# Cardiovascular Events Prediction on a Synthetic Cohort in Low-Income Setting with Machine Learning

MSR10

Carrasquilla Sotomayor M[1], Chiavegatto Filho ADP[1], Salcedo Mejía F[2], Alvis Zakzuk NJ[3]

1. Universidade de São Paulo, São Paulo, Brazil, 2. ALZAK, Cartagena, Bolívar, Colombia, 3. Universidad de la Costa, Barranquilla, Atlántico, Colombia

## BACKGROUND

Globally, Cardiovascular Diseases (CVD) remain the leading cause of morbidity and mortality. Current clinical guidelines for primary prevention of CVD emphasize the need to identify asymptomatic patients who may benefit from preventive actions based on predicted risk. Machine learning (ML) model, can incorporate a diverse array of features, leading to more accurate risk predictions tailored to individual patients. The study's objective was to identify the optimal ML model for the prediction of cardiovascular events (CVE) in patients enrolled in a cardiovascular program during 2013-2018.

## METHOD

We performed two ML predictive models: 1) With the first fatal or non-fatal CVE as the primary outcome, and 2) For subsequent events. Demographic, clinical, anthropometric, and epidemiological covariates were included as predictors. The models were trained using synthetic data for 20.000 patients simulated from an original data set of 93,552 patients enrolled in a cardiovascular cohort program from Colombia. Data were simulated to replace sensitive values and causing minimal distortion of the statistical distribution.

We trained four ML algorithms for structured data on 70% of the sample: Random Forest (RF), XGboost (XGB), LightGBM (LGBM) and Catboost; then were tested on the remaining 30%. Hyperparameters were selected using Random-search and variable selection was optimized with Boruta. For model selection we identified the highest AUC-ROC, accuracy and recall.

**Graph 1. Main metrics of tested algorithms**

| Algorithm | Accuracy | Recall | AUC-ROC |
|---|---|---|---|
| **First CVE model** | | | |
| Random Forest | 0.94 | 0.01 | 0.79 |
| XGBoost | 0.94 | 0.09 | 0.81 |
| LightGBM | 0.94 | 0.08 | 0.80 |
| Catboost | 0.94 | 0.09 | **0.82** |
| **Subsequent CVE model** | | | |
| Random Forest | 0.74 | 0.29 | 0.72 |
| XGBoost | 0.76 | 0.39 | **0.74** |
| LightGBM | 0.76 | 0.43 | 0.68 |
| Catboost | 0.75 | 0.39 | 0.70 |

## RESULTS

For the first CVE model the AUC-ROC metrics were: 0.79 for RF, 0.81 for XGB, 0.80 for LGBM, and the highest performance was for Catboost (0.829). However, the model reported a recall of 0.0912 when using the unbalanced outcome sample (<5.8% occurrence). From 555 patients with CVE history, the second model (30.63% subsequent CVE occurrence) obtained a lower performance, with little predictive potential for subsequent CVE risk. Catboost algorithm showed the highest performance for both evaluated outcomes (0.695). Nevertheless, for subsequent CVE XGB and RF showed an improved AUC-ROC (>0.72) and recall (>0.39). All models identified as strong predictors the time to event, high risk categorization, age, microalbuminuria and creatinine, cholesterol, and glycemic levels.

## CONCLUSIONS

ML models offer a few advantages especially when dealing with large datasets and unbalanced events, maintaining high performance and reliability in their predictions. Our findings provide an additional tool to help decision-making on prevention routes in primary and secondary care to enhance patients' quality of life.

**Graph 2. Shapley for optimized algorithm***