BIRMINGHAM CITY
University

Using supervised learning for the binary classification of Type 2 Diabetes

Lewis Higgins - Student ID 22133848

CMP6202 - Artificial Intelligence & Machine Learning

Module Coordinator: Nouh Elmitwally

Contents

Abstract

Probably would benefit from an abstract. You can't really write this until the very end though, so return to it then. **The example work is from a previous year wherein this assessment was a group task. You can see that each group member developed one ML model, but you seem to be developing all of them yourself, so don't be mislead by the report titles only mentioning one model.**

Your EDA can be very extensive, and you could potentially have pages and pages and pages of it; this isn't a bad thing. The vast majority of any ML-related work is EDA because it gives you the background information on the dataset to then apply when training the model, such as the identification of non-numerical columns and encoding them into numerical equivalents where possible so that they become useful training data for the model, as ML models cannot interpret strings.

Notes, remove before final

0.1 Mihai W7

0.1.1 Overall notes - Beginning of talk

- Find a problem before a dataset (No)
- Identify if said problem is regression or classification
- It's hard to use a classification dataset where the target column is just a bunch of labels (???)
  - This may be in reference to Week 7's wine set where you had 7 different levels of quality but converted them to 2 for binary classification.
- Examples on Moodle
- Datasets on Moodle but it's almost 100% that someone else will have used them.
- I also don't know if you're even allowed to use them.
- A dataset can have more than one target column (though the ones you use likely won't).
- Mihai strongly warns that ML is a very iterative process, and your first attempt will probably be poor.
  - This is why you use **pipelines** as in CMP6230 but that's not this module.
- You will be asked questions on it in the presentation **RELATED TO WHAT YOU'VE DONE IN CLASS**, likely by Mihai himself, and he is trying to catch you out.

0.1.2 Section 2

- After you have an identified dataset, begin EDA.
- Describe how you split the dataset into training and testing dataset.
- If the machine sees the test data, it may "overfit".
  - You've seen this before where the line of best fit isn't really a line and just connects every single point, meaning it's really good at guessing the data it already knows but not new data.
- You split your data *BEFORE* EDA to avoid "conceptual overfitting".
- When splitting data, you need to think about data imbalance i.e. training set having too much of the one option and few of the other (too many True, not enough False).
- SKLearn may try to fix that for you.
- Identify outliers, missing data.

- "Missing information can be information in itself". Consider the source of the data (not Kaggle but rather where the Kaggle author got it from)
  - You might be able to impute data rather than deleting it.
  - Conserve as much data as you can, deleting data should be a last option.
- Erroneous data
  - Another reason why your data source is important.
  - Could just be mistyped, see what the erroneous data actually is to see if you can correct it.
- Outliers
  - Is it significant enough to remove it? Is it definitely an outlier; could it feasibly be true? (speed camera example where one guy went 30 but another went 100, but that's still plausible.)
  - Boxplots can identify outliers.
- If your dataset is bad, your model will be, too.

0.1.3 Section 3

- Identify the right algorithm.
  - Decision trees are good at classification.
  - Random forest is also used for it.
  - Naive Bayes and KNN work for prediction and classification.
  - Some of these may perform worse with higher amounts of data, look into them.
- Performance won't matter a massive amount but you still need to be able to justify why it was good.
  - Also justify its downsides and limitations.
  - And the limtiations of the dataset itself.

0.1.4 Section 4

- Encoding is important here because ML doesn't use text.
  - Side-note: Even if it does that's actually just an abstraction and it's just encoding it under the hood.
- Fine-tuning
  - Playing around with parameters of the functions.
  - KNN Neighbours and such

  - Often comes after an initial test run
  - If your accuracy is really low (20% was the example), fine-tuning won't help and you just need to redo the entire work.
  - Could maybe give you an extra 5% accuracy.
  - Decision trees may have multiple versions. SKLearn's decision tree may be different from a CUDA one.

- Evaluation metrics
  - Accuracy is not Precision. Which do you need?
  - You want both to be high, if one lags massively behind the other then that's bad.
  - ML is an iterative process until you can get the best performing model.

0.1.5 Section 5

- Visualisation of evaluation results
  - For a correlation matrix, a heatmap is better than a bar plot for example.

Introduction

Diabetes mellitus, or type 2 diabetes, accounts for 90% of the 4.4 million cases of diabetes in the UK, and it is estimated that there are 1.2 million undiagnosed cases of type 2 diabetes across the country (Diabetes UK, 2024). The rate of type 2 diabetes per 100,000 individuals is rapidly increasing, with Khan et al. (2020)'s analysis projecting that by 2030, the rate will reach 7,079 per 100,000. Many people with diabetes suffer immensely reduced quality of life, with approximately 50% of patients suffering from peripheral neuropathy (Dhanapalaratnam et al., 2024), an irreversible disability which causes immense pain due to nerve damage from high blood sugar (NHS, 2022), which can occur when the patient was unaware they even had diabetes.

Therefore, it is imperative that systems are put in place to enable the swift diagnosis of diabetes, especially type 2 diabetes given its major prevalence. This can be accomplished by training machine learning models on existing clinical datasets to identify common trends in those with and without type 2 diabetes. This report will document the planning, development and evaluation of multiple machine learning models in their classification of whether individuals have type 2 diabetes based on multiple clinical factors, specifically through the stages of:

- Dataset Identification
- Data Integration
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Model Development
- Model Evaluation
- Research Conclusions

1.1 Dataset Identification

Machine learning models require large amounts of data to train upon, meaning a dataset must be identified consisting of many rows and features. This project identified two datasets which could be integrated into one larger dataset, the first of which being the well-reputed Pima Indian[1] Diabetes Database (UCI Machine Learning, 2024), sourced from Kaggle, a platform for students and researchers alike to download and upload datasets and code for research purposes. The dataset contains data on Pima Indian women in Phoenix, Arizona, USA, and has previously seen wide use across academic literature relating to machine learning (AlZu'bi et al. (2023), Zou, Zhang, and Chen (2024), Joshi and Dhakal (2021), Hayashi and Yukita (2016)), where other researchers have also aimed to solve the problem of diabetes classification via supervised learning. This dataset contains 768 rows with 9 features.

This project also includes a second dataset, also from Kaggle, that has been previously used in literature by Zou, Zhang, and Chen (2024). This dataset (John DaSilva, 2024) is based on data from female patients in Frankfurt, Germany, and includes the same 9 features as the Pima Indian dataset, but includes 2000 rows. By integrating these two datasets into one larger dataset of 2768 rows, it will be possible to give the machine learning models more data to train upon.

Table 1.1 details the 9 features seen in both datasets and their descriptions.

Feature	Description
Pregnancies	The number of pregnancies the patient has had.
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
BloodPressure	Diastolic blood pressure in mm/Hg.
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin.
BMI	Body Mass Index, calculated from the patient's weight and height.
DiabetesPedigreeFunction	The product of a function to ascertain the probability of diabetes based on family genetics. (Akmeşe, 2022)
Age	The patient's age.
Outcome	Whether the patient is likely to develop diabetes.

Table 1.1: The features seen in both datasets.

1.2 Supervised learning task identification

As previously mentioned, it is possible for patients to have diabetes without knowing. Therefore, it is paramount that swift and simple diagnosis methods are put in place, which can

[1] "Pima Indian" refers to a specific Native American ethnic group rather than people from India.

be achieved through the use of supervised learning classification models. This requires the existence of the "ground truth", which refers to the label given to data that indicates its class (c3.ai, 2024). Within these datasets, the ground truth is present as the 'Outcome' feature, which will be used as the target variable for the produced classification models.

Exploratory Data Analysis

This chapter details the EDA processes undertaken with the datasets, including key questions that will be answered by the process, as well as the splitting of the data into training, testing sets.

2.1 Question identification

The key factors involved in the diagnosis of diabetes are critical to understand, which can be solved through EDA on these datasets.

2.2 Data Integration

The two datasets must first be merged into one to allow for an overall analysis to be performed. This is a simple process because they both contain the same 9 features, and is detailed in Figure **??**.

2.3 Splitting the dataset

2.4 EDA process and results

2.5 EDA conclusions

Experimental Design

This chapter details the planned algorithms to be leveraged against this dataset, as well as the metrics to evaluate them. Furthermore, the data cleaning and preprocessing stages will be deeply explored, as well as some potential limitations relating to their use.

3.1 Identification of chosen algorithms

3.2 Identification of appropriate evaluation techniques

3.3 Data Cleaning and Pre-processing Transformations

3.4 Limitations and Options

Model Development

This chapter details the training and evaluation processes of the original produced models before any iterative improvements such as hyperparameter tuning.

4.1 Predictive modelling process

4.2 Results on seen data

Evaluation and further improvements

This chapter details the extensive evaluation of each model, as well as iterative improvements that were made to enhance their performance.

Conclusion

6.1 Summary of results

6.2 Reflection on Individual Learning

Bibliography

Akmeşe, Ömer Faruk (Mar. 30, 2022). "Diagnosing Diabetes with Machine Learning Techiques". In: *Hittite Journal of Science and Engineering* 9 (1), pp. 9–18. ISSN: 2148-4171. DOI: 10.17350/HJSE19030000250.

AlZu'bi, Shadi, Mohammad Elbes, Ala Mughaid, Noor Bdair, Laith Abualigah, Agostino Forestiero, and Raed Abu Zitar (Feb. 2023). "Diabetes Monitoring System in Smart Health Cities Based on Big Data Intelligence". In: *Future Internet* 15 (2). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 85. ISSN: 1999-5903. DOI: 10.3390/fi15020085.

c3.ai (2024). *What is Ground Truth? | Machine Learning Glossary Definition*. C3 AI. URL: https://c3.ai/glossary/machine-learning/ground-truth/ (visited on 12/09/2024).

Dhanapalaratnam, Roshan, Tushar Issar, Leiao Leon Wang, Darren Tran, Ann M. Poynten, Kerry-Lee Milner, Natalie C.G. Kwai, and Arun V. Krishnan (Aug. 21, 2024). "Effect of Metformin on Peripheral Nerve Morphology in Type 2 Diabetes: A Cross-Sectional Observational Study". In: *Diabetes* 73 (11), pp. 1875–1882. ISSN: 0012-1797. DOI: 10.2337/db24-0365.

Diabetes UK (2024). *How many people in the UK have diabetes?* Diabetes UK. URL: https://www.diabetes.org.uk/about-us/about-the-charity/our-strategy/statistics (visited on 11/27/2024).

Hayashi, Yoichi and Shonosuke Yukita (Jan. 1, 2016). "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset". In: *Informatics in Medicine Unlocked* 2, pp. 92–104. ISSN: 2352-9148. DOI: 10.1016/j.imu.2016.02.001.

John DaSilva (2024). *Frankfurt Diabetes Dataset*. diabetes. URL: https://www.kaggle.com/datasets/johndasilva/diabetes (visited on 11/25/2024).

Joshi, Ram D. and Chandra K. Dhakal (July 9, 2021). "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches". In: *International Journal of Environmental Research and Public Health* 18 (14), p. 7346. ISSN: 1661-7827. DOI: 10.3390/ijerph18147346.

Khan, Moien Abdul Basith, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi (Mar. 2020). "Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends". In: *Journal of Epidemiology and Global Health* 10 (1), pp. 107–111. ISSN: 2210-6006. DOI: 10.2991/jegh.k.191028.001.

NHS (Oct. 16, 2022). *Peripheral neuropathy - Causes*. nhs.uk. Section: conditions. URL: https://www.nhs.uk/conditions/peripheral-neuropathy/causes/ (visited on 12/04/2024).

UCI Machine Learning (2024). *Pima Indians Diabetes Database*. Pima Indians Diabetes Database. URL: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (visited on 11/25/2024).

Zou, Qiong, Yang Zhang, and Chang Sheng Chen (Feb. 13, 2024). "Construction and Application of a Machine Learning Prediction Model Based on Unbalanced Diabetes Data Fusion". In: *Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence*. JCRAI '23. New York, NY, USA: Association for Computing Machinery, pp. 114–123. ISBN: 979-8-4007-0770-4. DOI: 10.1145/3632971.3633348.