TECHNICAL REPORT / MACHINE LEARNING

Diabetes risk estimation in Mexican patients

A model to prioritize confirmatory diabetes tests without using glucose or HbA1c, with household-grouped validation, bias analysis and an operating plan.

Alan Adelfo Del Angel Figueroa
Software Development Engineering, Universidad Tecmilenio
Final project for Machine Learning Algorithms / September 2026
Code: github.com/AlanDelAngel/diabetes_pacientes_mexicanos

Summary

This project builds a machine learning model that ranks Mexican adult patients by their probability of being at high risk of diabetes, using data that is usually available before a glucose test: age, sex, weight, height and a biochemical profile (lipids, insulin, uric acid, creatinine and albumin). The goal is to decide who should get a confirmatory test (fasting glucose or HbA1c) first when that resource is limited. **The model does not diagnose**: it supports a decision that is always reviewed by healthcare staff.

The final model is a support vector machine (SVM) with an RBF kernel, tuned through hyperparameter search and evaluated with household-grouped cross-validation. On a test set of 365 patients from households the model never saw, with a threshold of 0.26 chosen to guarantee a recall of at least 80%, it achieves:

0.825	0.765	0.481	0.764
High-risk recall (99 of 120 cases detected)	Test ROC-AUC (baseline model: 0.505)	Precision: acceptable in a screening setting with review	Cross-validation ROC-AUC (+/- 0.027, 15 folds)

Performance is even between men and women (recall of 0.83 and 0.82), but drops for adults aged 20 to 39 (0.47 on the test set). That limitation is documented in the model card and in the risk register.

> **How to use this report with the repository.** The whole analysis is in the notebook `Proyecto_Final_Riesgo_Diabetes.ipynb`. Each section of this report points to the notebook section where its figures and numbers are generated. When it runs, the notebook creates the `scripts/`, `artifacts/`, `results/` and `docs/` folders with the files mentioned here (section 11). Variable and file names keep their original Spanish names so they match the dataset and the code.

Contents

1. Problem and success criterion
2. Data and data quality
3. Validation design
4. Pipeline and feature engineering
5. Model comparison and tuning
6. Decision threshold
7. Test set results
8. Interpretability
9. Errors, bias and responsible use
10. Putting the model into operation
11. Reproducibility
12. Model card

1. Problem and success criterion

The target variable is `riesgo_diabetes_cat`, with three levels: 0 = low, 1 = moderate and 2 = high. The operating decision is binary: identify **high-risk** patients versus the rest, because they benefit the most from an early confirmatory test.

Success criterion: detect high risk without using glucose or HbA1c, with a ROC-AUC clearly above a baseline model (`DummyClassifier`) and a high-risk recall of at least 0.80. Recall comes first because, in

healthcare, missing a high-risk patient costs more than an extra confirmatory test.

2. Data and data quality

Notebook: Part I (loading, glossary, exploratory analysis and data quality).

The project uses the *Diabetes Mexico Data Set* (Félix Jiménez & Sánchez Lee, 2026), derived from an ENSANUT-style national health survey: **2,549 adults from 2,055 households** and 24 columns. The target is highly imbalanced: 1,697 low-risk patients (66.6%), 12 moderate-risk (0.5%) and 840 high-risk (33.0%).

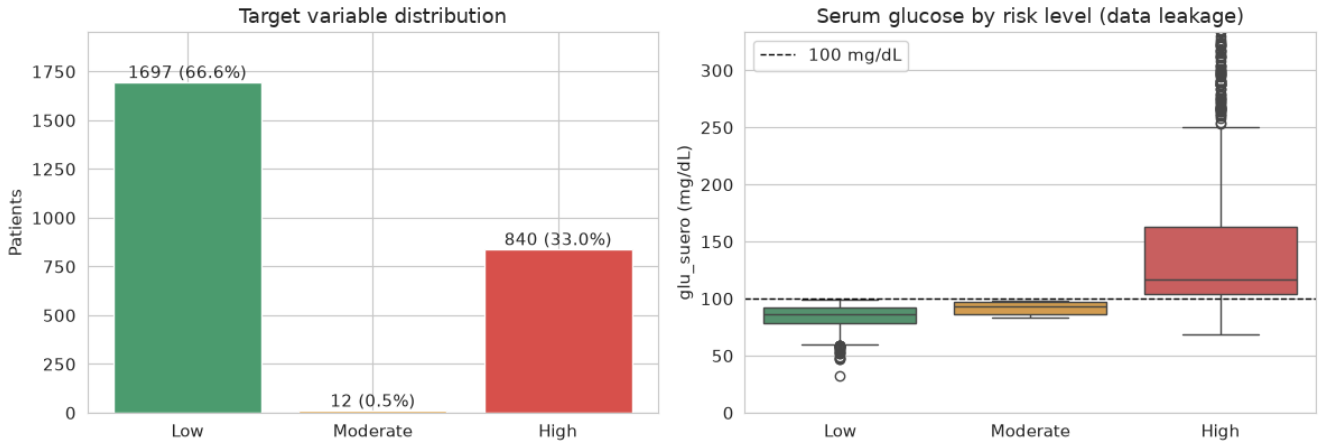


Figure 1. Target distribution (left) and serum glucose by risk level (right), which shows why glucose causes data leakage.

Data leakage

Serum glucose (`glu_suero`) and HbA1c (`hb1ac`) are the variables used to build the target. Including them would produce a model that looks almost perfect on paper and is useless in practice, since the whole point is to prioritize who should get a glucose test. That is why they **were excluded**, together with the TyG and HOMA-IR indices, which also depend on glucose.

Data quality

- **Sentinel values:** 11 records with weight = 222.22 and 15 with height = 222.2 produced fake BMI values close to 45. They were treated as missing.
- **Missing values:** 28% of patients have no weight or height; the original `imc` column is 98.6% missing and was replaced with a calculated BMI.
- **No duplicates**; `folio_int` uniquely identifies each patient.

Excluded variables	Reason
`glu_suero`, `hb1ac`	Data leakage (they define the target)
`folio_i`, `folio_int`	Identifiers (the household ID is only used to group the validation)
`Ciudad`, `entidad_cve`	High cardinality; the model without state performs the same and is easier to interpret (section 5)
`imc`, `muestra_suero`	98.6% missing; constant variable
`ponde_venosa`, `estrato`, `est_sel`, `ponde_hemo`	Survey sampling design variables

3. Validation design

Notebook: Part I (household split) and Part II, section 6.

- **Household split:** members of the same household share diet, physical activity and environment. Using `StratifiedGroupKFold` (seed 42), the data was split into a **development set of 2,184 patients** and a **test set of 365**, with no household shared between them.
- **Repeated cross-validation:** 5 household-grouped folds, repeated with 3 seeds (42, 43 and 44), for a total of 15 evaluations per model. A single split was unstable: with so few moderate-risk cases, a fold could have one or none.
- **Primary metric:** high-risk vs. rest ROC-AUC. Secondary: high-risk recall, precision and F1, accuracy and macro F1.
- **Untouched test set:** it was not used to compare models, select features, tune hyperparameters or choose the threshold. It was used only once, in section 7.

4. Pipeline and feature engineering

Notebook: Part II, sections 4 and 5. Transformer code: `scripts/diabetes_transformers.py`.

All preprocessing lives in a single scikit-learn `Pipeline`, so training and inference apply exactly the same rules:

1. `ConstructorCaracteristicas`: sentinel cleaning and creation of new features.
2. `ColumnTransformer` with four branches: numeric (p0.5-p99.5 winsorizing, median imputation and scaling), skewed numeric (the same plus `log1p`), binary (One-Hot) and ordinal (`OrdinalEncoder`).
3. The model.

All parameters (winsorizing limits, medians, categories) are fitted only on the training data of each fold.

New feature	Construction	Clinical meaning	Leakage risk
`ratio_trig_hdl`	trig / col_hdl	Indirect marker of insulin resistance	Low
`col_no_hdl`	colest - col_hdl	Atherogenic cholesterol (LDL + VLDL)	Low
`grupo_edad`	20-39, 40-59, 60+	Life stages with increasing risk	None
`categoria_imc`	<25, 25-29.9, >=30 (WHO)	Normal weight, overweight, obesity	None

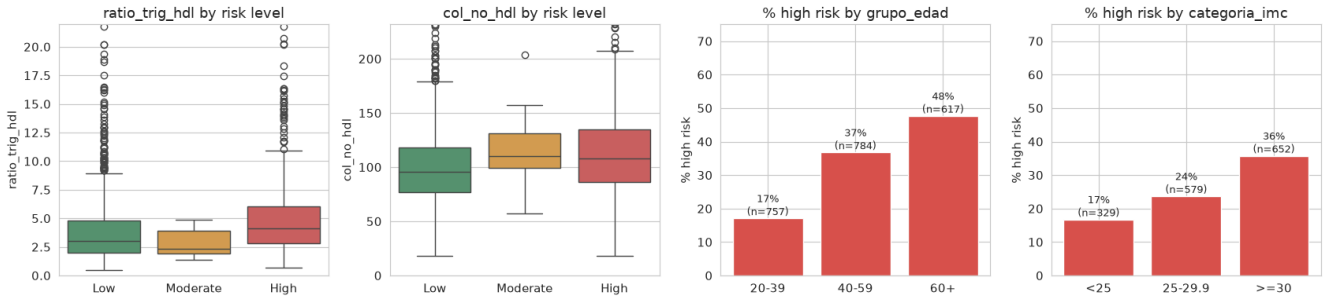


Figure 2. Relationship between the new features and risk level (development data).

5. Model comparison and tuning

Notebook: Part II, sections 7 to 14. Full log: `results/registro_experimentos.csv`.

5.1 Model families

Eight models were compared with the same pipeline and the same cross-validation (values with each model's default configuration):

Model	Family	High ROC-AUC	Std.	High recall	High F1	Inference (ms/1000)
DummyClassifier	Baseline	0.505	0.026	0.338	0.337	22.2
Logistic regression	Linear	0.723	0.022	0.642	0.566	17.5
Decision tree	Tree	0.674	0.033	0.608	0.525	27.7
Random Forest	Ensemble	0.756	0.026	0.579	0.561	72.3
Gradient Boosting (HGB)	Ensemble	0.746	0.026	0.642	0.571	35.0
Linear SVM	Margin-based	0.752	0.023	0.684	0.576	47.8
RBF SVM	**Margin-based**	**0.764**	**0.026**	**0.734**	**0.601**	**133.0**
Neural network (MLP)	Neural network	0.720	0.033	0.306	0.378	16.6

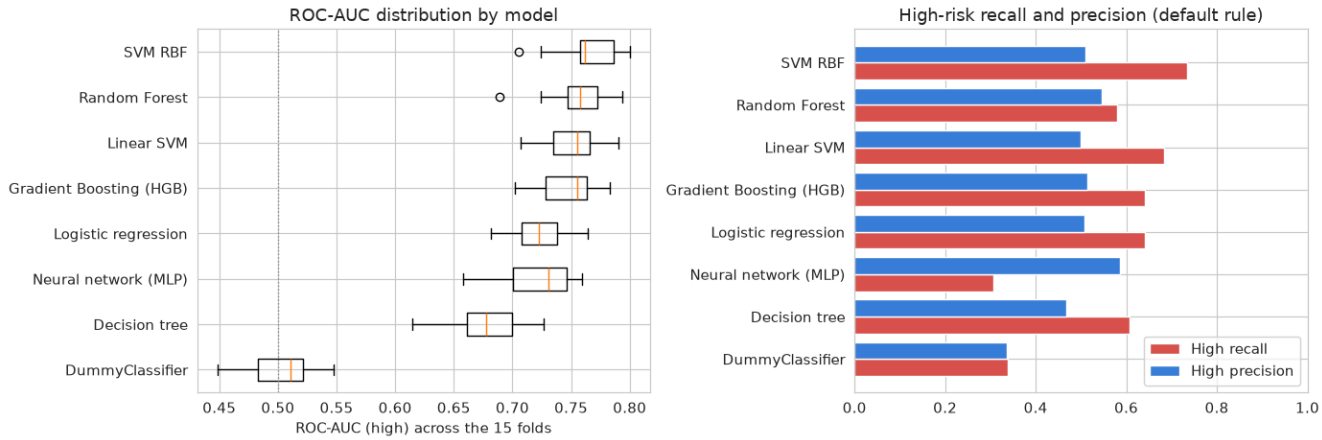


Figure 3. High-risk ROC-AUC across the 15 folds (left) and recall and precision with the default rule (right).

5.2 Feature selection

Two approaches were tested inside the cross-validation: `SelectFromModel` with Random Forest importances, and a reduced model without the state variable (15 clinical and demographic features).

Configuration	Features	High ROC-AUC	Std.	Interpretability
RBF SVM, full model	41	0.764	0.026	Medium: includes One-Hot states
RBF SVM + SelectFromModel	21	0.761	0.027	Medium: the selection changes with every retraining
RBF SVM, reduced model	**15**	**0.762**	**0.028**	**High: only clinical and demographic features**
Random Forest, full model	41	0.756	0.026	Medium
Random Forest, reduced model	15	0.755	0.027	High

The reduced model practically matches the full one (a difference of 0.002 or less in ROC-AUC), removes the state columns, which are hard to explain clinically, and is cheaper to run. It is used from here on.

5.3 Hyperparameter tuning

Model	Search space	Combinations	Best configuration	High ROC-AUC
RBF SVM (GridSearchCV)	**C {0.1, 0.3, 1, 3, 10} x gamma {scale, 0.01, 0.03, 0.1}**	**20**	**C = 3, gamma = 0.03**	**0.764**
Linear SVM (GridSearchCV)	C {0.01, 0.03, 0.1, 0.3, 1}	5	C = 0.1	0.747
Random Forest (RandomizedSearchCV)	n_estimators, max_depth, min_samples_leaf, max_features	20	400 trees, min_samples_leaf = 3, max_features = 0.3	0.757

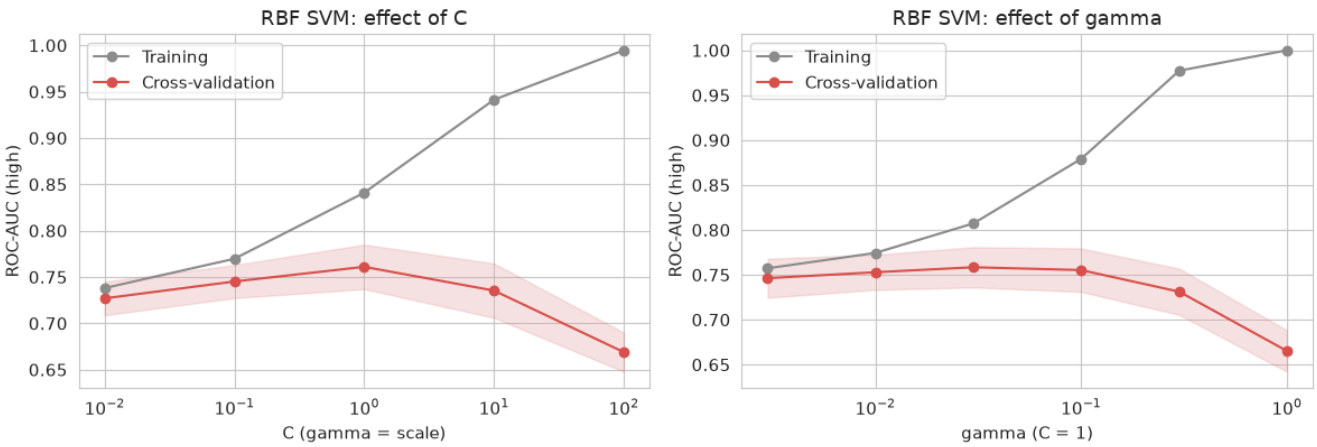


Figure 4. RBF SVM validation curves: higher C or gamma values raise the training ROC-AUC but lower the validation ROC-AUC (overfitting).

5.4 Decision matrix between finalists

The two best tuned models from different families were chosen as finalists, both with the reduced pipeline. The decision was not based only on the highest metric: each criterion was scored from 1 to 5 using cross-validation evidence.

Criterion	Weight	RBF SVM	Random Forest	Evidence (cross-validation)
Recall	25%	5.0	2.64	0.710 vs. 0.491 (baseline 0.338)
ROC-AUC	20%	5.0	4.91	0.764 vs. 0.758 (baseline 0.505)
F1-score	15%	5.0	3.98	0.590 vs. 0.526
Precision	10%	3.92	5.0	0.506 vs. 0.569
Stability	10%	4.9	5.0	ROC-AUC standard deviation: 0.027 vs. 0.026
Interpretability	10%	3.0	4.0	RF supports exact SHAP; the SVM only model-agnostic methods
Operating cost	10%	4.0	5.0	Inference 82 vs. 65 ms/1000; artifact 0.27 vs. 18.16 MB
Weighted total	**100%**	**4.58**	**4.14**	**The recall advantage, the highest-weighted criterion, decides**

6. Decision threshold

Notebook: Part II, section 15. File: `artifacts/umbral_v1.json`.

The threshold was chosen with out-of-fold predictions from the development set (each patient scored by a model that did not see them during training), without touching the test set. Rule: among the thresholds that detect at least 80% of high-risk cases, choose the highest one, which produces the fewest false positives. For reference, the Youden index and the minimum-cost threshold (a false negative costs 3 times as much as a false positive) were also calculated.

Threshold	Precision	Recall	F1	Patients flagged	Criterion
0.21	0.448	0.882	0.595	64.8%	Minimum cost
0.26	**0.470**	**0.813**	**0.595**	**57.0%**	**Chosen: recall >= 0.80**
0.30	0.495	0.756	0.598	50.3%	
0.37	0.539	0.661	0.594	40.4%	Youden index
0.50	0.587	0.440	0.503	24.7%	Default rule

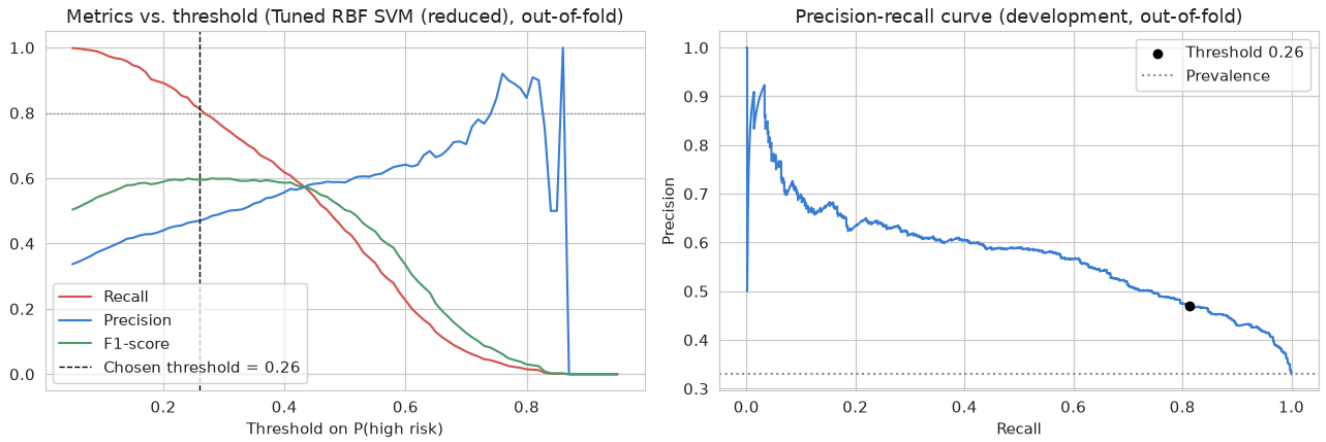


Figure 5. Effect of the threshold on precision, recall and F1 (left) and precision-recall curve with the chosen threshold (right), out-of-fold.

7. Test set results

Notebook: Part II, sections 16 to 18. File: `results/evaluacion_prueba_finalistas.csv`.

Each finalist was trained on the full development set and evaluated only once on the 365 test patients, with its operating threshold:

Model	Threshold	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Tuned RBF SVM (reduced)	**0.26**	**0.649**	**0.481**	**0.825**	**0.607**	**0.765**	**0.588**
Tuned Random Forest (reduced)	0.34	0.660	0.490	0.808	0.610	0.759	0.589

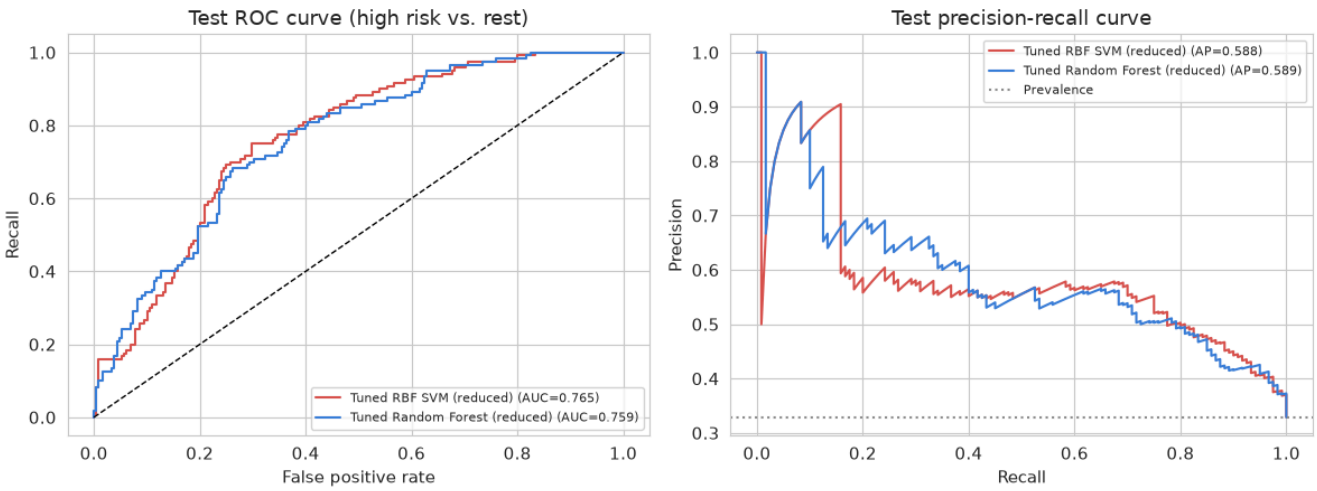


Figure 6. ROC and precision-recall curves of the finalists on the test set.

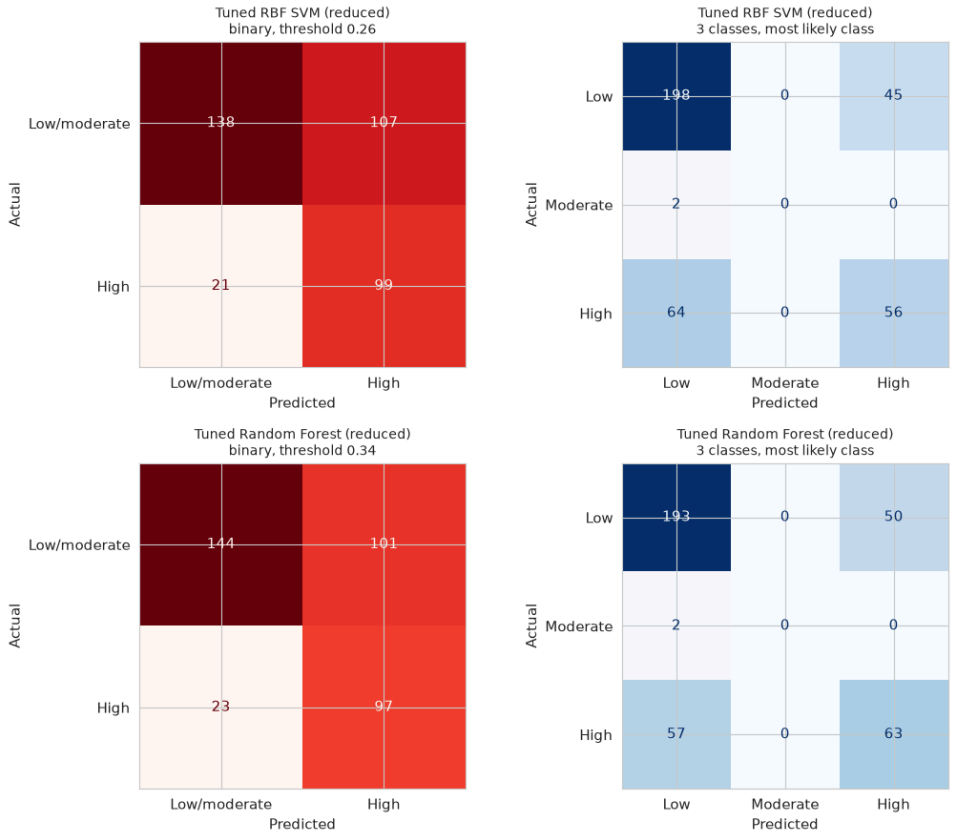


Figure 7. Test set confusion matrices: binary decision with the operating threshold (left) and 3-class classification with the most likely class (right).

Final model

The **tuned RBF SVM (reduced)** was selected:

Criterion	Evidence	Conclusion
Performance	ROC-AUC of 0.764 in cross-validation and 0.765 on the test set (Random Forest: 0.758 and 0.759)	Equal to or better than the other finalist and far above the baseline model (0.505)
Generalization	Test ROC-AUC within the cross-validation range (0.764 +/- 0.027)	No signs of overfitting
Errors	Recall of 0.825 with threshold 0.26: 21 false negatives and 107 false positives	Meets the priority of not missing high-risk cases
Complexity	15 input features; 0.5 s training time	Moderate, with no regional variables
Operation	About 82 ms of inference per 1,000 records; 12 input fields	Suitable for batch inference

The model does not detect the moderate class (12 cases in the whole dataset), so it is only used for the high-risk vs. rest decision.

8. Interpretability

Notebook: Part II, sections 19 and 20.

Two complementary methods were used: **permutation importance** on the full pipeline (how much the test ROC-AUC drops when each feature is shuffled, with 20 repetitions) and **SHAP** (KernelSHAP) on the transformed features.

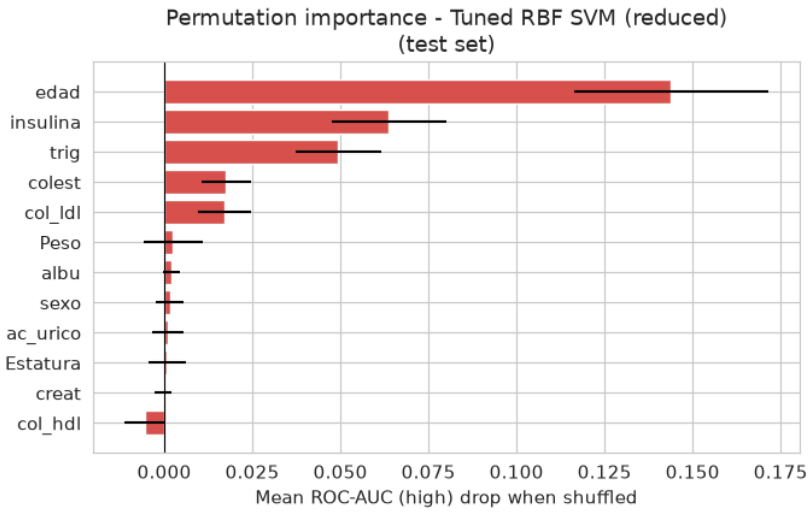


Figure 8. Permutation importance of the final model on the test set.

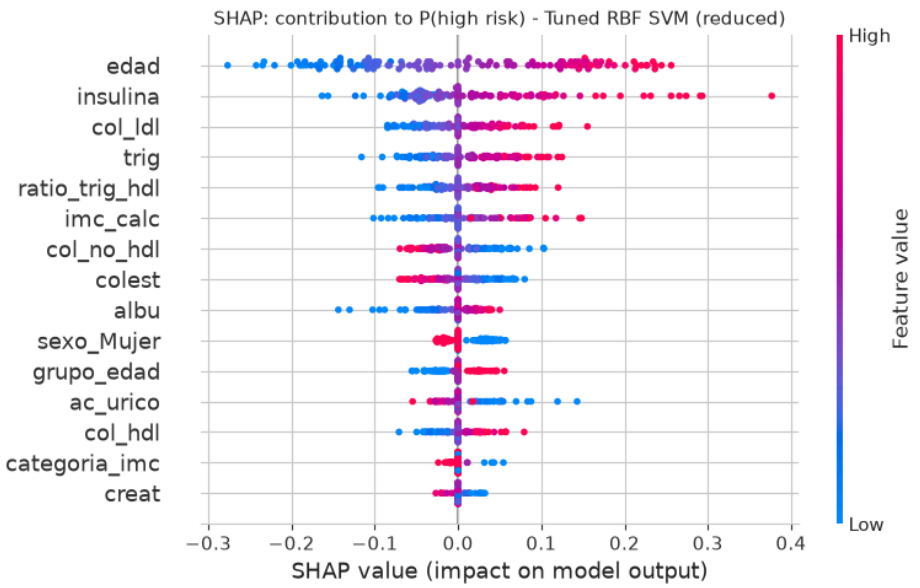


Figure 9. Contribution of each feature to the probability of high risk (SHAP). Red: high feature values; blue: low values.

Age is by far the most influential feature (a 0.144 drop in ROC-AUC when shuffled), followed by **insulin** (0.064) and **triglycerides** (0.049). High values of age, insulin, triglycerides, LDL cholesterol, triglyceride/HDL ratio and BMI push the probability toward high risk. High values of non-HDL cholesterol and total cholesterol push it slightly down; since total cholesterol and LDL are highly correlated (0.88), these contributions should be read together rather than separately.

Limitations. Permutation importance underestimates correlated features, because when one is shuffled the model compensates with the other. KernelSHAP is an approximation with a summarized background and assumes feature independence. Neither method implies causality: they show which features the model uses, not what would change a patient's actual risk.

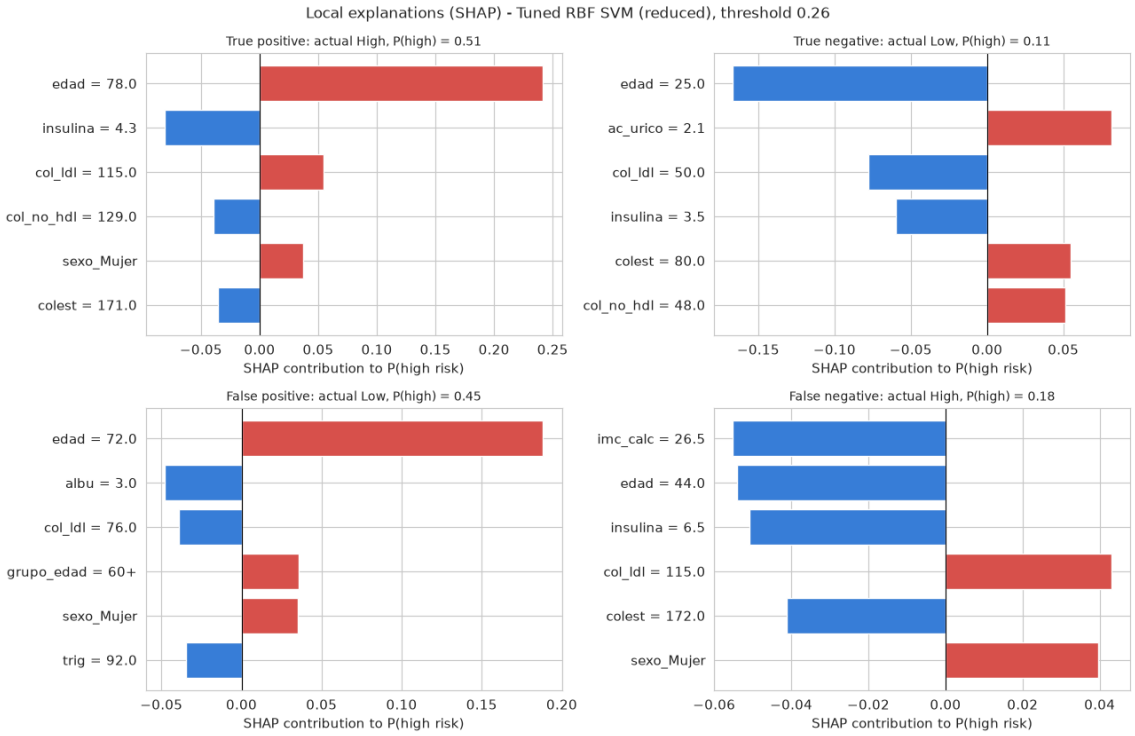


Figure 10. Individual explanations for a representative case of each outcome type (threshold 0.26).

9. Errors, bias and responsible use

Notebook: Part II, sections 17, 21 and 22. Files: `results/casos_error.csv` *and* `results/analisis_sesgos.csv`.

9.1 Error analysis

On the test set, with threshold 0.26, the model gets 99 true positives, 138 true negatives, 107 false positives and 21 false negatives. Typical false negatives are relatively young patients with lipid and biochemical profiles in the normal range and, in several cases, no BMI recorded: exactly the high-risk patients who do not show the usual signals. False positives tend to have high insulin or uric acid and be older, factors the model correctly associates with higher risk.

Error	Impact in this problem	How to reduce it
False negative	The most costly: a high-risk patient gets no follow-up. It happens more often in young patients without BMI	Lower the threshold for that group or make weight and height capture mandatory
False positive	An extra confirmatory test, with a lower clinical cost	Acceptable under the recall >= 0.80 criterion; communicated as a priority, not a diagnosis

9.2 Bias assessment

Groups by sex and age were compared using two sources: the test set (final evaluation, small groups) and the out-of-fold predictions from the development set (more patients, more stable).

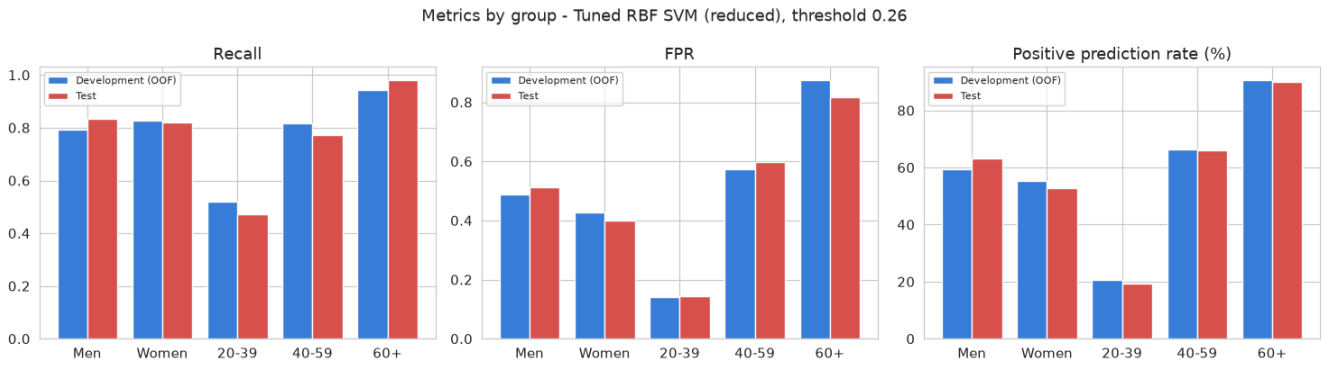


Figure 11. Recall, false positive rate and positive prediction rate by group.

Group (test)	Patients	Actual high risk	Recall	Precision	FPR	FNR
Men	128	37.5%	0.833	0.494	0.512	0.167
Women	237	30.4%	0.819	0.472	0.400	0.181
Aged 20-39	130	14.6%	0.474	0.360	0.144	0.526
Aged 40-59	121	36.4%	0.773	0.425	0.597	0.227
Aged 60 or older	111	50.5%	0.982	0.550	0.818	0.018

By sex, performance is even: recall differs by 1.4 percentage points. By age, the difference is marked: recall for adults aged 20 to 39 (0.47 on the test set, 0.52 out-of-fold) is much lower than for those over 60 (0.98 and 0.94). This is partly explained by high risk being much less frequent among younger patients (14.6% vs. 50.5%) and by their less clear biochemical markers. The report does not claim the model discriminates unfairly, but it documents this as an operating limitation that healthcare staff must know about.

9.3 Risk register

Risk	Cause	Consequence	Likelihood	Impact	Mitigation
False negatives	High risk with a "normal" profile; no BMI	At-risk patient without a confirmatory test	Medium	High	Threshold with recall >= 0.80; does not replace regular screening
False positives	Low threshold to favor recall	Extra tests and concern	High	Low to medium	Communicate as "review priority"
Data drift	Changes in the population or the laboratory	Miscalibrated probabilities	Medium	High	Monthly PSI and recall monitoring
Incomplete data	Missing weight and height (28%)	Less reliable imputation	High	Medium	Input validation with warnings
Automated use	Treating the classification as a final decision	Denying or delaying care	Medium	High	Mandatory human review
Group bias	Different prevalence and data quality	Lower recall for younger patients	Medium	Medium to high	Quarterly report by group; review if the gap exceeds 0.10

9.4 Conditions of use

- **Intended uses:** prioritizing confirmatory tests and preventive actions in a population that already has a lipid profile; supporting campaign planning.
- **Non-recommended uses:** diagnosing diabetes; denying, delaying or conditioning care, insurance or employment; applying it to people under 20, pregnant women or populations other than the survey's; estimating moderate risk.
- **Human review:** every result is reviewed by healthcare staff; cases close to the threshold (+/- 0.05) and records with warnings get additional review.
- **Model suspension:** high-risk recall below 0.70 in two consecutive monthly cuts, PSI above 0.25 in a main feature, a recall gap between groups above 0.15, changes in the target definition or a failed artifact verification.

10. Putting the model into operation

Notebook: Part II, sections 23 to 29.

10.1 Artifacts

The full pipeline is serialized with Joblib together with the target encoding and the threshold. Each artifact is registered with its version and SHA-256 hash to verify its integrity before loading it.

Artifact	File	Size
Final pipeline	`artifacts/pipeline_riesgo_diabetes_v1.joblib`	258.9 KB
Classifier	`artifacts/modelo_clasificador_v1.joblib`	252.1 KB
Target encoding	`artifacts/codificacion_objetivo_v1.json`	0.1 KB
Threshold	`artifacts/umbral_v1.json`	0.1 KB
Metadata and input schema	`artifacts/metadatos_modelo_v1.json`	4.1 KB

> **Security:** .joblib files can execute code when loaded. Artifacts from unverified sources must never be loaded; before every load, the SHA-256 hash is compared with the registered one.

10.2 Inference with input validation

The standalone script `scripts/inferencia_riesgo_diabetes.py` loads the artifact, receives new records (dictionary or CSV), validates their structure, calculates the probability of high risk, applies the threshold and returns the classification ("high priority" or "no high priority") with a human review reminder. Its predictions exactly match those of the model before serialization on the 365 test records.

Test case	Input	Expected result	Actual result
Valid record	Complete record	Valid	Valid with a warning (extra column `glu_suero` ignored)
Missing required field	No `trig`	Rejected	Rejected: missing field "trig"
Missing optional field	No weight	Valid with a warning	Valid with a warning: imputed with the median
Unknown category	sexo = "Otro" (other)	Rejected	Rejected: unrecognized category
Wrong type	edad = "cuarenta" (forty)	Rejected	Rejected: wrong type in "edad"

10.3 Simulation with new patients

Profile	P(high risk)	Classification	Suggested action
Low risk	0.061	No high priority	Regular follow-up and preventive advice
Medium risk	0.277	High priority	Review at the appointment; staff decide whether to move the test forward
High risk	0.697	High priority	Prioritize a confirmatory test and counseling
Uncommon combination (man, 60+, BMI < 25)	0.370	High priority	Estimate less supported by data: full clinical review
Real case close to the threshold	0.260	High priority	Borderline case: healthcare staff decide

10.4 Monitoring plan

Indicator	Frequency	Alert	Action
Missing values per field	Weekly	+10 points above what was seen in training	Review data capture
Records rejected by validation	Weekly	More than 5%	Review formats and units
Data drift (PSI) in age, BMI, triglycerides, HDL and insulin	Monthly	PSI > 0.10 warning; > 0.25 alert	Analyze the cause; recalibrate or retrain
High-risk recall	Monthly	< 0.75 warning; < 0.70 in two cuts: suspend	Adjust the threshold or retrain
Recall gap between groups	Quarterly	> 0.10 warning; > 0.15 suspend	Bias analysis
Response time	Continuous	More than 2 s per batch of 1,000	Review infrastructure

The PSI calculation was simulated: between development and test the population is stable (age PSI = 0.021). With a simulated batch with deliberate drift (triglycerides 30% higher, 35% missing weight and height, and an older population), the age PSI rises to 0.827 and correctly triggers the alert.

11. Reproducibility

Steps to reproduce all results:

1. Clone the repository github.com/AlanDelAngel/diabetes_pacientes_mexicanos.
2. Download the dataset from Kaggle and save `Diabetes_Mexico.csv` next to the notebook or in `data/`.
3. Install the dependencies with `pip install -r requirements.txt`.
4. Run `Proyecto_Final_Riesgo_Diabetes.ipynb` from start to finish (it also works in Google Colab).

Generated folder	Contents
`scripts/`	Custom pipeline transformer and inference script
`artifacts/`	Serialized pipeline, classifier, threshold, encoding and metadata
`results/`	Experiment log, model comparison, test evaluation, bias, errors and figures
`docs/`	Model card in Markdown

Item	Detail
Dataset version	`Diabetes_Mexico.csv`, 2,549 records x 24 columns, md5 `eb696bc3e9a018f03290f6265e7cabfe`
Split	StratifiedGroupKFold by household, seed 42: development 2,184 / test 365
Experiments	19 configurations logged with date, dataset hash, pipeline, hyperparameters and metrics
Libraries	Python 3.11, scikit-learn 1.8.0, pandas 3.0.2, NumPy 2.4.4, Joblib 1.5.3, SHAP 0.51.0
Training date	September 30, 2026

12. Model card

Field	Content
Name and version	Diabetes risk prioritization model, v1.0.0 (tuned RBF SVM, reduced model)
Purpose	Rank which adult patients should get a confirmatory diabetes test first, using demographic and anthropometric data and a biochemical profile without glucose or HbA1c
Intended users	Epidemiology and health planning staff, primary care physicians and preventive program coordinators
Dataset	2,549 adults from 2,055 households; development 2,184 and test 365
Hyperparameters	RBF kernel, C = 3, gamma = 0.03; 15 input features
Metrics	Cross-validation: ROC-AUC 0.764 +/- 0.027. Test: accuracy 0.649, precision 0.481, recall 0.825, F1 0.607, ROC-AUC 0.765, PR-AUC 0.588
Threshold	0.26 on the probability of high risk, chosen for recall >= 0.80
Limitations	Moderate performance (ROC-AUC around 0.76); does not detect the moderate class; requires a lipid and insulin profile; 28% without weight or height; data from a single survey; lower recall for ages 20-39
Bias analysis	Test recall: men 0.83, women 0.82; ages 20-39 0.47; ages 40-59 0.77; ages 60+ 0.98
Monitoring	Weekly data quality; monthly PSI and share of flagged patients; quarterly recall, FPR, FNR and fairness
Updates	Recall < 0.75 in two cuts, PSI > 0.25 or a recall gap between groups > 0.15; annual review
Owners	Alan Adelfo Del Angel Figueroa (technical) and epidemiology coordination (clinical)

Lessons learned

This project let me go through the full lifecycle of a model applied to a real public health problem: from finding errors in the data itself (the meaning of the target variable, a sentinel value slipping into the

model) to defining how it should be monitored once in use.

What made me reflect the most was the bias analysis. The model is even by sex, but much weaker for young adults, simply because high risk is rare in that group and there is less signal to learn from. A global metric like a 0.76 ROC-AUC can hide important weaknesses in subgroups, and reporting them transparently is a core part of building a responsible model. I also learned that choosing the threshold with an explicit clinical criterion is as important a decision as choosing the model, and that a model like this should never be used to deny care or to diagnose, but to support a healthcare professional's decision.

References

American Diabetes Association. (2023). Classification and diagnosis of diabetes: Standards of Care in Diabetes 2023. *Diabetes Care, 46*(Supplement 1), S19-S40. https://doi.org/10.2337/dc23-S002

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Félix Jiménez, A. F., & Sánchez Lee, V. S. (2026). *Diabetes Mexico Data Set* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/14619160

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.