

REAL WORLD ANALYTICS DS3104



PROJECT REPORT

Health Analytics: A Machine Learning Approach for Risk Prediction

22CDS0439

Harol Maxilan

Academic Advisor

Professor S. Vasanthapriyan

Dean, Faculty of Computing

Sabaragamuwa University of Sri Lanka

Department of Data Science, Faculty of Computing Sabaragamuwa

University of Sri Lanka

2025 November 30

Contents

1. Project Overview	3
2. Data Preparation and Cleaning.....	3
3. Exploratory Data Analysis (EDA).....	4
4. Preprocessing Pipeline.....	5
5. Modeling and Evaluation	6
6. Key Interpretability Insights	7
7. Conclusion	8
8. Reference	8
9. Appendix.....	9

1. Project Overview

This project applies a complete machine learning workflow to analyze patient health indicators and predict a health risk score.

A synthetic medical dataset containing 20,000 patient records was used to train multiple regression models and interpret the factors contributing to high health risks.

The pipeline includes:

- Data cleaning
- Feature engineering
- Risk score formulation
- Model training (Linear, Ridge, Lasso, Decision Tree)
- Evaluation via error metrics
- SHAP-based interpretability
- Insights on disease indicators

The final model accurately identifies the most influential health factors, helping support preventive healthcare analytics.

2. Data Preparation and Cleaning

The dataset contained 20,000 records with attributes such as age, sugar percentage, glucose levels, cholesterol, blood pressure, and diabetic conditions.

Key Preprocessing Steps

- Blood Pressure Processing:
 - The blood_pressure attribute (e.g., “120/80”) was split into systolic_bp and diastolic_bp.
- Boolean Conversion:
 - Fields like has_eye_disease and has_diabetic_retinopathy were converted into binary (0/1) numeric values.
- Duplicate Analysis:
 - The dataset contained 0 duplicate records, ensuring high data quality.

Missing Values:

- No missing values were detected in the dataset.

Basic Statistical Summary

The variables displayed realistic medical ranges:

- Mean glucose: 135 mg/dL
- Mean cholesterol: 199 mg/dL
- Mean obesity percentage (BMI-derived): 27.49%
- Systolic BP between 100–160 mmHg
- This confirmed the dataset is comprehensive and medically relevant.

3. Exploratory Data Analysis (EDA)

The EDA phase revealed important relationships:

- Strong Correlation Patterns
- Glucose %, Sugar %, Obesity %, and Blood Pressure contributed significantly to risk levels.
- Patients with diabetic retinopathy showed sharply elevated medical indicators.
- Heart rate distributions varied predictably across age groups.
- Categorical Insights
- Age was grouped into:
 - Young (<40)
 - Middle-aged (40–60)
 - Senior (>60)
- BMI categories were derived from obesity percentage:
 - Underweight, Normal, Overweight, Obese

4. Preprocessing Pipeline

A structured preprocessing workflow ensured model readiness:

Feature Scaling

- For regression models (Linear, Ridge, Lasso), numerical features were standardized using StandardScaler.
- Decision Tree operated on raw values (not scaled) due to its structure.

Dataset Splitting

- Train–test split: 80% training, 20% testing

Target Variable Creation

A composite risk_score was calculated using weighted contributions from:

- Sugar percentage
- Glucose level
- Cholesterol
- Obesity
- Blood pressure
- Eye disease
- Diabetic retinopathy

5. Modeling and Evaluation

Four regression models were trained and evaluated:

Model	MSE	MAE	R ²
Linear Regression	0.0000	0.0000	1.0000
Ridge Regression	0.0000	0.0005	1.0000
Lasso Regression	0.0531	0.1846	0.9997
Decision Tree Regressor	0.0820	0.2304	0.9995

Key Observations

- Linear & Ridge Regression performed perfectly ($R^2 = 1.0000$), indicating extremely strong linear relationships in data.
- Lasso Regression performed slightly worse due to coefficient shrinkage.
- Decision Tree Regressor still performed excellently but was slightly less stable.

Conclusion

The Linear Regression and Ridge Regression models are the best-performing models, offering perfect predictive accuracy.

6. Key Interpretability Insights

To understand why the model predicts certain risk levels, interpretability tools were applied.

Decision Tree Feature Importance

Top contributing features:

- Diabetic Retinopathy — 0.8767
- Eye Disease — 0.1103
- Sugar Percentage — 0.0062
- Systolic BP — 0.0034
- Obesity % — 0.0033

Medical interpretation:

- Diabetic retinopathy and general eye disease dominated the risk score, reflecting real-world clinical importance.
- Metabolic indicators also contributed but at a lower magnitude.

SHAP Analysis

- Global Explanation (Summary plot):
 - Retinopathy and eye disease had the largest absolute SHAP values, confirming global importance.
- Local Explanation (Waterfall Plot):
 - For individual patients, increases in glucose, obesity, or systolic BP directly increased risk.
 - SHAP precisely showed how much each factor contributed to a single prediction.

SHAP validated the clinical reasoning embedded in the engineered risk score.

7. Conclusion

This project successfully built a comprehensive machine learning pipeline for predicting patient health risk scores.

Highlights:

- Data was clean, consistent, and medically interpretable.
- Multiple regression models achieved excellent accuracy, with Linear and Ridge Regression producing perfect fit.
- SHAP and feature importance analysis identified critical medical indicators:
 - Diabetic retinopathy
 - Eye disease
 - Sugar percentage
 - Blood pressure

Impact

These insights can help health professionals prioritize patients with severe indicators and build early-warning systems for chronic medical conditions.

8. Reference

- Scikit-learn Documentation — <https://scikit-learn.org>
- Pandas Documentation — <https://pandas.pydata.org>
- Seaborn Data Visualization — <https://seaborn.pydata.org>
- SHAP Library — <https://shap.readthedocs.io>
- Matplotlib Documentation — <https://matplotlib.org>

9. Appendix

1. <https://www.kaggle.com/datasets/jockeroika/eye-health>