Data Mining-Machine Learning

Data Science Engineering

Project Portfolio

Project Summary:

Diabetes mellitus, a chronic metabolic disorder, is characterized by the body's impaired ability to utilize blood sugar (glucose) effectively. The American Diabetes Association categorizes diabetes into two primary types.

- **Type 1 diabetes (previously known as insulin-dependent diabetes mellitus, IDDM):** This form often manifests in childhood. It results from an autoimmune response where the body's immune system mistakenly attacks and destroys insulin-producing beta cells in the pancreas. This destruction leads to a deficiency in insulin production. The etiology of this autoimmune response is likely multifactorial, potentially involving a combination of genetic predisposition, environmental factors, and viral infections.
- **Type 2 diabetes (previously known as non-insulin-dependent diabetes mellitus, NIDDM):** This is the more prevalent type, typically diagnosed in adulthood. It arises due to either insufficient insulin secretion or the development of insulin resistance within the body's cells. Risk factors associated with type 2 diabetes include a positive family history, obesity, and physical inactivity.

Beyond these primary types, less common forms of diabetes can occur due to genetic defects, pancreatic dysfunction, or exposure to medications or chemicals.

Gestational diabetes mellitus (GDM) is a temporary type of diabetes that can develop during pregnancy. Hormonal and metabolic changes during gestation can lead to insulin resistance, causing the body to utilize blood sugar less efficiently. While GDM typically resolves after childbirth, it increases the mother's risk of developing type 2 diabetes later in life.

Maternal inheritance of diabetes and its impact on offspring:

- Gestational diabetes itself is unlikely to directly cause diabetes in the baby.
- If the mother has pre-existing type 2 diabetes, the child has an elevated risk of developing type 2 diabetes later in life due to genetic predisposition.
- Mothers with type 1 diabetes have a slightly increased risk of their child having type 1 diabetes at birth, though this risk remains relatively low.

Diabetes is a multi-factorial disease. Many machine learning models have been built to assist doctors in the diagnosis of diabetes for future patients using different features. Many of these models have been built on the well: PIMA Indian diabetes dataset.

In this project, we will build such a model using a recent study [Chou et al., J.Pers.Med. 2023] of 15000 women aged between 20 and 80 selected as the subjects in the Taipei Municipal medical center. These women were patients who had gone to the medical center during 2018–2020 and 2021–2022 with or without the diagnosis of diabetes.

The dataset – TAIPEI_diabetes.csv

The dataset provides attributes for 15000 women on 8 features:

- Pregnancies: Number of times pregnant
- PlasmaGlucose: Plasma glucose concentration after 2 hours in an oral glucose tolerance test
- DiastolicBloodPressure: Diastolic blood pressure (mm Hg)
- TricepsThickness: Triceps skin fold thickness (mm)
- SerumInsulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigree: A function that scores the probability of diabetes based on family history
- Age: Age in years the species

And the variable to predict is in the last column of the table:

Diabetic: 1 = diabetes diagnosed, 0 = no diabetes diagnosed

Using this dataset, you should build a model to predict the Diabetic outcome (Diabetic) using the different features. You will apply the steps of a machine learning pipeline as seen in class to build and deploy a small web application that takes an input on an interface and returns the prediction for the diabetic outcome. The details of the web application implementation (framework, style, etc.) are left to you.

Each student in the group is expected to submit the following deliverables. Failure to do so is considered an incomplete submission and will result in a grade of 0:

- A Jupyter Notebook or Python script that includes exploratory analysis of the data, feature engineering and selection, model training, comparison and evaluation.
- The complete code of the web application containing the best chosen Machine Learning model as well as an interface that accepts standard inputs (this could be something as simple as a text box handling manual input values or a drop zone accepting CSV files containing rows of input) and returns predictions (in the format of your choosing, e.g., in a text box, table, downloadable CSV file, etc.). You can refer to Flask or Stremlit to get started with web application development in Python.
- A **5-10 page** report detailing your analysis and process structured in the following manner:
  - Introduction: Short description of the dataset and the problem at hand
  - Methodology: Section containing the steps of the pipeline, most notably:
    - Exploratory Data Analysis
    - Feature Engineering and Selection
    - Model Selection, Comparison and Evaluation
  - Results: Analysis of the model performances and justification of the best model choice as well as interpretation of the prediction results
  - Deployment: Short description of the web application
  - Conclusion: Concluding thoughts with insights of improving the project
- A short video of the web application running with a sample input provided and a sample prediction returned on the web interface.

You may use resources from those that are suggested in the "Project Resources" section or others as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. For example, you can validate your model using the PIMA Indian diabetes dataset. **However, this comes with a big responsibility: any submission that is over-plagiarized or does not reflect personal work will not be accepted**.

Project Resources:

Here are resources that may be helpful for the project:

- **Original study for the project dataset**: Chou CY, Hsu DY, Chou CH. Predicting the Onset of Diabetes with Machine Learning Methods. J Pers Med. 2023 Feb 24;13(3):406. doi:10.3390/jpm13030406. PMID: 36983587; PMCID: PMC10057336.
- PIMA Indian dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- WHO Diabetes webpage: https://www.who.int/news-room/fact-sheets/detail/diabetes

Project Evaluation:

The project will be evaluated using the following rubric. It contains the required items for a complete submission as well as bonus elements. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Jupyter Notebook (or Python script) containing entire machine learning pipeline [18 point]
- Complete web application code [18 point]
- Report (in PDF format) [18 point]
- Web application short demo video [18 point]
- GitHub repository [18 point]
- BONUS: Best group model performance in class [10 point]

Prepared By Dr. KHIM Chamroeun