Project MileStone - Machine Learning for Diabetes Prediction

Team Details:

S. No	Name	SJSU ID	Email
1.	Vishnu Vardhan Reddy Ireddy	016816176	vishnuvardhanreddy.ireddy@sjsu.edu
2.	Vani Vineela Aremanda	016131284	vanivineela.aremanda@sjsu.edu
3.	Lakshmi Sri Nitya Sunkara	017459390	lakshmisrinitya.sunkara@sjsu.edu

Project Github Link :

https://github.com/VardhanReddy2/data-mining-course-project

Objective:

Our goal is simple, we want to make a tool that can tell if someone has diabetes by looking at their health info. We're focusing on women of Pima Indian heritage who are 21 and older because they have a higher chance of getting diabetes. We're using machine learning to find the most important signs of diabetes and to see how well different methods work in finding it early. This study is important for predicting the disease early and helping people stay healthy.

Dataset:

We're working with data from the National Institute of Diabetes and Digestive and Kidney Diseases. It's all about predicting diabetes in Pima Indian women who are over 21. The data tells us things like their glucose levels, blood pressure, body mass index (BMI), and more. There are 768 people in the data, each with 8 health stats plus an outcome that says if they have diabetes or not.

Downloaded from: https://www.kaggle.com/code/ohseokkim/diabetes-three-ensemble-models/input

Baseline Modules:

These are the baseline modules we’ve used so far:

- **Cleaning the data**: We've been fixing up the data so it's ready for analysis. We are dealing with missing info and making sure everything's in a good state to work with.

- **Machine Learning Basics**: We've started with some common methods like Logistic Regression, Random Forest, and a couple more to train our tool and make predictions.

- **Feature Engineering with PCA**: Using PCA, we've been able to figure out which health stats are most important for predicting diabetes.

Methodology:

- **Step 1: Data Collection and Data Fixing**: The initial step involved collecting data from the Pima Indian Diabetes Database on Kaggle and cleaning it up.

- **Step 2: Feature Engineering**: We’ve employed Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, thus improving computational efficiency and model accuracy by focusing on the most significant predictors.

- **Step 3: Training our Tool**: We tried out different ways to predict diabetes, like Logistic Regression and Random Forest.

- **Step 4: Comparative Analysis**: We’ve compared each algorithm based on accuracy, precision, and recall metrics. This comparative analysis helped us to select the most effective model for predicting diabetes.

- **Step 5: Validation and Testing**: We tested our final choice to make sure it gives good results in real life.

Achievements and Current Status:

Our project has made substantial progress in developing a system for predicting diabetes using machine learning techniques. The primary achievements include:

- **Gathered Data**: We’ve sourced data from the Pima Indian Diabetes Database on Kaggle, the dataset includes key health metrics like glucose levels, blood pressure, and BMI, indicative of diabetes presence.

- **Prepared the Data**: We’ve employed Pandas and Numpy for data cleaning and pre-processing to ensure data quality and reliability for model training.

- **Tested Different Models**: We’ve and compared four supervised machine learning algorithms—Logistic Regression, Random Forest, k-Nearest Neighbours (kNN), and Support Vector Machine (SVM)—for predicting diabetes.

- **Feature Engineering**: Applied Principal Component Analysis (PCA) to reduce dimensionality, enhancing model accuracy and computational efficiency.

But, we're still working on it. We've made considerable progress on our project, completing many of the tasks we needed to do. We're planning to try three more Machine Learning Algorithms and we want to analyze these algorithms (XGBoost Classifier, Decision Tree Classifier, Naïve Bayes) to see which one is best for predicting diabetes based on accuracy.

Challenges:

We’ve faced few challenges along the way:

Not all the data was complete, especially for key things like blood pressure. Most of our data was about healthy people, which made it hard to predict diabetes. Figuring out which health stats were really important to predict diabetes was a challenge. With so many prediction methods out there, choosing the right one was tough.

How We Fixed Them:

Here’s how we’ve tackled these issues:

We estimated missing values based on the rest of the data. We added more data for people with diabetes to even things out. PCA helped us focus on the health stats that are more important. We didn't just stop with the first method we tried. We tested a bunch and kept track of how well each one worked.

Future Work:

We're looking into trying a couple more predictive methods (XGBoost Classifier, Decision Tree Classifier, Naïve Bayes) to see if we can improve our accuracy even further. Once we find the best method, we'll fine-tune it and double-check to make sure it's as reliable as possible.

References:

[1] P. Verhulst, "Recherches mathématiques sur la loi d'accroissement de la population," *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, vol. 18, pp. 1–42, 1845.

[2] T. Oliphant, "Python for Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 10–20, May/June 2007, doi:10.1109/MCSE.2007.58.

[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi:10.1023/A:1010933404324.

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi:10.1007/BF00994018.

[5] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: Wiley, 2012.