Interpretable and Fair Machine Learning for Diabetes Risk Prediction Using Clinical Data

(Data Science project)

Prepared by:

(CS4-G1)

Murad Bishr	23160018
Abdullah Al-Kbous	23160045
Riyadh Marish	23164062
Abdulsalam Muzafar	23164099

Date:

January 15th, 2026

1. Abstract

Diabetes is a growing worldwide health challenge where early risk detection enables timely intervention. However, many ML approaches, while producing accurate predictions, remain opaque to the clinician, limiting their adoption in the real world. In this paper, we present an easy-to-reproduce ML pipeline for diabetes risk prediction using the publicly available NHANES 2017–2018 dataset. We combine interpretable and conventional models -calibrated Logistic Regression and XGBoost- and apply homogeneous preprocessing, imputation, and scaling. Making the predictions actionable, we integrated explainable AI, SHAP, for global and per-patient interpretation, evaluated model calibration to obtain reliable risk scores, and assessed model fairness across demographic subgroups: age, gender, and race. Our results show that a calibrated, interpretability-focused pipeline achieves competitive predictive performance while offering clear explanations and clinically useful subgroup differences that will facilitate equitable deployment. We make the code and artifacts publicly available to help ensure reproducibility and foster further validation and clinical translation.

Key words: Diabetes Risk Prediction, Machine Learning, Explainable AI(SHAP), Fairness/Bais Analysis, Healthcare Decision Support

2. Introduction

Diabetes mellitus is one of the most important current public health problems in the world [2]. It inflicts considerable suffering and very high costs on healthcare systems worldwide. Early detection of persons at high risk enables measures that can either prevent or delay the development of the disease [2]. ML has been found to be quite important in diabetes risk prediction due to the fact that it brings to the surface intricate patterns from clinical data of enormous size. In real clinical applications, however, ML models should be interpretable, well-calibrated, and fair among different demographic groups, besides being powerful in predictive ability [5], [18].

Previous ML studies have often used small or convenience datasets, such as the Pima Indians dataset, and most have employed large, complex models yielding high discrimination scores. However, most of these studies offer limited interpretability, without calibration analysis, and rarely consider fairness in predictions across age, gender, or race. Small differences in pre-processing steps and labeling further limit reproducibility by making the comparison of results, or their translation into clinical practice, non-trivial [20]. Clinicians therefore view black-box models with skepticism due to their uncertain reliability in a diverse range of populations [5].

These gaps denote persistent challenges: inadequate model transparency, missing calibration assessments, insufficient fairness evaluation, and a lack of reproducible pipelines using large public datasets. Taken together, these issues limit the clinical readiness of ML models for diabetes screening and prevention. We present here a simple, reproducible ML pipeline using the public NHANES 2017–2018 dataset [1]to address these limitations. We apply calibrated Logistic Regression and XGBoost with standardized preprocessing [14], evaluate discrimination, calibration [9], [12], and subgroup performance, and use SHAP to provide both global and per-patient explanations [4]. We release all code and artifacts to ensure reproducibility [20].

Our contributions are as follows: a complete, reproducible pipeline for NHANES data; a comparison of calibrated Logistic Regression and XGBoost; SHAP-based explainability; extensive calibration and equity analysis; and practical artefacts and advice towards clinical decision support applications.

3. Related Work

Machine learning has rapidly advanced the field of diabetes risk prediction: early work relied heavily on small, convenience datasets and singular models while more recent studies have emphasized multi-model benchmarking and external validation on population samples. A useful illustration of this evolution is provided by Amirian et al., who benchmark six supervised classifiers, three anomaly detectors and a stacking ensemble against the established FINDRISC score using a large prospective cohort with external validation on NHANES and the PIMA dataset [3]. Their study is methodologically rigorous—harmonizing preprocessing, applying stratified cross-validation, and reporting bootstrapped confidence intervals to provide an honest assessment of transportability—but its main focus is discrimination (AUC/PR) and external generalizability rather than the full set of concerns that determine clinical readiness. In particular, while Amirian et al. use SHAP to report global feature importance and provide sensible sensitivity analyses, they do not elevate probability calibration, per-patient explanation, or systematic subgroup equity to primary evaluation objectives [3], [4].

Parallel work has shown how explainability and usability can be married to prediction in practical systems. Maimaitijiang et al. design an explainable AI framework that couples a CatBoost classifier with SHAP explanations and an interactive chatbot interface to deliver personalized, user-facing explanations and recommendations [6]. This line of work demonstrates a plausible route to clinician- or patient-facing deployment by making model outputs actionable, but it relies on a relatively small clinical sample and emphasizes discrimination and usability over rigorous calibration and comprehensive fairness evaluation. Similarly, a range of applied NHANES-based and high-risk cohort studies compare classical classifiers (logistic regression, random forest, SVM, and gradient boosting) and show that tree-based ensembles often achieve the strongest discrimination in clinically targeted samples [7]. These applied analyses are valuable for feature discovery and model selection but commonly report only discrimination and feature rankings; they typically do not provide calibrated risk estimates, do not produce patient-level, clinician-ready explanations linked to each prediction, and rarely report systematic performance differences across demographic groups [9], [12], [4], [18].

Taken together, the recent literature has moved in positive directions—larger cohorts, honest external validation, and adoption of explainability tools such as SHAP—but three consistent gaps remain. First, discrimination is not a substitute for clinically useful risk scores: many works omit rigorous calibration analysis—what is expected in terms of calibration curves, Brier score, and post-hoc recalibration—when the predicted probabilities will be used to triage patients or to inform shared decision-making [9], [12]. Second, while global feature importance is increasingly reported, per-patient interpretability—the explanations that justify an individual's predicted risk and can be communicated to clinicians and patients—is seldom integrated with evaluation metrics or presented as an explicit deployment output [4]. Third, systematic fairness evaluation—quantifying how both discrimination and calibration vary across age, gender, and race/ethnicity—is still uncommon, despite being a centerpiece for equitable clinical translation [18].

The cited study directly addresses these gaps. Building on the methodological rigor that benchmarking literature and explainability-driven deployments have come to represent, we present a reproducible NHANES-based pipeline that treats calibration, per-patient explainability, and subgroup fairness as first-class evaluation objectives. Specifically, we: 1) report discrimination alongside calibration metrics (Brier score and calibration curves) and apply calibration correction where appropriate [9], [12]; 2) use SHAP for global insight and for concise, clinician-oriented patient-level explanations tied to each prediction [4]; and 3) quantify model performance and calibration across age, gender, and race subgroups to surface and quantify disparities [18]. Finally, we release preprocessing scripts, modeling notebooks, and model artifacts to enable replication and independent validation, as transparent end-to-end reproducibility is key for responsible deployment [20].

Study (representative)	Dataset (size / type)	Models evaluated	Explainability used	Calibration evaluated	Fairness / subgroup analysis	Reproducibility (code / pipeline)	Key findings	How our work extends/addresses
Amirian et al. — Internal & External Validation.	Large prospective cohort (n≈9,171) + NHANES & PIMA external	LR, RF, GBM, SVM, NN, stacking, anomaly detectors	SHAP (global importance)	Not emphasized as primary evaluation	Not systematically quantified	Methodology described; code/artifacts not released in paper	ML outperforms FINDRISC on discrimination when lab data available; external AUCs decline	Adds per-patient SHAP, explicit calibration, subgroup fairness, public NHANES pipeline
Maimaitijiang et al. — Explainable AI + Chatbot.	Small clinical dataset (≈520 patients)	CatBoost	SHAP (global & per-user) + chatbot interface	Not reported / not primary	Limited subgroup insight	Deployment prototype reported; full pipeline/code not emphasized	High discrimination; user-facing explainability demonstrated	Scales explainability to nationally representative data, adds calibration and fairness reporting
Yang et al. — ML in high-risk NHANES cohorts.	NHANES cycles (thousands; high-risk subset)	LR, RF, XGBoost, SVM, etc.	Feature importance or coefficient reports	Typically not evaluated	Not systematically evaluated	Often no complete public pipeline	Tree ensembles show strong discrimination; key predictors: waist, BMI, age, BP	Adds patient-level explainability (SHAP), calibration, and subgroup equity analyses

Table 1 – related work comprasion

4.Methodology and Model Design

Data source and provenance. We used the publicly available National Health and Nutrition Examination Survey (NHANES) 2017–2018 public-use files [1]. Data were merged across the Demographics (DEMO), Examination (EXAM), and Questionnaire (QUEST) modules using the participant identifier (SEQN). Key variables included RIDAGEYR (age), RIAGENDR (gender), RIDRETH1 (race/ethnicity), BMXBMI (BMI), BPXSY1–3/BPXDI1–3 (blood pressure), and SMQ020 (smoking). Full variable mappings and the variable dictionary are provided in the repository and Appendix.

Cohort selection. We retained adults aged ≥18 years to focus on the adult population, since type 2 diabetes (T2DM) is rare and pathophysiologically different in children and adolescents. Participants missing label-defining information (no DIQ010 and no laboratory measurements usable to form the label) were excluded. After merging modules and applying inclusion/exclusion criteria, N = 5856 participants remained for analysis.

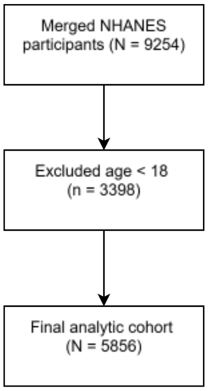

Merged NHANES participants (N = 9254)
Excluded age < 18 (n = 3398)
Final analytic cohort (N = 5856)

Figure 1 – Cohort Flowchart

Label definition and leakage prevention. The binary diabetes label (T2DM) was defined as DIQ010 = "Yes" (self-reported physician diagnosis) OR HbA1c ≥ 6.5% OR fasting plasma glucose ≥ 126 mg/dL, consistent with the American Diabetes Association (ADA) criteria for diagnosing type 2 diabetes [2]. This combination ensures that both self-reported diagnoses and biochemically confirmed cases are captured, increasing sensitivity while maintaining clinical validity.

Laboratory variables used to define the label (HbA1c, fasting glucose) were excluded from the primary model input to prevent target leakage — that is, to avoid allowing the model to "cheat" by seeing the outcome-defining features. Excluding these variables also allows the model to simulate a mostly questionnaire-based screening scenario, which is more practical in resource-limited or population-level screening contexts.

By defining the label this way, the study ensures that the outcome is clinically meaningful, robust, and reproducible, while maintaining rigorous separation between predictors and outcome.

Feature sets. The primary feature set comprised demographic, anthropometric, vital sign, and questionnaire variables routinely available without many blood tests. The main model_features used were: ['age','gender','race','bmi','chol_total','chol_hdl','smoke_now','family_diabetes','physically_active','bp_sys_mean','bp_dia_mean']. Laboratory-augmented experiments including HbA1c and fasting glucose were conducted as secondary analyses to quantify lift from lab data.

Exploratory data analysis. We inspected distributions, missingness, and bivariate associations for the EDA column set (['age','gender','race','bmi','hba1c','glu','chol_total','chol_hdl','smoke_now','family_diabetes','physically_active','bp_sys_mean','bp_dia_mean']). Visualizations included histograms, density plots, correlation heatmaps, and stratified boxplots. EDA informed preprocessing choices and label plausibility checks.

Feature (EDA)	*NHANES Module*	*NHANES Variable*	*Description*	*Type*	*Units / Values*	*Used in Model?*	*Notes*
age	DEMO	RIDAGEYR	Age of participant	Numeric	Years	Yes	Non-lab feature
gender	DEMO	RIAGENDR	Sex of participant	Categorical	Male/Female	Yes	Non-lab feature
race	DEMO	RIDRETH1	Race / ethnicity (NHANES coding)	Categorical	1–5	Yes	Non-lab feature
bmi	EXAM	BMXBMI	Body Mass Index	Numeric	kg/m²	Yes	Non-lab feature
diag_self	QUEST	DIQ010	Self-reported physician diagnosis of diabetes	Categorical	Yes/No	No	Label-defining only
hbA1c	LAB	LBXGH	Hemoglobin A1c	Numeric	%	No	Lab feature, excluded to prevent leakage
glu	LAB	LBXGLU	Fasting plasma glucose	Numeric	mg/dL	No	Lab feature, excluded to prevent leakage
chol_total	LAB	LBXTC	Total cholesterol	Numeric	mg/dL	Yes	lab feature
chol_hdl	LAB	LBDHDD	HDL cholesterol	Numeric	mg/dL	Yes	lab feature
smoke_now	QUEST	SMQ020	Current smoking status	Categorical	Yes/No	Yes	Non-lab feature
family_diabetes	QUEST	DIQ170	Family history of diabetes (parent/sibling)	Categorical	Yes/No	Yes	Non-lab feature
physically_active	QUEST	PAQ650	Regular physical activity	Categorical	Yes/No	Yes	Non-lab feature
bp_sys_mean	EXAM	BPXSY1–3	Mean systolic blood pressure	Numeric	mmHg	Yes	Non-lab feature
bp_dia_mean	EXAM	BPXDI1–3	Mean diastolic blood pressure	Numeric	mmHg	Yes	Non-lab feature

Table 2 -features mapping

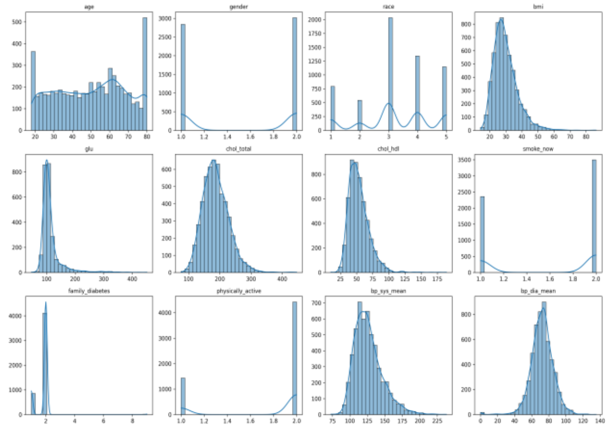


Figure 2 – features description

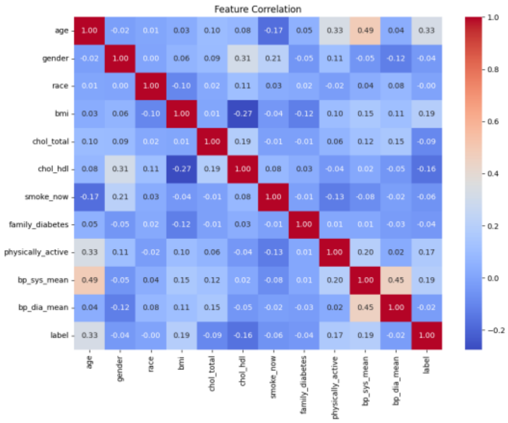


Figure 3 – features correlation

Preprocessing pipeline. Preprocessing was implemented as a scikit-learn–compatible pipeline fitted on training data only. Steps: (1) type coercion with pd.to_numeric(..., errors='coerce'); (2) numerical imputation with SimpleImputer(strategy='median'); (3) categorical imputation with the most-frequent value; (4) categorical encoding via one-hot encoding (ColumnTransformer) unless integer encoding was explicitly preferred for a model; (5) scaling via StandardScaler() [15] fitted to the training set; (6) persistence of feature_names and preprocessing artifacts with joblib.dump() for reproducible inference. All preprocessing parameters were saved with the model artifacts.

Train/test split and resampling. The dataset was split into training and test sets using an 80/20 stratified split (preserving label prevalence) with random_state=42. Model selection and hyperparameter tuning used 5-fold stratified cross-validation on the training set. To address class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied only to training folds during cross-validation; no oversampling was applied to validation or test sets [13].

Models and hyperparameter tuning. We compared two primary model families: (1) logistic regression (LR) as an interpretable baseline, and (2) XGBoost (XGB) [14] as a high-performance comparator. For LR, hyperparameters (regularization strength $C$and solver) were optimized using 5-fold stratified cross-validation with GridSearchCV, selecting the configuration with the highest cross-validated ROC-AUC. The selected LR model's predicted probabilities were then calibrated using isotonic regression (CalibratedClassifierCV(method='isotonic')) on a held-out calibration split from the training data, and final performance was evaluated on the held-out test set.

XGBoost models were trained with xgboost.XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42) and hyperparameters were tuned using RandomizedSearchCV over 50 randomized draws (typical ranges: learning rate 0.01–0.2; n_estimators 100–500; max_depth 3–8; subsample 0.6–1.0; colsample_bytree 0.5–1.0) optimizing ROC-AUC. A stacking ensemble combining base learners with a logistic meta-learner was evaluated as a complementary approach.

Anomaly detection methods (Local Outlier Factor, One-Class SVM, Isolation Forest) were included in sensitivity analyses for low-incidence scenarios. Final model performance was assessed on the held-out test set, reporting ROC-AUC, PR-AUC, Brier score, and classification metrics at clinically relevant thresholds.

Evaluation metrics and uncertainty. Models were evaluated on the held-out test set. Primary evaluation metrics included ROC-AUC and PR-AUC, while secondary metrics included Brier score (calibration) [12], accuracy, precision, recall (sensitivity), specificity, F1-score, and confusion matrices at selected thresholds. Uncertainty for each metric was estimated using nonparametric bootstrap confidence intervals (500 resamples for individual metrics; 2,000 resamples for paired ΔAUC comparisons). Paired bootstrap tests were applied where appropriate to assess statistical differences in correlated ROC-AUCs [16]. PR-AUC was reported to account for the imbalanced prevalence of diabetes in the cohort.

Calibration and decision analysis. Calibration was assessed with calibration curves (sklearn.calibration.calibration_curve, n_bins=10) and the Brier score [12]. Because screening often prioritizes sensitivity, we evaluated multiple thresholds (0.05, 0.8, n=76 evenly spaced) and reported resulting recall/precision tradeoffs. Decision Curve Analysis (DCA) was used to estimate net benefit across thresholds relative to "treat-all" and "treat-none" strategies [17], enabling assessment of clinical utility over a range of operating points.

Interpretability (SHAP). To provide global and per-patient explanations we computed SHAP values: shap.TreeExplainer for XGBoost and shap. KernelExplainer for the calibrated logistic model (or LinearExplainer on an uncalibrated LR copy for faster offline analysis) [4]. Outputs included mean-|SHAP| global importance (top 10 features), beeswarm plots, and per-patient breakdowns. For interactive contexts, a modest background sample (n≈100) was used to balance fidelity and compute time. SHAP analyses informed clinical interpretability and feature prioritization.

Fairness and subgroup analysis. We evaluated model performance and calibration across demographic subgroups: age bins (18–39, 40–59, 60+), gender, and race/ethnicity (NHANES coding). For each subgroup we report ROC-AUC with bootstrap 95% CIs, recall, FPR, FNR, and sample size. Differences in subgroup performance were assessed via bootstrap differencing and flagged when clinically meaningful or statistically significant [18].

In addition to reporting metrics at a fixed global threshold, we evaluated recall as a function of decision threshold within each subgroup to reflect screening operating-point choices. We also computed subgroup-specific thresholds achieving a target sensitivity (e.g., recall ≥ 0.80) to illustrate how threshold selection can equalize sensitivity across groups while changing false-positive burden.

Implementation and reproducibility. All analyses were implemented in Python [15]. Key packages and versions are documented in requirements.txt. Preprocessing artifacts and trained models were saved via joblib.dump(); a reproducible dictionary (imputer, scaler, model, feature_names) is included in the repository [20]. The code, notebooks, and trained artifacts are available at here .We used random_state=42 for all randomized procedures.

Ethics and limitations. NHANES is de-identified public data and does not require IRB approval for secondary analysis [1]. Limitations include the cross-sectional nature of NHANES, potential label noise from mixing self-reported and laboratory definitions, and limited generalizability beyond the US population. Excluding laboratory variables improves deployability in low-resource settings but reduces discrimination relative to lab-augmented models. Residual confounding by socioeconomic and access-to-care variables may affect subgroup assessments.

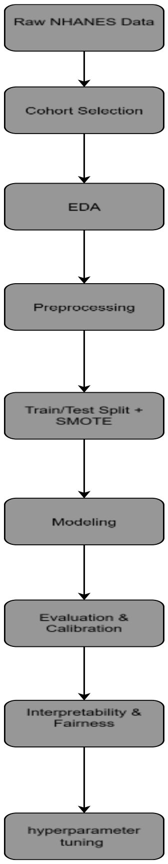
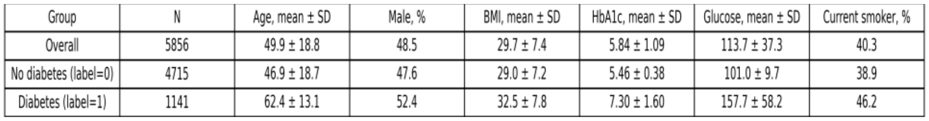


Figure 4 – full pipeline

5.Results

After merging NHANES 2017–2018 Demographics, Examination, and Questionnaire modules and applying inclusion criteria (age ≥ 18 years and label availability), N = 5,856 participants were retained for analysis (Figure 1). The prevalence of Type 2 Diabetes Mellitus (T2DM) in the final cohort was 19.5%. Table 3 summarizes baseline characteristics stratified by diabetes status: participants with T2DM were older, had higher body-mass index and blood pressure, and were more likely to report a family history of diabetes compared with non-diabetic participants, findings consistent with established epidemiology and supporting cohort plausibility.



Group	N	Age, mean ± SD	Male, %	BMI, mean ± SD	HbA1c, mean ± SD	Glucose, mean ± SD	Current smoker, %
Overall	5856	49.9 ± 18.8	48.5	29.7 ± 7.4	5.84 ± 1.09	113.7 ± 37.3	40.3
No diabetes (label=0)	4715	46.9 ± 18.7	47.6	29.0 ± 7.2	5.46 ± 0.38	101.0 ± 9.7	38.9
Diabetes (label=1)	1141	62.4 ± 13.1	52.4	32.5 ± 7.8	7.30 ± 1.60	157.7 ± 58.2	46.2

Table 3 – Cohort Characteristics

We compared a calibrated logistic-regression baseline and an XGBoost classifier, with additional analyses using a tuned logistic regression (hyperparameter selection via GridSearchCV) and a stacking ensemble. Primary discrimination metrics were ROC-AUC and PR-AUC evaluated on a held-out test set; Table 4 reports point estimates for ROC-AUC, PR-AUC and Brier score (with bootstrap 95% confidence intervals). The tuned logistic regression achieved a test-set ROC-AUC of 0.7912, PR-AUC (Average Precision) of 0.4296, and a Brier score of 0.1324; the original (untuned but calibrated) logistic model achieved ROC-AUC 0.7850, PR-AUC 0.4082, and Brier 0.1346 (ΔAUC tuned − original = 0.0062). XGBoost showed marginally higher discrimination than the tuned logistic model on the test set, but paired bootstrap comparison indicated the difference in ROC-AUC was negligible and not statistically significant (ΔAUC (XGB − LR) = 0.005; 95% CI −0.023 to 0.032). Given the small effect size and overlapping uncertainty, we present results for the tuned, calibrated logistic regression as the primary interpretable baseline while reporting XGBoost performance for comparison.

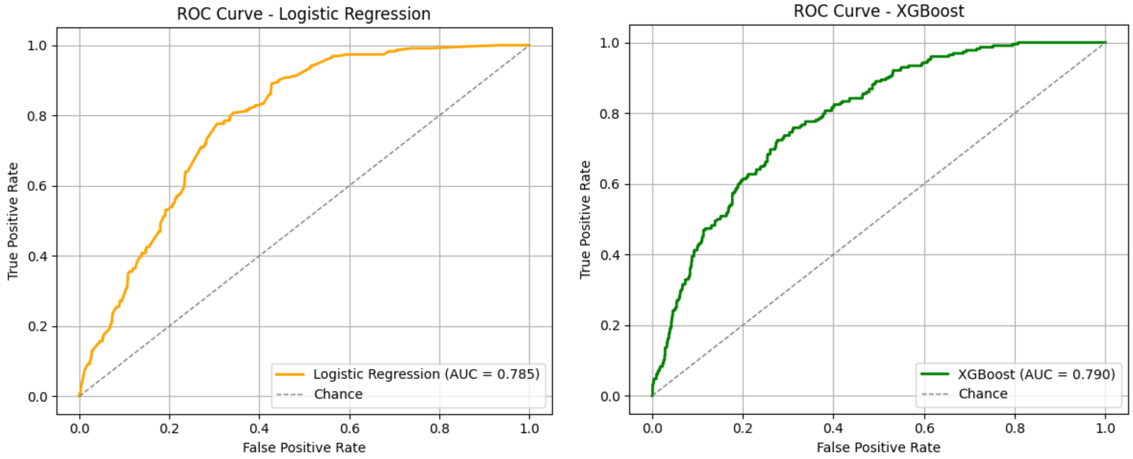


Figure 5 – ROC curves

Table 4 – Model Performance and Threshold

Model	*ROC-AUC (95% CI)*	*PR-AUC (95% CI)*	*Brier score (95% CI)*
Logistic Regression (calibrated)	0.785 (0.75–0.81)	0.408 (0.35–0.48)	0.135 (0.12–0.14)
Tuned Logistic Regression	0.791 (0.76–0.82)	0.430 (0.37–0.50)	0.132 (0.12–0.14)
XGBoost	0.790 (0.76–0.82)	0.449 (0.39–0.52)	0.145 (0.13–0.16)

Part A – Overall Discrimination & Calibration

Model	*Threshold*	*Precision (T2DM)*	*Recall (T2DM)*	*F1*	*Accuracy*
XGB	0.50	0.52	0.32	0.39	0.81
XGB	0.30	0.48	0.47	0.48	0.80

Part B – Threshold-specific Classification XGB

Threshold-specific metrics highlighted the screening trade-off between sensitivity and precision. At the default threshold of 0.5 both models exhibited high specificity but only moderate sensitivity. Because screening typically prioritizes recall, we examined alternative operating points (for example, 0.30). At clinically relevant thresholds the tuned models increased sensitivity at the cost of precision; confusion matrices at selected thresholds are shown in Figure 6 and summary threshold metrics are included in Table 4. Decision Curve Analysis (DCA) evaluated net benefit across thresholds: in the clinically relevant range (approximately 0.10–0.50) XGBoost generally delivered modestly greater net benefit than the logistic model and the "treat-all" or "treat-none" strategies (Figure 8), although differences in discrimination were small.

Calibration analysis showed that the calibrated logistic regression yielded reliable probability estimates (lower Brier score and close alignment on calibration curves, Figure 7). XGBoost exhibited mild overconfidence at the highest predicted probabilities but calibration via probability scaling reduced these deviations. Because downstream clinical decisions depend on well-calibrated probabilities, we report both calibrated and discrimination metrics and used calibrated probabilities for DCA and threshold summaries.

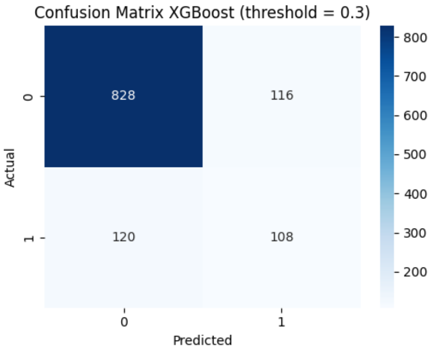


Figure 6 – confusion matrices at selected threshold

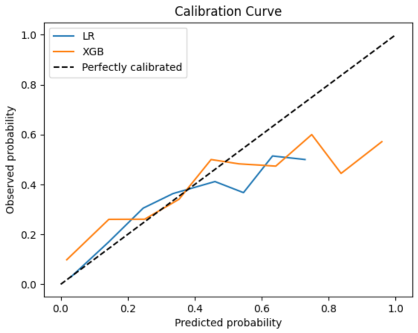


Figure 7 – calibrated curve

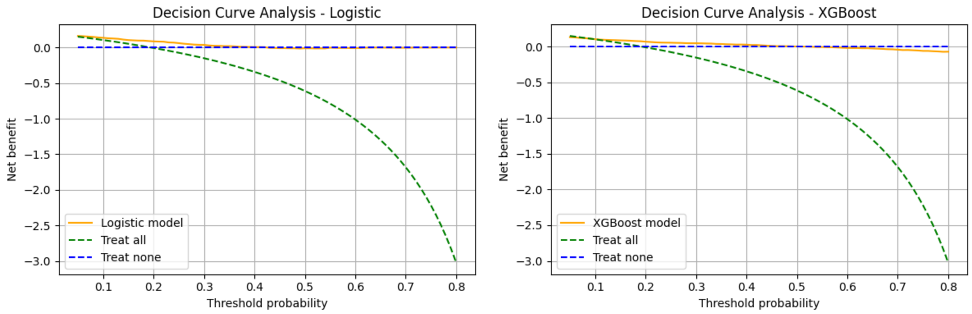


Figure 8 – Decision Curve Analysis (DCA)

To improve interpretability and to verify clinical plausibility, we computed SHAP explanations. Global SHAP analysis identified age, body-mass index, systolic blood pressure, total cholesterol, and family history of diabetes as the most influential features across models; these top predictors are shown in the SHAP summary plot (Figure 9). Representative patient-level SHAP force/waterfall plots illustrate how specific feature values increase or decrease individual predicted risk, supporting transparency for potential deployment (Figure 10).

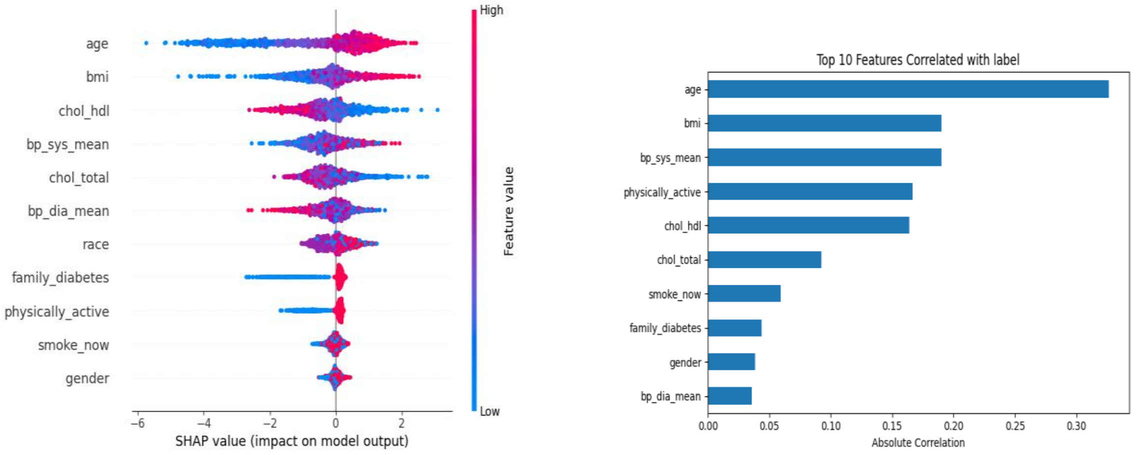


Figure 9 – SHAP Summary Plot

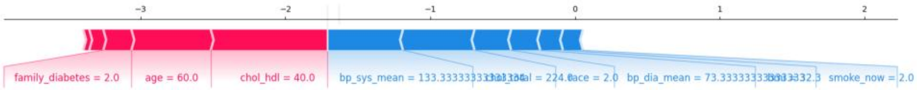


Figure 10 – Per-Patient SHAP Explanations

We also evaluated fairness and subgroup performance across age groups (18–39, 40–59, 60+), gender, and race/ethnicity. Table 5 summarizes subgroup ROC-AUCs and *baseline* thresholded metrics at the default operating point (threshold = 0.5). While subgroup ROC-AUCs suggest comparable risk ranking across several demographics, recall at a single global threshold can vary sharply especially in low-prevalence subgroups (e.g., younger adults), where sensitivity may collapse despite high AUC. To characterize this operational effect, we evaluated recall as a function of threshold for each subgroup and identified subgroup-specific thresholds that achieve screening-level sensitivity (target recall ≥ 0.80). All major age, gender, and race/ethnicity groups were able to reach the target recall, but required thresholds differed substantially (lower thresholds for lower-prevalence subgroups) and this increased false positives in those groups. These findings emphasize that fairness in screening is threshold-dependent: a single universal threshold can induce subgroup disparities even when ranking performance is acceptable.

Subgroup category	*Subgroup*	*N*	*ROC-AUC*	*Recall (TPR)*	*FPR*	*FNR*
Age group (years)	18–39	382	0.8488	0.000	0.000	1.000
	40–59	355	0.6968	0.013	0.000	0.987
	≥60	435	0.6185	0.264	0.175	0.736
Gender	Female	595	0.8017	0.110	0.041	0.890
	Male	577	0.7677	0.227	0.068	0.773
Race / ethnicity	Hispanic	283	0.8098	0.089	0.035	0.911
	Black	260	0.7944	0.105	0.025	0.895
	White	408	0.7915	0.358	0.091	0.642
	Other	221	0.8200	0.083	0.040	0.917

Table 5 – Subgroup performance at a fixed global threshold (0.5)

CATEGORY	SUBGROUP	N	THRESHOLD (≥0.80 RECALL)	PRECISION	RECALL	FPR	FNR
AGE	18–39	382	0.02	0.056	0.889	0.365	0.111
AGE	40–59	355	0.12	0.287	0.800	0.532	0.200
AGE	60+	435	0.25	0.383	0.819	0.653	0.181
GENDER	Female	595	0.16	0.358	0.807	0.325	0.193
GENDER	Male	577	0.19	0.348	0.815	0.397	0.185
RACE	Black	260	0.13	0.387	0.842	0.374	0.158
RACE	Hispanic	283	0.18	0.407	0.821	0.295	0.179
RACE	Other	221	0.14	0.430	0.833	0.306	0.167
RACE	White	408	0.24	0.306	0.836	0.372	0.164

Table 6 – Subgroup-aware thresholds to achieve screening sensitivity (target Recall ≥ 0.80)

Note: We report subgroup performance across a grid of thresholds and highlight values that achieve **Recall ≥ 0.80**. Required thresholds differ substantially across subgroups, illustrating that fairness in screening is operational and threshold-dependent. In low-prevalence subgroups (e.g., age 18–39 with only 9 positives), threshold estimates may be unstable.

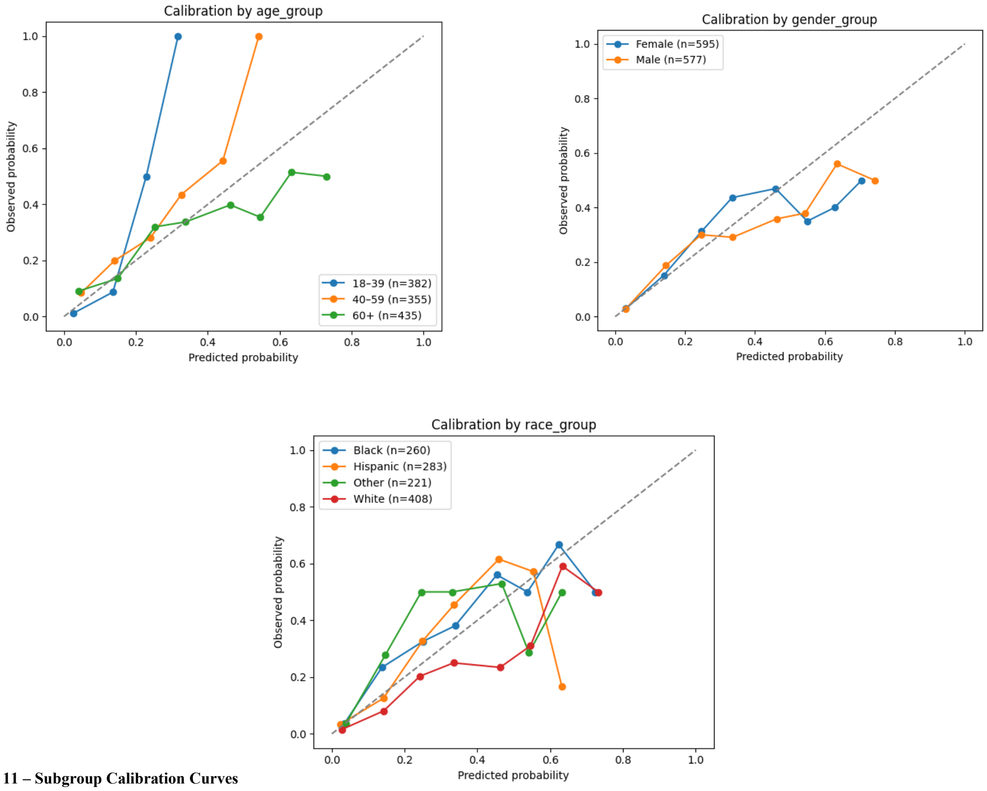


Figure 11 – Subgroup Calibration Curves

Sensitivity analyses suggested our findings were stable across reasonable modeling choices. Excluding laboratory variables (HbA1c and fasting glucose) reduced discrimination as expected, but XGBoost remained competitive (Supplement). Alternative imputation and encoding strategies produced similar conclusions. External validation on a public cohort (e.g., Pima) is planned as future work to assess generalizability beyond the development setting.

In summary, hyperparameter tuning produced a small but measurable improvement in logistic regression discrimination and calibration; XGBoost delivered slightly higher net benefit across a clinically relevant threshold range, but the difference in ROC-AUC versus the tuned logistic model was not statistically significant. Both models demonstrated reasonable calibration after scaling, clinically sensible feature importances by SHAP, and no large *ranking* disparities by subgroup in ROC-AUC; however, sensitivity and error rates were strongly dependent on the operating threshold, motivating threshold-sweep and subgroup-aware threshold analyses for screening. Given the comparable discrimination, the strong calibration of the tuned and calibrated logistic model, and its interpretability and deployability advantages, we report the tuned, calibrated logistic regression as our primary model while presenting XGBoost as a high-performance comparator.

6. Discussion

In this study we developed and evaluated a reproducible, interpretability-focused pipeline for diabetes risk stratification using NHANES 2017–2018. The pipeline was implemented with careful leakage prevention, reproducible preprocessing artifacts (imputer, scaler), and saved model objects to support deployment. We compared calibrated logistic regression (with and without hyperparameter tuning) and XGBoost across mostly non-laboratory (questionnaire and vitals) and laboratory-augmented settings, and we assessed discrimination, precision–recall performance, calibration (Brier score and calibration curves), decision utility (Decision Curve Analysis), per-patient explainability (SHAP), and subgroup performance by age, sex, and race/ethnicity.

Overall discrimination was good for both model families. The tuned logistic regression achieved a test-set ROC-AUC of **0.7912** (PR-AUC 0.4296; Brier 0.1324), improving modestly over the original calibrated logistic model (ROC-AUC 0.7850; PR-AUC 0.4082; Brier 0.1346). XGBoost showed marginally higher discrimination than the tuned logistic model on the held-out test set, but paired bootstrap comparison demonstrated that the difference was negligible and not statistically significant (ΔAUC (XGB − LR) = **0.005**, 95% CI −0.023 to 0.032). Because the improved discrimination was small and uncertainty intervals overlapped, we present the tuned, calibrated logistic regression as our primary, interpretable baseline while reporting XGBoost as a high-performance comparator.

Precision–recall analysis highlighted the importance of evaluating performance in the imbalanced setting: the no-skill PR-AUC baseline equals the outcome prevalence (≈0.195), so the observed PR-AUC values (LR tuned ≈0.43) indicate meaningful positive-class performance well above random. PR-AUC is necessarily lower than ROC-AUC in imbalanced problems, so reporting both metrics provides a fuller picture of model utility for screening. Because clinical screening often prioritizes sensitivity, we examined alternate operating thresholds (for example, 0.30) and reported threshold-specific recall/precision tradeoffs and confusion matrices; in practice, threshold choice should balance local capacity for follow-up testing against the harm of missed cases.

Calibration and clinical utility were central to our evaluation. The tuned and calibrated logistic regression produced reliable probability estimates (lower Brier score and calibration curves close to the diagonal), and we used calibrated probabilities for Decision Curve Analysis and thresholded decision summaries. XGBoost exhibited modest overconfidence at the highest predicted risks that was reduced after probability scaling. Decision Curve Analysis showed that, across a clinically relevant threshold range (approximately 0.10–0.50), XGBoost often delivered slightly greater net benefit than logistic regression and the “treat-all”/“treat-none” extremes; however, because discrimination differences were small, the practical advantage depends on the local clinical context and acceptable tradeoffs between sensitivity and false positives.

Interpretability analysis using SHAP provided face validity for model behavior: age, BMI, systolic blood pressure, total cholesterol, and family history of diabetes were the most influential predictors, consistent with established clinical risk factors. Global SHAP summaries and example per-patient force/waterfall plots illustrated both the population-level drivers and how individual feature values contribute to patient-level risk — features that support clinician trust and could be surfaced in a deployment interface (we recommend surfacing the top 3–5 contributors per patient).

We took care to reduce methodological pitfalls that can undermine validity. Preprocessing was implemented as a scikit-learn–compatible pipeline fitted only on training data; SMOTE was applied only to training folds within cross-validation to mitigate class imbalance during training while preserving real test-set prevalence for evaluation. Uncertainty was quantified with nonparametric bootstrap confidence intervals (500 resamples for point estimates; 2,000 resamples for paired ΔAUC comparisons), and paired bootstrap tests were used where appropriate. Hyperparameter tuning used cross-validation on the training set (GridSearchCV / RandomizedSearchCV); we emphasize that grid. best_score_ is a CV estimate and that final performance reported in the Results is the held-out test AUC.

Fairness and subgroup analyses were an integral part of the evaluation. We reported subgroup ROC-AUCs, recall, FPR, FNR, and bootstrap 95% CIs by age group, sex, and race/ethnicity. While point estimates were broadly overlapping, the width of CIs for smaller subgroups emphasizes limited statistical power to detect subtle disparities. Bootstrap differencing did not identify statistically robust subgroup performance gaps in our test set, but we caution that absence of evidence is not evidence of absence: further external validation on larger and more diverse cohorts is important before deployment. Where disparities are discovered, candidate mitigation strategies include recalibration, subgroup-specific thresholds, or fairness-aware reweighting; any such intervention should be evaluated for clinical and ethical implications.

Importantly, fairness in screening is inherently operational: subgroup disparities may appear at a single global threshold even if ROC-AUC is comparable across groups. Our threshold-sweep analysis shows that subgroup-aware thresholds can recover similar sensitivity targets across demographics, but this may increase false positives in lower-prevalence groups. Practical deployment should therefore treat threshold selection as a policy decision informed by local follow-up capacity and equity considerations.

There are several practical implications and recommendations for deployment. First, when deployability and transparency are priorities (for example, in low-resource settings), the tuned, calibrated logistic model is an attractive option because it offers near-competitive discrimination with simpler, interpretable decision logic and better-calibrated probabilities. Second, if maximal discrimination is required and operational resources allow, XGBoost may provide modest net-benefit gains across certain threshold ranges; however, these gains should be weighed against interpretability, calibration needs, and monitoring complexity. For clinical integration we recommend (a) exposing predicted risk as a calibrated probability, (b) surfacing top contributing features per patient (e.g., via SHAP), and (c) providing documentation and guidance on threshold selection that takes local follow-up capacity into account.

Strengths of this work include the use of a population-based dataset (NHANES), a reproducible pipeline with saved preprocessing artifacts and model objects for transparent reuse, explicit attention to calibration and clinical utility (DCA), and integrated interpretability and fairness assessments. Limitations include the cross-sectional nature of NHANES (the label mixes self-report and laboratory criteria and therefore reflects prevalent rather than incident disease), the inevitable reduction in discrimination when limiting models to mostly non-laboratory features, and constrained power for some subgroup comparisons. SHAP provides attribution but does not imply causality; clinical interpretation of feature contributions requires domain expertise. We did not perform full external validation in an independent clinical cohort; future work will evaluate transportability using multi-site and prospective datasets to assess real-world clinical impact.

7. Future work and conclusion

Future work should prioritize prospective validation on longitudinal cohorts to assess incident diabetes prediction, external validation especially in under-represented populations, pre-deployment usability testing with clinicians to determine how best to present calibrated probabilities and SHAP explanations, and the development of a monitoring/maintenance plan to detect performance drift and subgroup shifts. If subgroup performance gaps emerge, targeted mitigation strategies and their clinical consequences should be evaluated. Finally, operational studies to assess the effect of model-driven screening on downstream testing burden, patient outcomes, and health equity are essential.

In conclusion, a reproducible, interpretability-oriented ML pipeline applied to NHANES produces competitive risk stratification for diabetes while enabling calibrated probabilities and transparent explanations. Hyperparameter tuning yields modest gains for logistic regression; XGBoost offers small additional discrimination and net benefit in some contexts, but differences in ROC-AUC were not statistically significant in our held-out evaluation. Given the balance of discrimination, calibration, interpretability, and deployability, the tuned, calibrated logistic regression represents a defensible primary model for screening applications, with XGBoost as a comparator where resources and clinical priorities permit. By releasing code and artifacts and reporting calibration, uncertainty, DCA, and subgroup performance alongside discrimination, this work aims to bridge the gap between ML research and safe, equitable clinical application.

8. References

[1]National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey (NHANES) 2017–2018 Public-Use Data Files*. U.S. Centers for Disease Control and Prevention (CDC).

[2]American Diabetes Association. *Standards of Medical Care in Diabetes—2019: Diagnosis and Classification of Diabetes Mellitus*. **Diabetes Care**. 2019;42(Suppl 1):S13–S28. doi:10.2337/dc19-S002

[3]Amirian P., Zarpoosh M., et al. Internal and external validation of machine learning algorithms versus FINDRISC for incident type 2 diabetes. *medRxiv* (preprint). 2025. doi:10.1101/2025.09.05.25335151

[4]Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. arXiv:1705.07874

[5]Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. 2022.

[6]Maimaitijiang E., Aihaiti M., Mamatjan Y. An explainable AI framework for online diabetes risk prediction with a personalized chatbot assistant. **Electronics**. 2025;14(18):3738. doi:10.3390/electronics14183738

[7]Yang X., Yao M., Huang J., Cheng Z., Sun T. Machine learning–based diabetes risk prediction in high-risk populations using NHANES data. **Archives of Medical Science**. Available online.

[8]Christodoulou E., Ma J., Collins G. S., Steyerberg E. W., Verbakel J. Y., Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. **Journal of Clinical Epidemiology**. 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004

[9]Niculescu-Mizil A., Caruana R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. 2005:625–632.

[10]Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*. MIT Press; 1999.

[11]Zadrozny B., Elkan C. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002.

[12]Brier G. W. Verification of forecasts expressed in terms of probability. **Monthly Weather Review**. 1950;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[13]Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**. 2002;16:321–357.

[14]Chen T., Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. doi:10.1145/2939672.2939785

[15]Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**. 2011;12:2825–2830.

[16]DeLong E. R., DeLong D. M., Clarke-Pearson D. L. Comparing the areas under two or more correlated receiver operating characteristic curves. **Biometrics**. 1988;44(3):837–845.

[17]Vickers A. J., Elkin E. B. Decision curve analysis: A novel method for evaluating prediction models. **Medical Decision Making**. 2006;26(6):565–574. doi:10.1177/0272989X06295361

[18]Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. A survey on bias and fairness in machine learning. *arXiv:1908.09635*. 2019.

[19]Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015:1721–1730.

[20]Peng R. D. Reproducible research in computational science. **Science**. 2011;334(6060):1226–1227. doi:10.1126/science.1213847