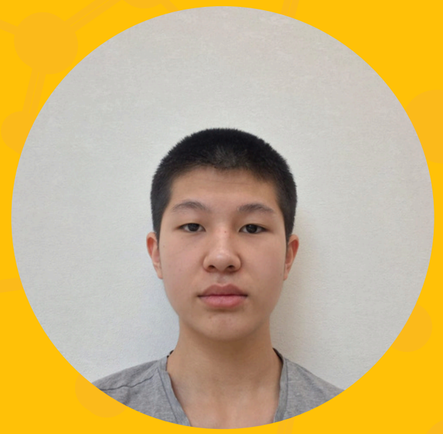
MACHINE LEARNING-BASED LDL CHOLESTEROL PREDICTION

Team: Hope of Kazakhstan

OUR TEAM



BALTABAYEV BEK

BIOLOGIST

KENES SANZHAR

TEAM LEADER

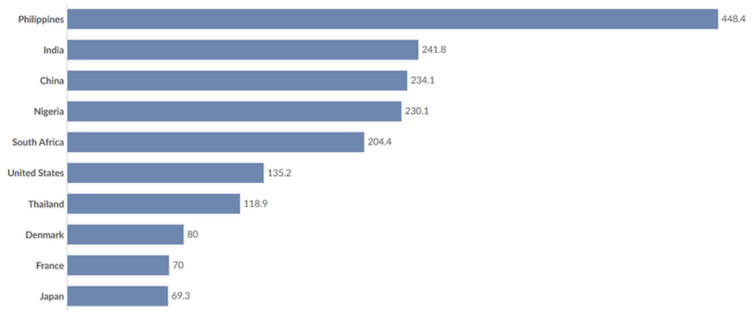
FRONTENDER

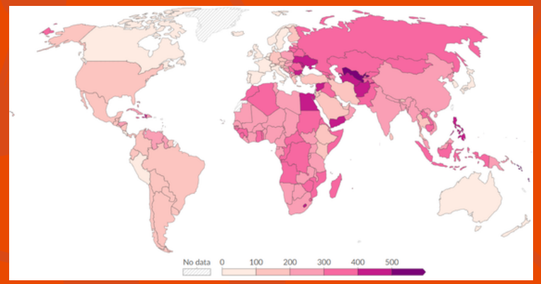
TYNYSHTYK BEKTAI

ML DEVELOPER

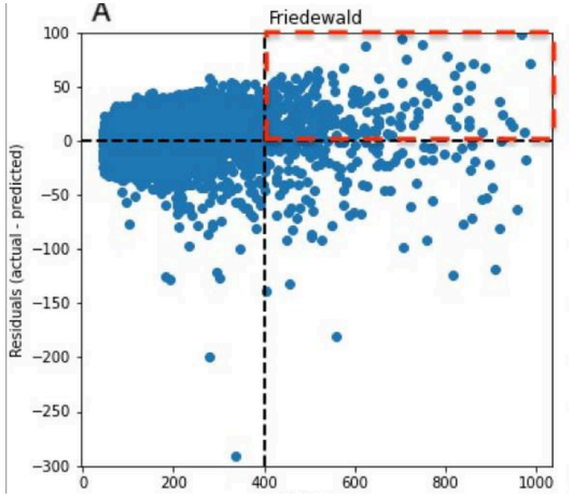
PROBLEMS

LDL-C IS CAUSING DEATHS WORLDWIDE THROUGH CVD



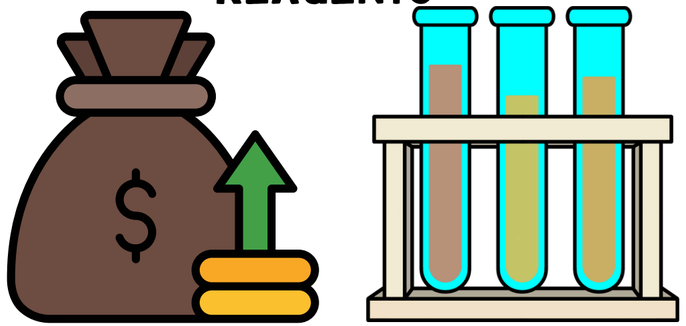


INACCURATE CALCULATION FORMULAS



STANDARD EQUATIONS FAIL AT TRIGLYCERIDE LEVELS > 400 MG/DL AND IGNORE COMPLEX BIOLOGICAL INTERACTIONS.

DIRECT LDL-C MEASUREMENT IS EXPENSIVE AND REQUIRES SPECIALIZED FACILITIES AND REAGENTS



PROHIBITIVE LABORATORY COSTS

PROBLEM: STANDARD EQUATIONS

$$LDL\text{-}C = TC - HDL\text{-}C - \frac{TG}{5}$$

Friedewald formula

$$LDL\text{-}C = TC - HDL\text{-}C - \frac{TG}{\zeta}$$

Martin formula

$$LDL\text{-}C = \frac{TC}{0.948} - \frac{HDL\text{-}C}{0.971} - \frac{TG}{8.56} + \frac{TG \times non\text{-}HDL\text{-}C}{2140} - 9.44$$

Sampson formula

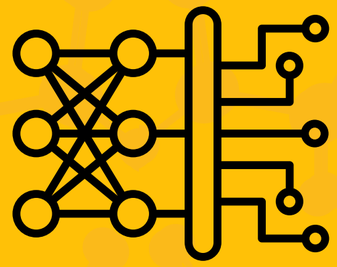
All three formulas – Friedewald, Martin, and Sampson – lose accuracy when triglyceride (TG) levels exceed 400 mg/dL, particularly in cases of hypertriglyceridemia or abnormal lipid profiles. Their ability to handle complex nonlinear relationships between lipid parameters is limited, because they rely on predefined mathematical assumptions rather than adaptive learning.

GOAL:

TO DEVELOP AND VALIDATE RANDOM FOREST REGRESSION THAT PROVIDE MORE ACCURATE LDL-C ESTIMATES THAN TRADITIONAL FORMULAS

TASKS

COLLECT AND PREPROCESS A DATASET OF LIPID PROFILES (TC, HDL-C, TG)



IMPLEMENT AND TRAIN RANDOM FOREST REGRESSION

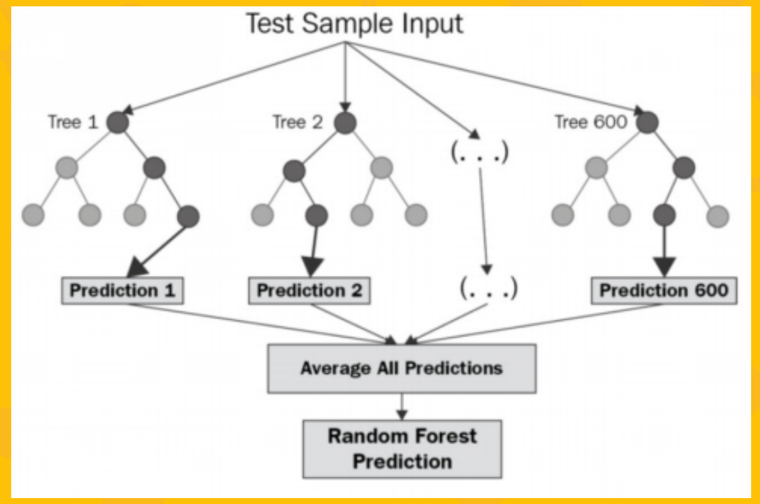
EVALUATE PERFORMANCE AGAINST DIRECTLY MEASURED LDL-C AND TRADITIONAL EQUATIONS

SOLUTION

THE SOLUTION - MACHINE LEARNING APPROACH

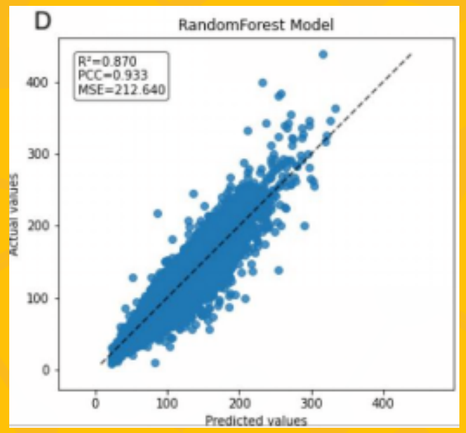
Advanced Machine Learning

USING RANDOM FOREST TO OUTPERFORM RIGID TRADITIONAL FORMULAS



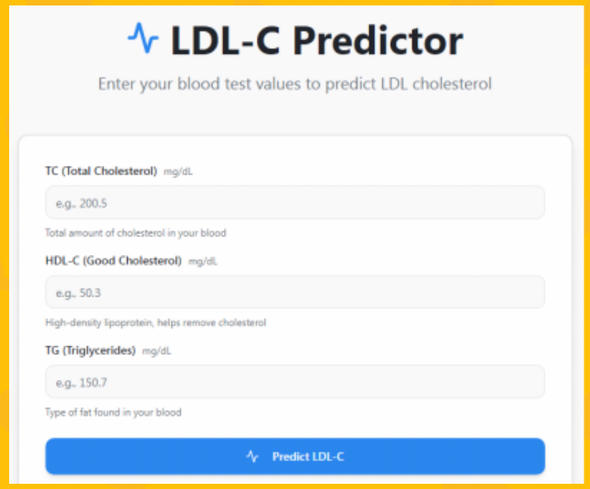
Dynamic Pattern Capture

CAPTURING NON-LINEAR LIPID PATTERNS FOR UNIVERSAL PRECISION

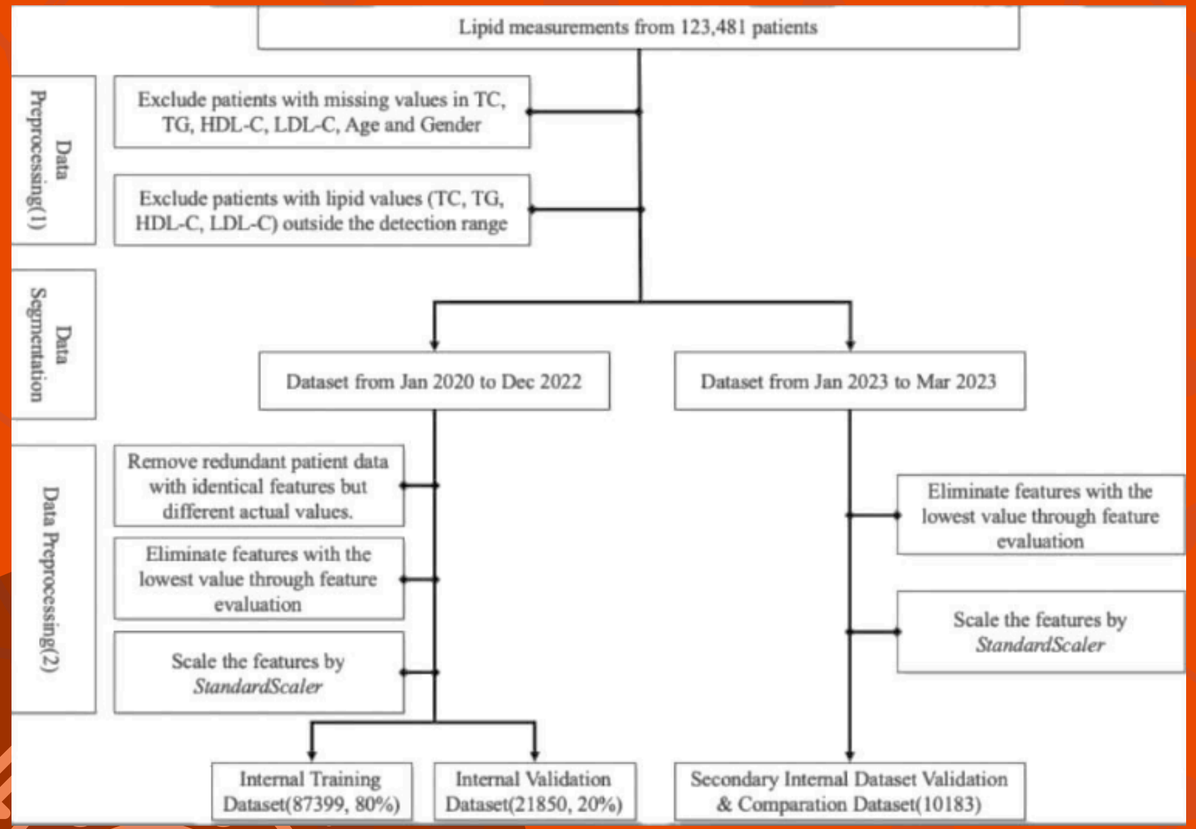


Accessible Clinical Tool

DEVELOPING A WEB APPLICATION (LIPIDAI.VERCEL.APP) FOR REAL-TIME



THE DATASET



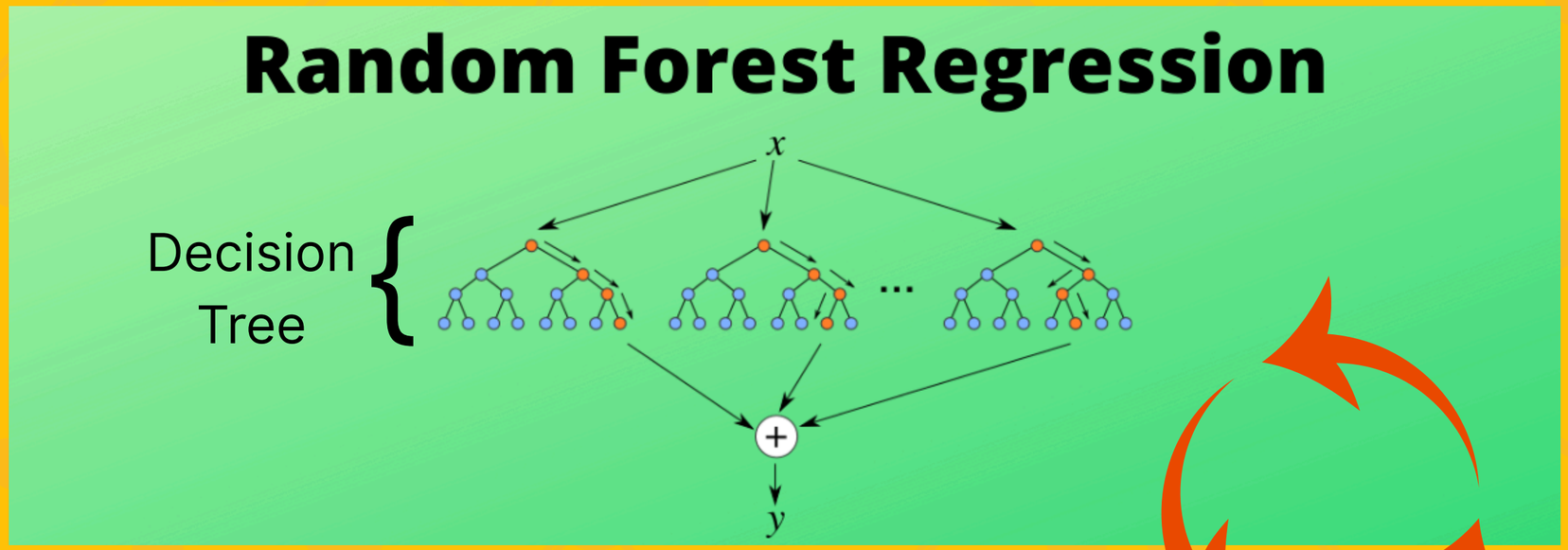
THE DATASET CONSISTS OF LIPID PROFILE MEASUREMENTS FROM REAL PATIENTS, INCLUDING:

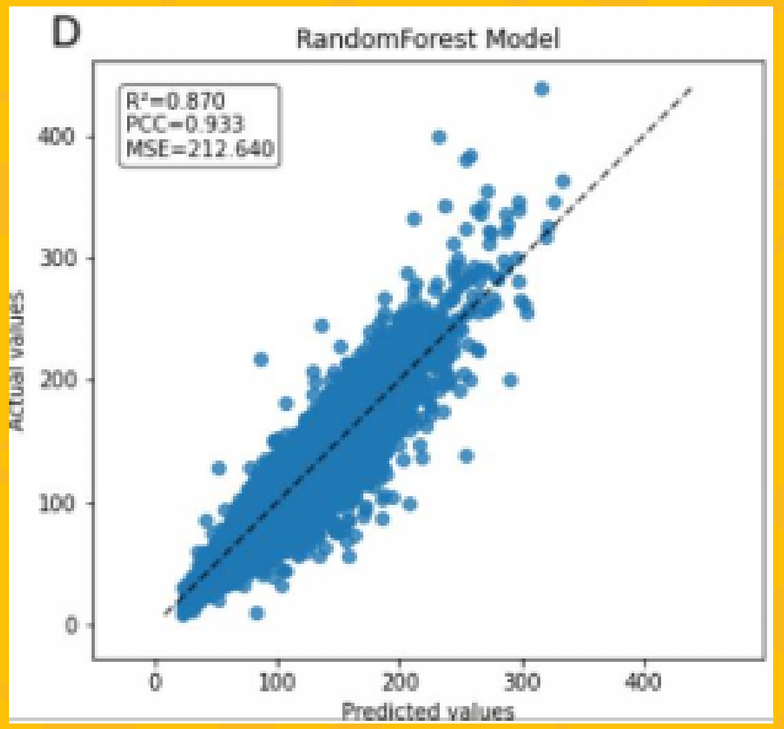
1. TOTAL CHOLESTEROL (TC)
2. HDL-C, TRIGLYCERIDES (TG)
3. TARGET VARIABLE LDL-C

THE NEXT PROCESSES:

1. DATA CLEANING
2. REMOVAL OF MISSING VALUES
3. MODEL TRAINING BY DATASET
4. MODEL EVALUATING BY DATASET WITH 80/20 TRAIN–TEST SPLIT, ENSURING RELIABLE PERFORMANCE ASSESSMENT AND ROBUST PREDICTION OF LDL-C LEVELS.

RANDOM FOREST REGRESSION MODEL





RANDOM FOREST REGRESSION PREDICTS A VALUE BY COMBINING THE RESULTS OF MANY DECISION TREES. EACH TREE IS TRAINED ON A RANDOM SUBSET OF THE DATA AND A RANDOM SUBSET OF FEATURES. WHEN MAKING A PREDICTION, EVERY TREE PRODUCES ITS OWN OUTPUT, AND THE FINAL RESULT IS CALCULATED AS THE AVERAGE OF ALL TREE PREDICTIONS.

THIS ENSEMBLE APPROACH IMPROVES ACCURACY AND STABILITY COMPARED TO A SINGLE DECISION TREE.

TECHNOLOGY

```
1. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
   print("Начинаю обучение на 100 000 строк...")
2. model = RandomForestRegressor(
       n_estimators=500,
       max_depth=15,
       min_samples_leaf=5,
       n_jobs=-1,
       random_state=42
   )

3. model.fit(X_train, y_train)
   print("Обучение завершено!")
4. y_pred = model.predict(X_test)
5. mae = mean_absolute_error(y_test, y_pred)
   r2 = r2_score(y_test, y_pred)
```

1. TRAIN–TEST SPLIT
2. RANDOM FOREST MODEL INITIALIZATION
3. MODEL TRAINING
4. PREDICTION ON TEST DATA
5. MODEL EVALUATION

FEATURES

DATA-DRIVEN

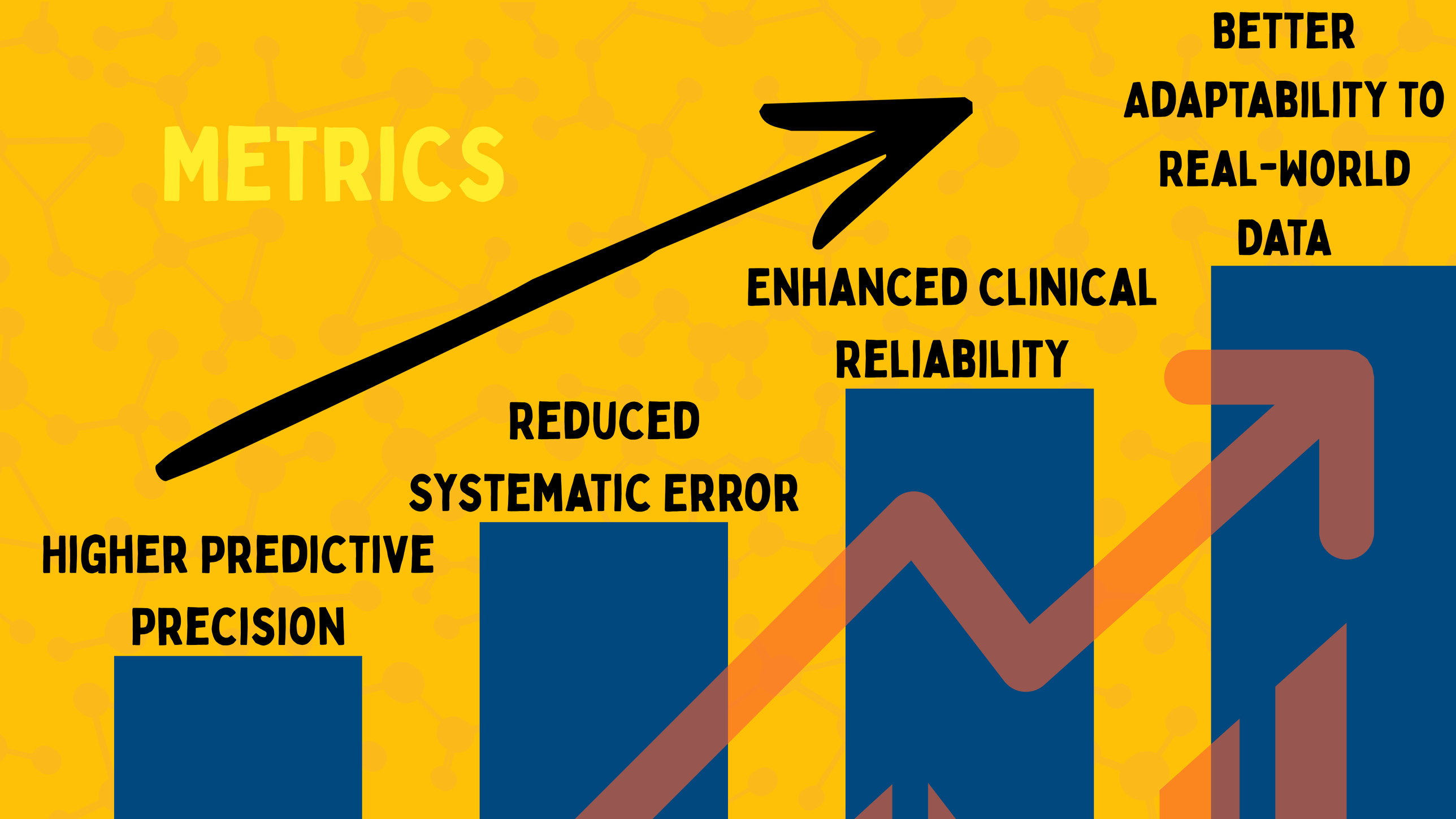
IT LEARNS PATTERNS DIRECTLY FROM REAL CLINICAL DATA.

STABLE AT HIGH TRIGLYCERIDE LEVELS

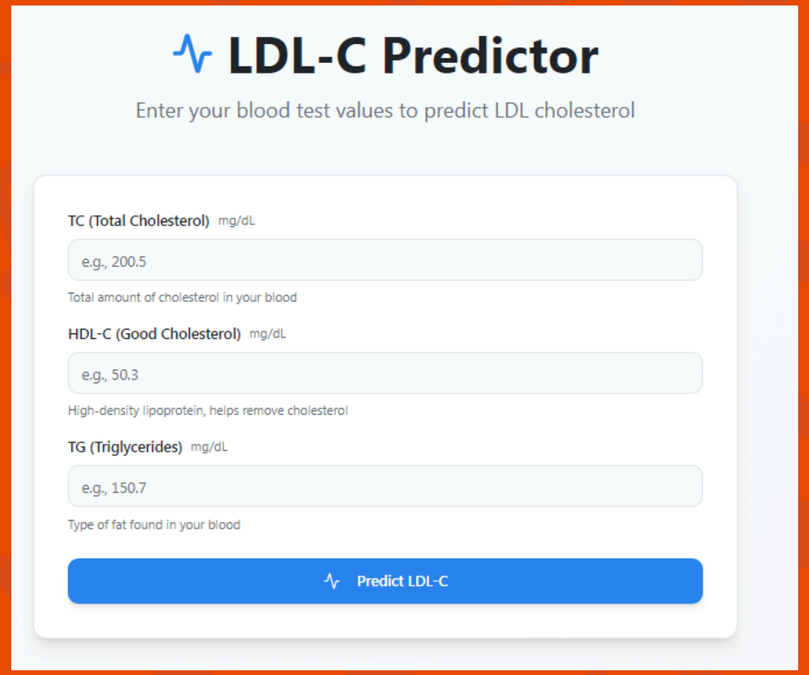
RANDOM FOREST MAINTAINS CONSISTENT PERFORMANCE ACROSS WIDER BIOCHEMICAL RANGES.

STABLE AT HIGH TRIGLYCERIDE LEVELS

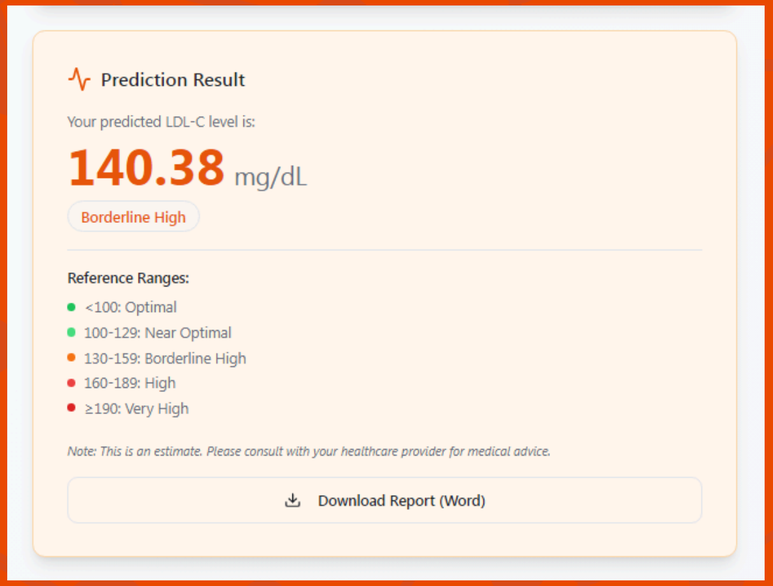
MACHINE LEARNING OPTIMIZATION ALLOWS IT TO ACHIEVE STRONGER AGREEMENT WITH LABORATORY LDL MEASUREMENTS


METRICS
HIGHER PREDICTIVE PRECISION
REDUCED SYSTEMATIC ERROR
ENHANCED CLINICAL RELIABILITY
BETTER ADAPTABILITY TO REAL-WORLD DATA

WEB-SITE



The website is built with React, hosted on Vercel



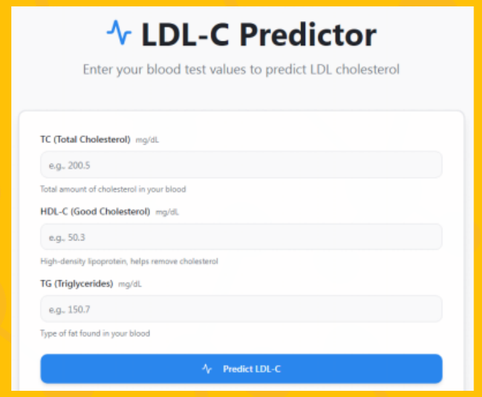
1. USER ENTERS THE BLOOD TEST VALUES

2. THE SITE CALCULATES PREDICTED LDL-C LEVEL

COMPARE

Fixed formulas and our model			
Method	**R2**	**MSE**	**PCC**
Friedewald formula	0.72	290.11	290.11
Martin formula	0.75	310.62	310.62
Sampson formula	0.77	270.52	270.52
Random Forest Regression Model	0.87	212.267	212.267

THE SCHEME OF WORK



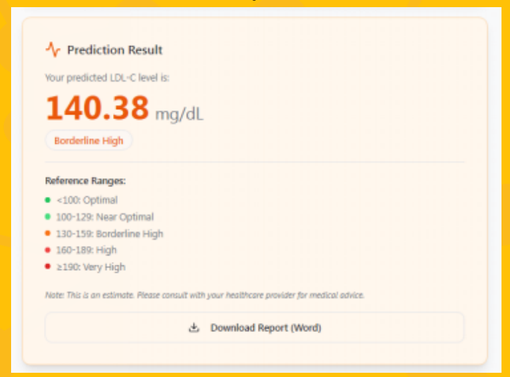
THE USER ENTERS VALUES ON THE WEBSITE: LIPIDAI.VERCEL.COM

FastAPI

THE SERVER RECEIVES THE VALUES AND ANALYZES THEM USING ML

MACHINE LEARNING

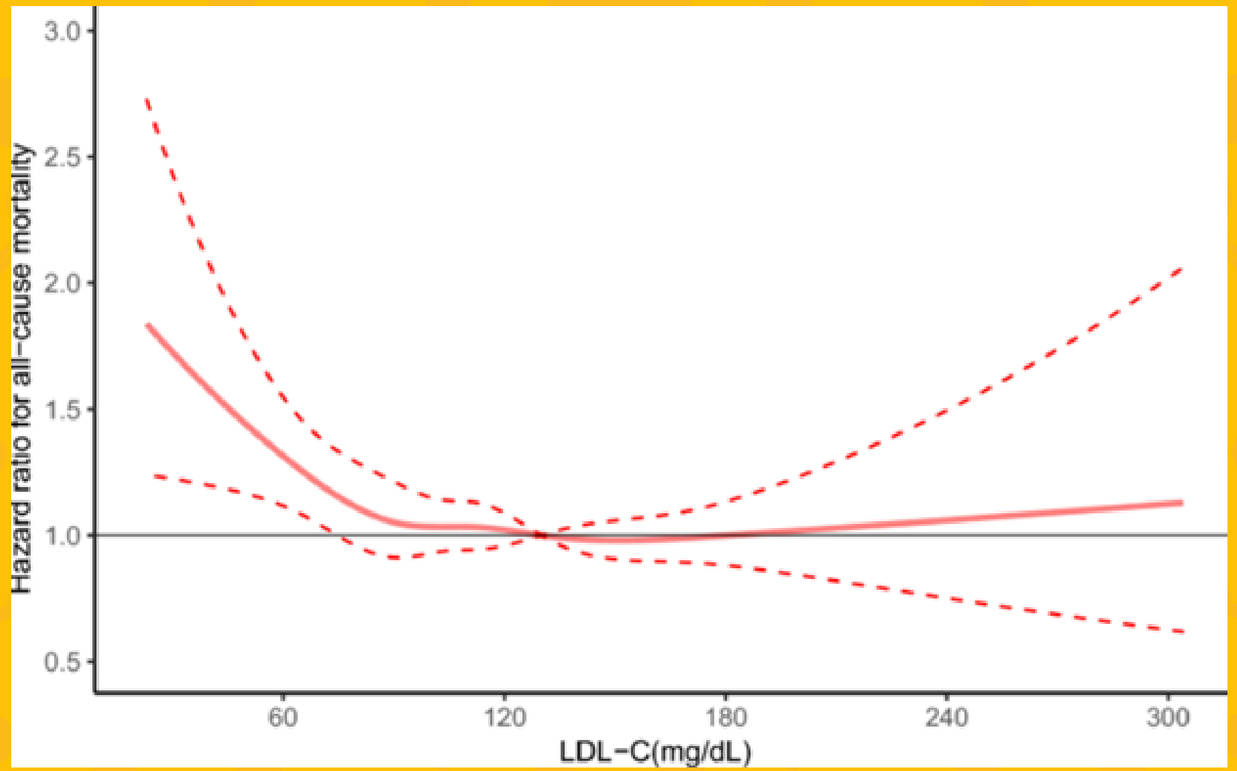
PREDICTION OF RANDOM FOREST REGRESSION MODEL



THE PREDICTED LDL-C VALUE IS DISPLAYED ON THE WEBSITE

RESULT

RANDOM FOREST REGRESSION MODEL IS NEEDED FOR ACCURATE LDL-C ESTIMATION. ADOPTING IT INTO MEDICINE WIIL BE A PLEASURE TO THOSE WHO SUFFER FROM HARMFUL CHOLESTEROL



FINAL

BETTER DATA - BETTER DECISIONS - BETTER PATIENT OUTCOMES

HOPE OF KAZAKHSTAN

OUR CONTACTS

+7 707 902 0610
+7 775 883 7493
+7 771 520 0603

sanzhar_k0602@akb.nis.edu.kz
tynyshtyk_b1017@akb.nis.edu.kz
baltabaev_b1114@akb.nis.edu.kz