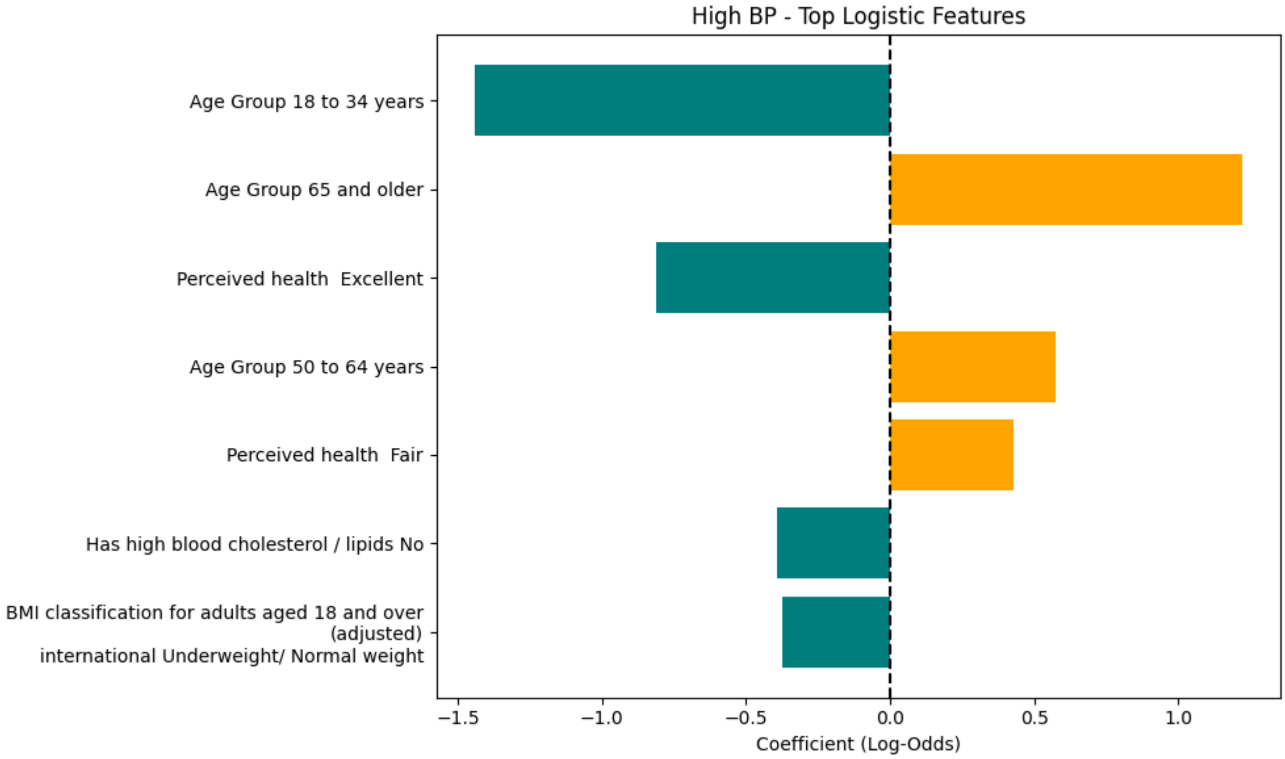
6.2 Model Comparison

- **Logistic Regression** offered the best trade-off between recall and interpretability among baseline models.
- **Decision Tree** and **Random Forest** showed decent performance, but interpretability decreased with complexity.
- **XGBoost** and **MLP** performed slightly better overall but lacked interpretability and required more computational resources.
- Based on balanced metrics, interpretability, recall optimization, and clinical usability, **Logistic Regression** emerged as the **optimal final model across all targets.**

6.3 Interpretation of Final Model

Top predictors identified from Logistic Regression included features such as age group, perceived health, BMI category, and chronic conditions (e.g. cholesterol).

Coefficient importance plots (Top 7) For High BP:

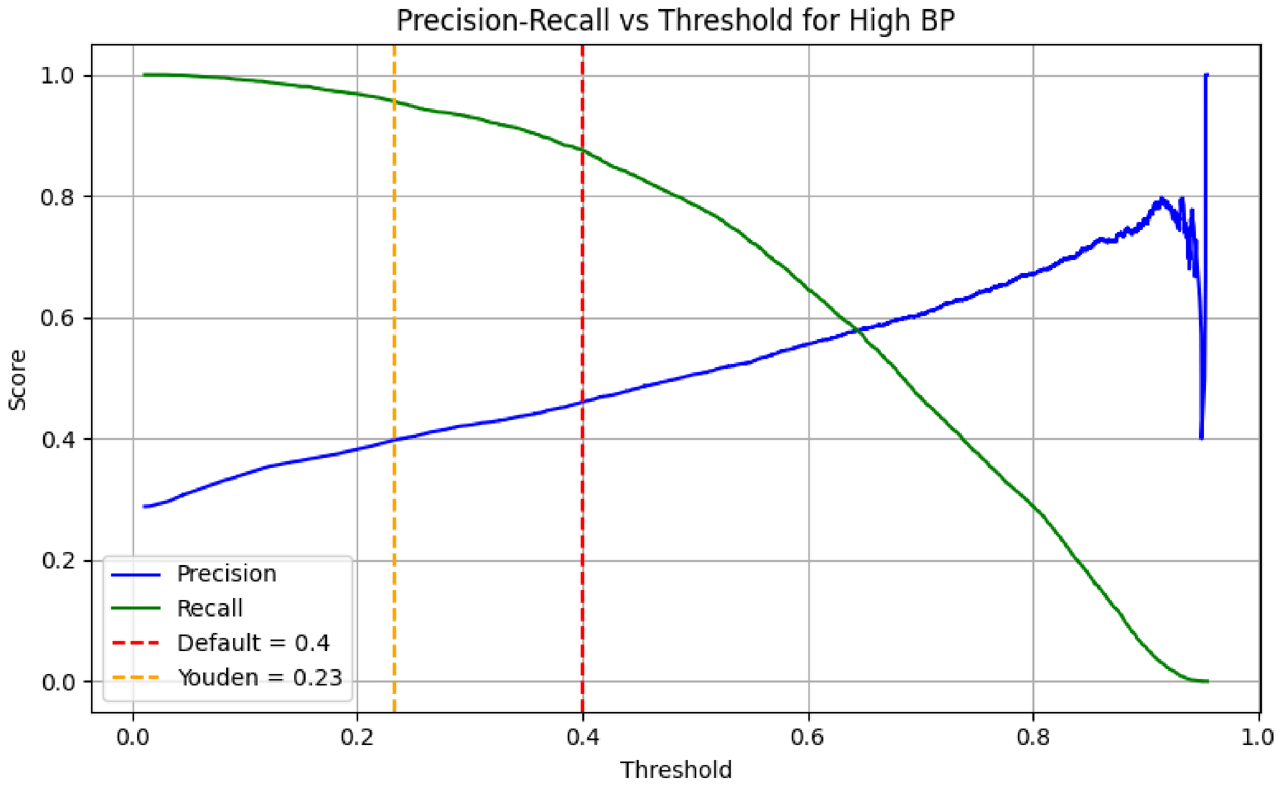


**Coefficient importance plots for Diabetes & Cardiovascular Condition are available in Appendix D.*

6.4 Threshold Optimization of Final Model

Modified Youden's J statistic [`that maximizes (Recall + Precision - 1)`] substantially improved recall of our final model (Logistic Regression). We Prioritize Recall to reduce missed cases and accept minor trade-off in precision.

Precision-Recall vs Threshold for High BP Plot:



* *Precision-Recall vs Threshold for Diabetes and Cardiovascular disease are in Appendix E.*

6.5 Deployment

The final logistic regression model was successfully deployed via Streamlit. The progressive prediction interface accurately reflected medical logic (High BP → Diabetes → Cardiovascular), proving suitable for real-world screening and risk estimation.

7. Conclusion

This project successfully developed and evaluated an interpretable, recall-focused machine learning pipeline using the Canadian Community Health Survey (CCHS) dataset to screen for chronic conditions. Our structured approach addressed critical challenges including data quality, interpretability, class imbalance, and threshold optimization. Logistic Regression, optimized using Youden's J statistic, demonstrated high clinical usability, achieving significantly improved recall across targets.

The Streamlit deployment provided practical value, facilitating personalized and accessible health-risk screening. Future improvements include integrating clinical validation, addressing potential biases inherent in survey data, and extending the pipeline to other public health datasets.

8. References

- Statistics Canada. 2020. *Canadian Community Health Survey (CCHS) 2019–2020 Public Use Microdata File*. Statistics Canada.
- M. Panahiazar, C. Taslimitehrani, M. J. Jadhav, and H. R. Pathak. 2015. Using EHRs and machine learning for diabetes risk prediction. In *AMIA Annual Symposium Proceedings*, 2015: 1899–1908.
- M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), 1135–1144. https://doi.org/10.1145/2939672.2939778
- N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 321–357. https://doi.org/10.1613/jair.953
- R. Fluss, D. Faraggi, and B. Reiser. 2005. Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal* 47, 4, 458–472. https://doi.org/10.1002/bimj.200410135
- Scikit-learn developers. 2023. *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/
- XGBoost contributors. 2023. *XGBoost: Scalable and Flexible Gradient Boosting*. https://xgboost.ai/
- Streamlit Inc. 2023. *Streamlit Documentation*. https://streamlit.io/

Appendix

Appendix A

Complete ABT Table:

No.	Feature	Domain	Description
1	Age Group	Demographics	Age category (e.g., 18–34, 35–49)
2	Sex at Birth	Demographics	Respondent's sex assigned at birth
3	Marital Status	Demographics	Current marital status
4	Suicide – Lifetime	Mental Health	Ever considered suicide
5	Suicide – 12 Months	Mental Health	Considered suicide in the last 12 months
6	Smoking Status	Lifestyle	Smoking behavior (current, former, never)
7	Cannabis Dependence	Substance Use	Severity of cannabis dependence
8	Cannabis Use – 12 Months	Substance Use	Cannabis use in the past 12 months
9	Usual Care Place	Access to Care	Has a regular place for minor health issues
10	Household Income	Socioeconomic	Total household income
11	BMI (Age 12–17)	Health Metrics	WHO BMI classification (only used for validation)
12	BMI (Adults)	Health Metrics	BMI class based on adjusted international standard
13	Pain Status	Health Status	Reports usual pain or discomfort
14	Perceived Health	Health Perception	Self-rated physical health
15	Mental Health Rating	Mental Health	Self-rated mental health
16	Life Satisfaction	Well-being	General life satisfaction
17	Flu Shot – Ever	Immunization History	Ever had a seasonal flu shot
18	Flu Shot – Last Time	Immunization History	Time since last flu shot
19	Drinker Type	Lifestyle	Alcohol consumption behavior

20	Binge Drinking Frequency	Lifestyle	Frequency of 5+/4+ drinks on one occasion
21	Sleep Apnea	Diagnosed Conditions	Diagnosed with sleep apnea
22	High Cholesterol	Diagnosed Conditions	Diagnosed with high blood cholesterol
23	Cholesterol Medication	Medication Use	Took cholesterol meds in past month
24	Chronic Fatigue	Diagnosed Conditions	Diagnosed with chronic fatigue syndrome
25	Mood Disorder	Mental Health	Diagnosed mood disorder (e.g., depression)
26	Anxiety Disorder	Mental Health	Diagnosed anxiety-related disorder
27	Respiratory Condition	Diagnosed Conditions	Asthma or COPD
28	Musculoskeletal Issue	Diagnosed Conditions	Arthritis, fibromyalgia, or osteoporosis
29	BP Medication	Medication Use	Took BP medication in the past month
30	High Blood Pressure	Diagnosed Conditions	Target: Has high blood pressure
31	Diabetes	Diagnosed Conditions	Target: Has diabetes
32	Cardiovascular Condition	Diagnosed Conditions	Target: Heart disease or stroke

Appendix B

Transformation Plan after 1st Round:

S. No.	Feature	Categories to Merge/Replace	New Value	Action	Why
a	Marital Status	Not stated	Married/Common-law	Merge	Avoid noise; only 375 cases.
b	Considered suicide - lifetime	Not stated, Refusal, Don’t know	Unknown	Merge	Unified non-responses.
c1	Considered suicide - last 12 months	Valid skip	No	Replace	Logically skipped → "No".

c2	Considered suicide - last 12 months	Not stated, Refusal, Don’t know	Unknown	Merge	Clarified non-responses.
d1	Smoking status	Abstainer, Experimental	Non-smoker (abstainer or experimental)	Merge	Reduce category noise.
d2	Smoking status	Not stated	Unknown	Rename	For consistency.
e1	Cannabis Dependence	Valid skip	No cannabis use	Recode	No cannabis use.
e2	Cannabis Dependence	Not stated	Unknown	Recode	Missing info grouped.
e3	Cannabis Dependence = 0	—	No dependence	Recode	Symptom-free group.
e4	1–4	—	Mild dependence	Recode	Grouped low-dep.
e5	5–10	—	Moderate dependence	Recode	Clinically meaningful.
e6	11–15	—	Severe dependence	Recode	Grouped for analysis.
f1	Used cannabis - 12 months	Not stated, Don’t know, Refusal	Unknown	Merge	Unified missing values.
g1	Usual care place	Don’t know, Refusal	Unknown	Merge	Cleaner category.
h1	Income	Not stated	Unknown	Recode	Preserve value.
i1	BMI 12–17	All rows = skip	Drop	Drop	No variance.
j1	BMI adult	Not stated	Unknown	Recode	Keep completeness.
k2	Pain health status	Not stated	Unknown	Recode	Clean missing.
l1	Perceived health	Not stated	Unknown	Recode	Same as above.
m1	Mental health	Not stated	Unknown	Recode	Same logic.
n1	Life satisfaction	Not stated	Unknown	Recode	Uniform missing tag.

o1	Flu shot - lifetime	Not stated, Don’t know, Refusal	Unknown	Merge	Consistent category.
p1	Flu shot - last time	Not stated, Don’t know, Refusal	Unknown	Merge	Same as above.
q1	Drinker type	Not stated	Unknown	Recode	Normalize category.
r1	5+/4+ drinks freq	Not stated, Don’t know, Refusal	Unknown	Merge	Group unclear freq.
s1	Sleep apnea	Don’t know, Refusal	Unknown	Merge	Simplified unclear tags.
t1	High cholesterol	Don’t know, Refusal	Unknown	Merge	Unified missing.
u1	Cholesterol med use	Don’t know, Refusal	Unknown	Merge	Keep clarity.
v1	Chronic fatigue	Don’t know, Refusal	Unknown	Merge	Ensure interpretability.
w1	Mood disorder	Don’t know, Refusal	Unknown	Merge	Merge mental flags.
x1	Anxiety disorder	Don’t know, Refusal	Unknown	Merge	Same reason.
y1	Respiratory condition	Not stated	Unknown	Recode	Unified missing.
z1	Musculoskeletal	Not stated	Unknown	Recode	Clean handling.
aa1	BP medication	Don’t know, Refusal	Unknown	Merge	Unified flag.
ab1	Has high BP	Don’t know, Refusal	Unknown	Merge	For modeling clarity.
ac1	Diabetes	Not stated, Don’t know, Refusal	Unknown	Merge	For consistency.
ad1	Cardiovascular condition	Not stated	Unknown	Recode	Clean handling.

Transformation Plan after 2nd Round:

S. No.	Feature	Categories to Merge/Replace	New Value	Action	Why
a1	Severity of Cannabis Dependence	Mild, Moderate, Severe dependence	Takes cannabis & dependent on it	Merge	Moderate (0.44%) and Severe (0.03%) too sparse; merging avoids sparsity while preserving dependence signal.
a2	Severity of Cannabis Dependence	No dependence	Takes cannabis but no dependence	Rename	Clarifies user takes cannabis but shows no dependence.
b1	Usual place for immediate care for minor problem	Unknown	Mode value	Replace	Only 0.31%; assumed missing-at-random; using mode maintains distribution integrity.
c1	Pain health status	Unknown	Mode value	Replace	Rare (0.32%); likely missing-at-random; replaced to avoid sparsity.
d1	Perceived health	Unknown	Mode value	Replace	Proportion too small (0.16%) to justify separate category; mode imputation is safe.
e1	Satisfaction with life in general	Very Dissatisfied, Dissatisfied	Dissatisfied	Merge	Very Dissatisfied (0.67%) is rare; merging improves category size and interpretability.
f1	Type of drinker	Unknown	Mode value	Replace	0.4% missing likely random; using mode avoids noise.
g1	Drank 5+/4+ drinks freq (12 months)	Unknown	Mode value	Replace	0.58% missing; replacing with mode avoids category fragmentation.
h1	Has sleep apnea	Unknown	Mode value	Replace	0.21% proportion is too low; assumed missing-at-random.

i1	High cholesterol - med use	Unknown	Mode value	Replace	0.41% rare; likely data collection gaps, safe to impute.
j1	Chronic fatigue syndrome	Unknown	Mode value	Replace	0.23% is small enough for mode replacement.
k1	Mood disorder	Unknown	Mode value	Replace	0.18%; likely non-response; handled by mode.
l1	Anxiety disorder	Unknown	Mode value	Replace	Very low (0.19%); replaced assuming random missingness.
m1	High BP - med use	Unknown	Mode value	Replace	0.23%; imputed to maintain modeling quality.
n1	Has high blood pressure	Unknown	Mode value	Replace	0.34% is sparse; replaced for consistency.

Appendix C

Appendix C.1

Top 5 Models – Diabetes:

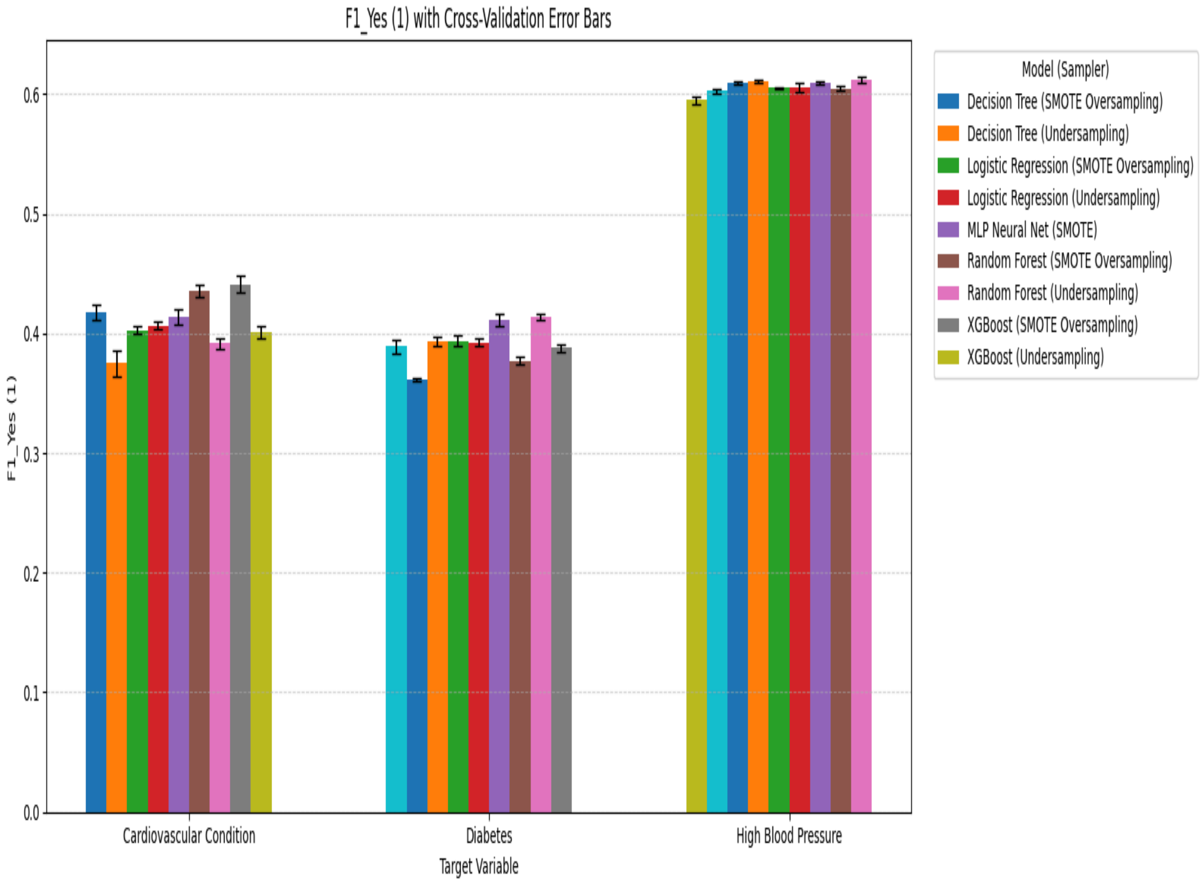
Model	AUC ± SD	Recall ± SD	F1 ± SD
XGBoost (Undersampling)	0.853 ± 0.003	0.822 ± 0.005	0.385 ± 0.003
Logistic Regression (Undersampling)	0.853 ± 0.002	0.799 ± 0.008	0.390 ± 0.005
XGBoost (SMOTE Oversampling)	0.848 ± 0.001	0.525 ± 0.007	0.418 ± 0.003
Logistic Regression (SMOTE Oversampling)	0.848 ± 0.002	0.785 ± 0.004	0.390 ± 0.004
Random Forest (Undersampling)	0.848 ± 0.002	0.826 ± 0.007	0.377 ± 0.003

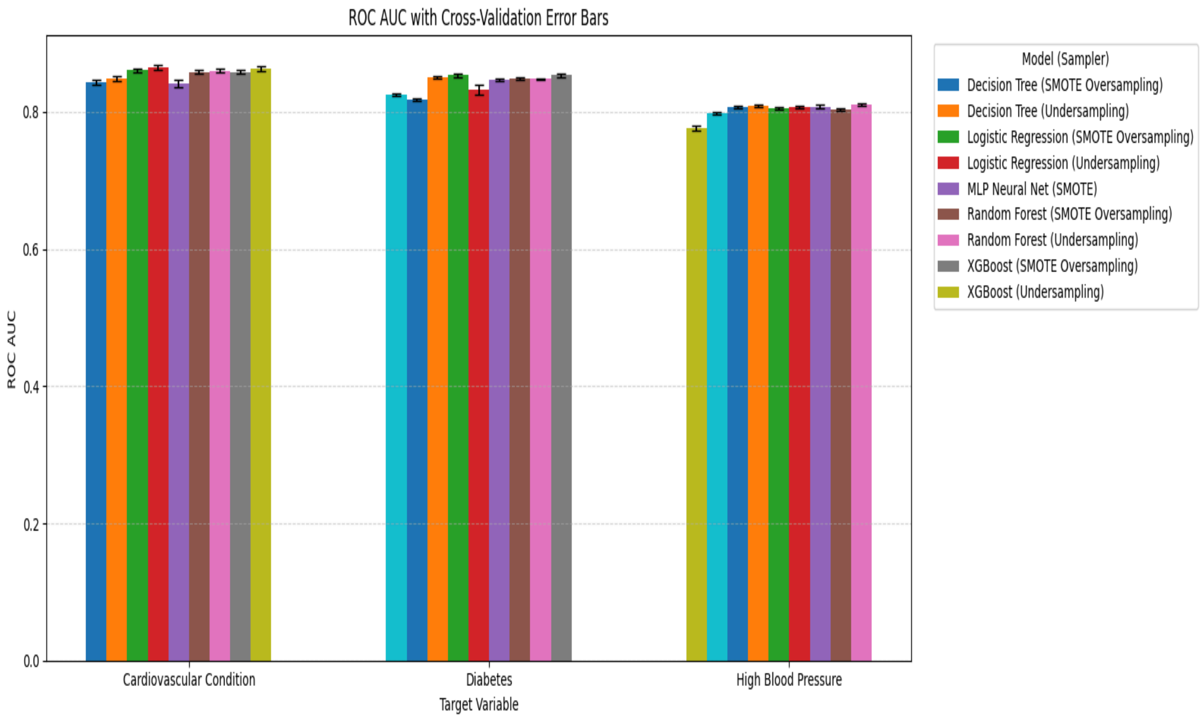
Top 5 Models – Cardiovascular Condition:

Model	AUC ± SD	Recall ± SD	F1 ± SD
Logistic Regression (Undersampling)	0.856 ± 0.004	0.814 ± 0.011	0.396 ± 0.003
XGBoost (Undersampling)	0.856 ± 0.004	0.835 ± 0.014	0.393 ± 0.005
Logistic Regression (SMOTE Oversampling)	0.853 ± 0.003	0.805 ± 0.012	0.395 ± 0.003
Random Forest (Undersampling)	0.853 ± 0.003	0.848 ± 0.011	0.388 ± 0.004
Random Forest (SMOTE Oversampling)	0.852 ± 0.003	0.694 ± 0.007	0.430 ± 0.005

Appendix C.2

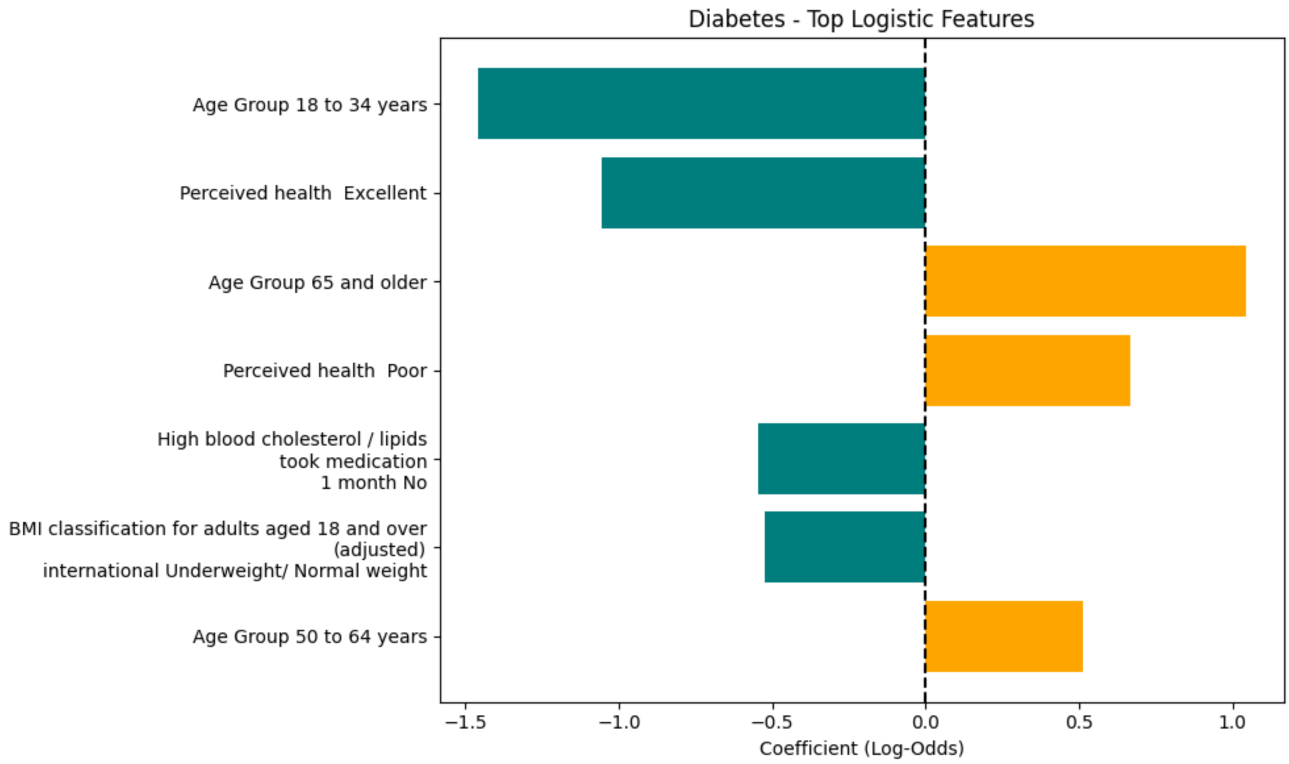
F1_Yes and ROC AUC Cross-validation error bars

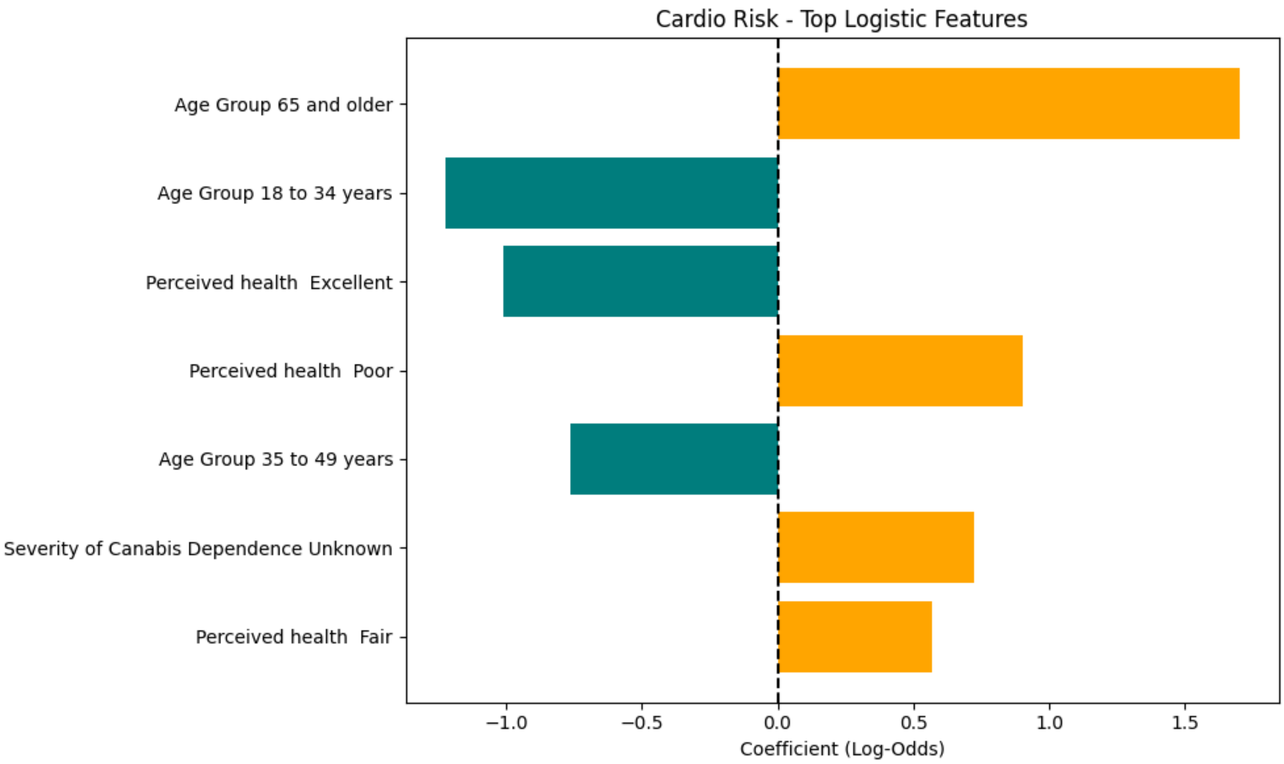




Appendix D

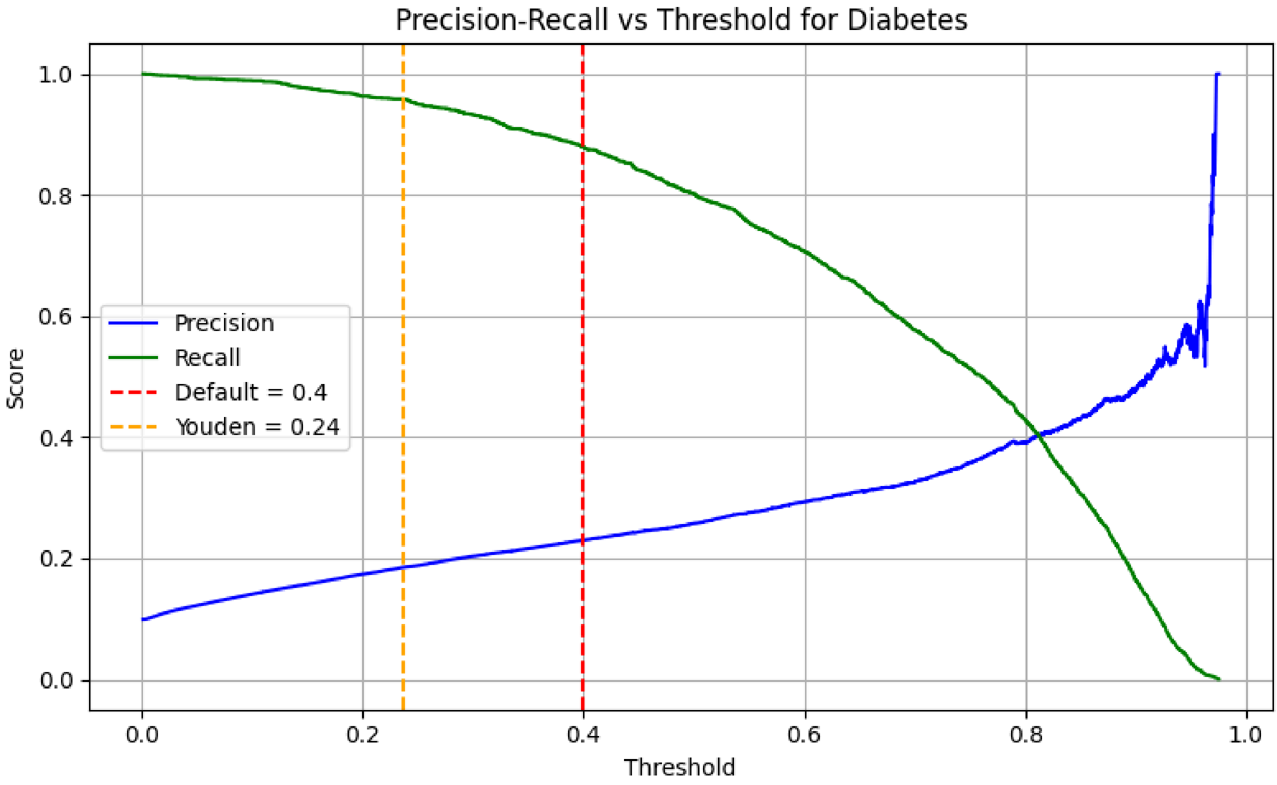
Coefficient importance plots for Diabetes & Cardiovascular Condition

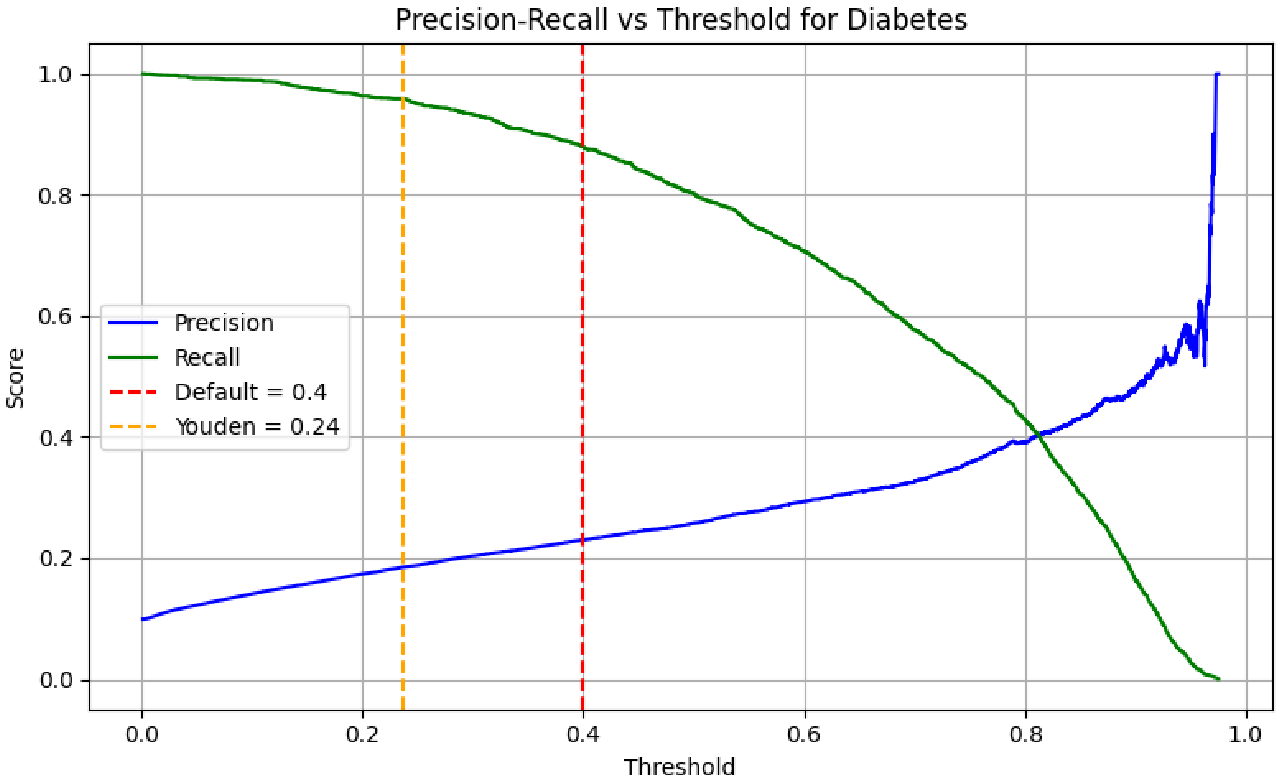




Appendix E

Precision-Recall vs Threshold for Diabetes and Cardiovascular disease




Precision-Recall vs Threshold for Diabetes
Score
Threshold
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision
Recall
Default = 0.4
Youden = 0.24