ML approaches to improve patient outcomes for Heart Disease and Diabetes diagnoses

Rachel Roggenkemper, Jacob Perez, Brendan Callender
Cal Poly Department of Statistics in collaboration with the American College of Cardiology

AMERICAN COLLEGE of CARDIOLOGY®

Background

Diabetes and Cardiovascular Disease (CVD) are closely linked, requiring integrated approaches for risk assessment. These conditions significantly impact global health outcomes.

Project Goals

1. Develop **predictive algorithms** that improve **diagnostic consistency** for these disease states.
2. Prioritize **equitable outcomes for male** and **female patients** to improve patient outcomes across sex.

Project Data

1. CDC **Diabetes** Health Indicators Dataset
2. **CVD** Data from a Multispecialty hospital in India
3. Sylhet **Diabetes** Hospital in Bangladesh dataset

Project Data

CDC Diabetes Dataset (N = 70,692):

- Classification Target: Diabetes vs No Diabetes
  - 50/50 split in data (positive/negative)
- **Demographic** and **Lifestyle** predictors
  - Easily accessible, minimal testing

High BP?	High Chol?	BMI	Sex	...	Age Group	Difficulty Walking?	Diabetes?
Yes	Yes	33	Male	...	55-59	Yes	Yes
No	Yes	24	Female	...	18-24	No	No

Cardiovascular Disease Dataset (N = 1,000):

- Classification Target: Heart Disease vs No Heart Disease
  - 58/42 split in data (positive/negative)
- **Demographic**, **Clinical**, **Biochemical**, and **Lifestyle** predictors
  - Patient testing required

Age	Sex	Chest Pain	Resting BP	...	Peak Exercise Slope	# Major Vessels	Heart Disease?
53	Male	Non-Anginal	171	...	Downsloping	3	Yes
40	Male	Typical Angina	94	...	Upsloping	1	No

Early-Stage (ES) **Diabetes Dataset** (N = 520):

- Classification Target: Diabetes vs No Diabetes
  - 60/40 split in data (positive/negative)
- **Demographic**, **Symptom-Based** predictors
  - Minimal testing required

Excessive Thirst?	Excessive Urination?	Sex	...	Age	Excessively Hungry?	Vision Blurring?	Diabetes?
Yes	Yes	Male	...	51	Yes	No	Yes
No	Yes	Female	...	43	No	Yes	No

CDC Diabetes Decision Tree Classifier

Only requires **4 easy-to-collect predictors**:

1. Whether patient has **high blood pressure**
2. Patient **BMI**
3. Whether patient has **difficulty walking or climbing stairs**
4. Whether the patient would describe their **current health** as "very good"

Correctly diagnoses
70%
69%/71%
of patients

Correctly diagnoses
75%
71%/77%
of patients who truly have diabetes

ROC-AUC
0.74
0.71/0.75

CVD Logistic Regression with Elastic Net

Most Important Predictors for:

Positive Heart Disease Diagnosis	Negative Heart Disease Diagnosis
ST Depression of **EKG**	Normal ST Slope on **EKG**
Presence of **Chest Pain**	Normal Resting Blood Pressure

Correctly diagnoses
96%
96%/95%
of patients

Correctly diagnoses
95%
94%/96%
of patients who truly have diabetes

ROC-AUC
0.99
0.99/0.98

ES Diabetes Logistic Regression with LASSO

Most Important Predictors for:

Positive Diabetes Diagnosis	Negative Diabetes Diagnosis
Having **Excessive Itching**	Having **Excessive thirst**
Having **Muscle Stiffness**	Having **Excessive urination**

Correctly diagnoses
90%
89%/92%
of patients

Correctly diagnoses
91%
81%/99%
of patients who truly have diabetes

ROC-AUC
0.94
0.95/0.91

Methods

1. **Exploratory Data Analysis**
   - Examined distribution of sex and diagnosis in data
   - Investigated predictor relationships with diagnoses
2. **Classification Models** (Supervised Learning)
   - Decision Tree Classifiers
   - Logistic Regression with Ridge/LASSO penalties
3. **Evaluation Metrics Used**
   - **Accuracy** – Overall correctness of model diagnosis predictions
   - **Sensitivity** – Correctness of model diagnosis for those who truly have a positive diagnosis
   - **ROC-AUC** – Measures model's ability of balancing the true positive rate and false positive rate. We expect a value of 0.5 for random guessing and 1 for a perfect model.

Limitations

External Validity of Results:

Due to **cultural differences** which influence individuals' **diet**, **health habits**, **perceptions of pain**, and **medical symptoms**, we advise only applying these models for the following populations:

- CDC Diabetes Model -- American adults
- CVD Model -- Indian adults
- ES Diabetes Model -- Indian adults

We also recognize these **data represent** individuals who do have **access to health care** and may **underrepresent marginalized groups** who lack access to health care.

Negative Model Impact:

- **False negatives** could lead to diseases being left **untreated**
  - This can potentially affect patients with atypical symptoms

Lastly, **FDA approval** and **additional model testing** is **required** before these models can be freely used by doctors

References

Doppala, Bhanu Prakash; Bhattacharyya, Debnath (2021), "Cardiovascular_Disease_Dataset", Mendeley Data, V1, doi: 10.17632/dzz48mvjht.1

Centers for Disease Control and Prevention (CDC), Behavioral Risk Factor Surveillance System Survey Data, [year of data], Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention

Early Stage Diabetes Risk Prediction [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5VG8H