Woodrow Reese

10/22/2024

Project Proposal: Diabetes Prediction

Diabetes is a disease that affects the body's ability to process glucose. Early prediction of diabetes risk allows for effective treatment, awareness and management. This project focuses on using machine learning techniques to predict the risk of diabetes for women in the Pima Indian community who are known to have a significantly high rate of diabetes. This dataset consists of 9 health related measurements for 768 women of age 21 and older. The columns are: Age, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree and Outcome.

A concern of this dataset is the missing/zero values in health indicators (5 glucose, 35 blood pressure, 227 skin thickness, 374 insulin, 11 BMI), meaning some entries were tested less for predictors than others. These incomplete entries require additional handling of while preprocessing the data to optimize model performance.

By analyzing this dataset, we have the goal of answering the following questions:

1. What is the most important factor in predicting whether someone is diabetic?
    a. This is a Supervised ML task that can be answered with a classification model such as Logistic Regression
2. How can we assess each entry and place them into risk groups (LOW, MED, HIGH) based on health indicators?
    a. This is an Unsupervised ML task that can be answered with a clustering model such as KNN
3. How are age and pregnancies related to diabetes risk?
    a. This is a Supervised ML task that can be answered with a regression model. One approach can be a logistic regression model to predict whether the person has diabetes or not. Another approach would be to use linear regression to give a score such as a pedigree to predict the risk of diabetes.
4. What is the relationship between BMI and insulin when predicting diabetes risk?
    a. This is a Supervised ML task that can be answered with a regression model. Specifically, I'd like to know how the outcome is relevant when both values are high vs low.
5. How does the performance of different predictive models compare: logistic regression vs decision trie vs neural network.
    a. This is a Supervised ML task that can be answered by classifying the model's performance. A measurement of performance can be based on a series of precision, accuracy and recall.

Link - Diabetes Dataset (kaggle.com)