CS539 Project Proposal
Advanced Predictive Modeling for Healthcare and Stock Market
Team 4

Omnia Abouhassan, Nate Hindman, Tanveer Kaur, Sepideh Sedghi, Yilu Yang, Haoyang An

Problem Definition

The problem that we will be working on is: Evaluating the effectiveness of different ML models (regression vs classification) across diverse prediction tasks.

The function of our tool: The tool will train and evaluate two different approaches to machine learning such as regression and classification on two different types of datasets: diabetes diagnosis and stock price prediction. The six of us will use different machine learning models to evaluate the performance of each of the models on each of the tasks to determine which model is more suitable for each problem type.

The tool will train and compare regression and classification models on two different tasks: predicting whether a patient has diabetes and forecasting stock prices. It will evaluate each model’s performance using appropriate metrics and determine which model is more effective for each type of problem.

Motivation

The healthcare industry struggles with early diagnosis and preventive care, often leading to worse outcomes for patients and higher costs. Early diagnosis in healthcare and accurate forecasting in finance are critical yet challenging tasks. Our tool is designed to help healthcare professionals and financial analysts better understand how to select suitable machine learning models for their domain-specific problems.

Additionally, this project serves as a learning framework for students, educators, and researchers in machine learning, offering hands-on insight into model behavior across problem types. It bridges the gap between theory and practical application, especially in model selection.

Existing approaches (literature)

There are several existing predictive tools in both healthcare and finance:

IBM Watson Health and Google Health uses advanced AI models for disease prediction and patient risk analysis.

In finance, tools like QuantConnect and Yahoo Finance APIs support machine learning-based forecasting.

Our tool is different because it directly compares multiple classic machine learning models (both classification and regression) on two completely different problem domains (healthcare and finance) using transparent, interpretable models like decision trees, logistic regression, and SVM. We’ re applying and evaluating models on both classification (diabetes) and regression (stock prediction) tasks in the same framework. We believe that our tool can be an educational framework and particularly useful for students, researchers, and data scientists who are learning to choose appropriate machine learning methods for various tasks. While most tools don’t offer cross-domain, model-type comparisons in a clear and interpretable way, our tool fills this niche.

The difficulty of our project is medium, we could use Random Forest, SVM, Logistic Regression, etc. as our tool to approach the solution. But we still have challenges like we need to ensure fairness when comparing two different models across different types of problem, it requires careful metric selection and problem framing. When it comes to preprocessing the data, we need to be careful to handle some noisy values.

Methodology:

We will build our tool using Python and standard machine learning libraries including scikit-learn, pandas, and NumPy.

The project will be divided into two parts:
1. Classification Task: Use logistic regression, decision tree, random forest or SVM to predict diabetes using the UCI Diabetes dataset.
2. Regression Task (optional, time-permitting): Use linear regression and other regression models to forecast stock prices using historical data from Yahoo Finance.
Steps:

- Data Preprocessing: Handle missing values, normalize features, and perform feature selection/engineering.
- Model Training: Train selected models using the respective datasets.
- Hyperparameter Tuning: Optimize models using grid search or similar techniques.
- Model Evaluation: Use task-appropriate metrics and visual tools to compare performance.

Data/Resources

We will use the following resources: Kaggle: Healthcare Data Analysis and Predictive Modeling for diabetes prediction. If time permits, we will use Yahoo Finance for stock market prediction.

- Diabetes Dataset:
  This dataset consists of medical records for 768 patients and includes 8 numerical features including Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The target variable is a binary variable (0 or 1) indicating whether a patient has diabetes.
- Stock Price Dataset:
  This dataset contains 1,258 daily records of stock market data, extracted from Yahoo Finance. The features include Date, Open price, High, Low, Close, Adjusted Close, and Volume. The "Close" price can be used as the target variable for regression-based prediction.

Evaluation

We will evaluate our tool by training and testing each selected model on the same dataset and comparing their performance using appropriate metrics. For classification (diabetes prediction): we can use F1-score. For regression (stock prediction, if included): we can use Mean Squared Error (MSE)to evaluate it. We can also apply k-fold cross-validation to ensure generalizability and avoid overfitting.

By comparing the two models we can find the strengths, trade-offs, and suitability for each model. We'll also incorporate visual tools like ROC curves for classification and predicted-vs-actual plots for regression to make it easier to interpret the evaluation of used models.

Citations

1. Dataset Source:
a. Walde, R. (n.d.). Healthcare Data Analysis and Predictive Modeling. Kaggle. Retrieved March 22, 2025, from https://www.kaggle.com/code/rishavwalde/healthcare-data-analysis-and-predictive-modeling/input
b. Yahoo Finance. (2023). *AAPL: Stock performance from 2013 to 2023*. Yahoo Finance. https://finance.yahoo.com/quote/AAPL/history?p=AAPL
2. Existing Tools:
a. IBM. (n.d.). Watson Health. Retrieved March 22, 2025, from https://www.ibm.com/watson-health
b. Google. (n.d.). Google Health. Retrieved March 22, 2025, from https://health.google
c. QuantConnect. (2023). QuantConnect [Software]. https://www.quantconnect.com
3. Machine Learning Algorithms:
a. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
b. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

c. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

d. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106. https://doi.org/10.1007/BF00116251