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.**

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, 2024a). 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 Diabetes Database (UCI Machine Learning, 2024), downloaded from Kaggle, a platform for students and researchers alike to download and upload datasets and code for research purposes. The data originates from the National Institute of Diabetes and Digestive and Kidney Diseases, who collected this data from Pima Indian[1] women aged 21 and over in hospitals in Phoenix, Arizona, USA, and it 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] "Pima Indian" refers to a specific Native American ethnic group rather than people from India.

1.2 Supervised learning task identification

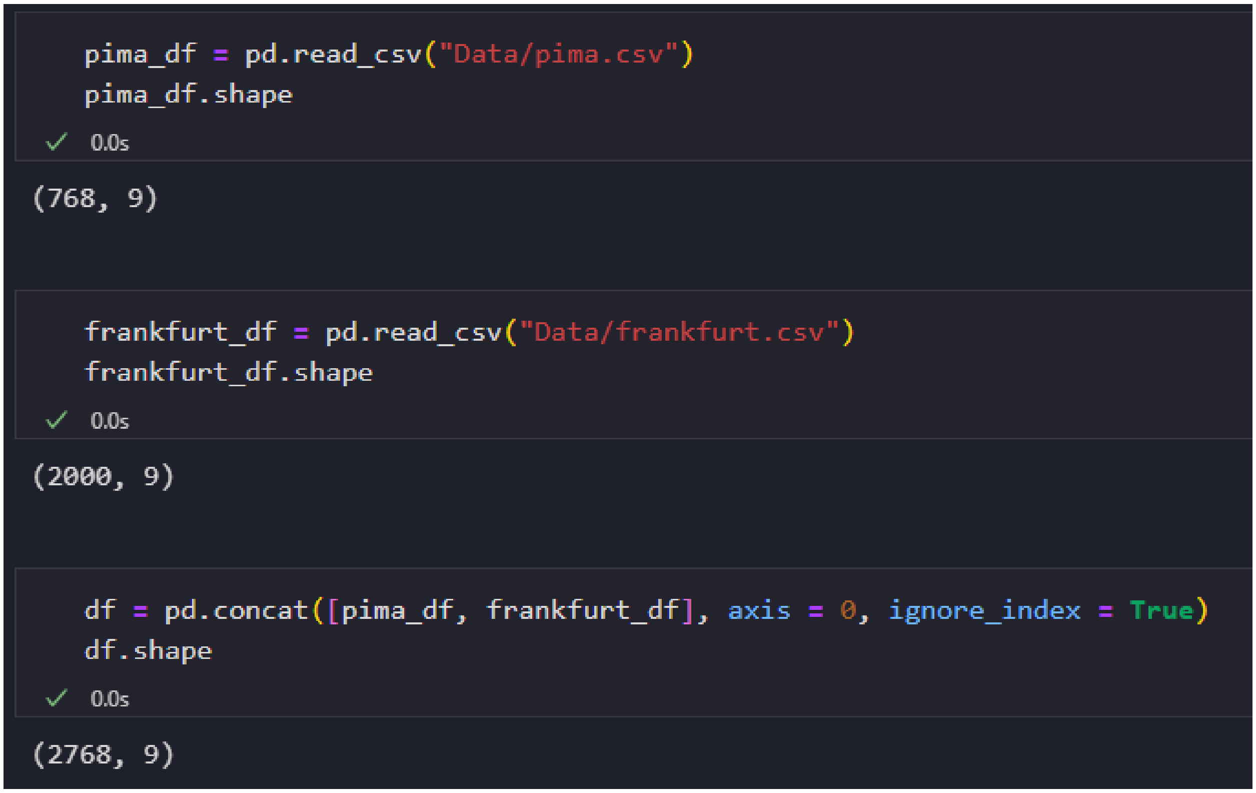
As previously mentioned, it is possible for patients to have diabetes without knowing. Therefore, it is paramount that swift and simple diagnosis methods are put in place, which can be achieved through the use of supervised learning classification models. This requires the existence of the "ground truth", 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 and testing sets.

2.1 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 2.1.



```
pima_df = pd.read_csv("Data/pima.csv")
pima_df.shape
```

✓ 0.0s

```
(768, 9)
```

```
frankfurt_df = pd.read_csv("Data/frankfurt.csv")
frankfurt_df.shape
```

✓ 0.0s

```
(2000, 9)
```

```
df = pd.concat([pima_df, frankfurt_df], axis = 0, ignore_index = True)
df.shape
```

✓ 0.0s

```
(2768, 9)
```

Figure 2.1: Integrating the two separate datasets into one larger dataset.

2.2 Question identification and assumptions

The key factors involved in the diagnosis of diabetes are critical to understand, which can be solved through EDA on these datasets. It is possible to make various assumptions based on topical background research of each of the features in the dataset, detailed in Table 2.1

Feature	Research-based assumptions
Pregnancies	Approximately 13.4% of pregnant women develop a temporary condition known as Gestational Diabetes Mellitus (GDM), which typically subsides after birth (Adam et al., 2023). However, research by (Dennison et al., 2021) indicates that 33% of women who develop GDM will go on to develop permanent diabetes mellitus within 15 years. Therefore, it is assumed that pregnancies will positively correlate with the diabetes outcome. It is also expected that pregnancies should naturally positively correlate with age.
Glucose	Glucose concentrations are an enormous factor in the diagnosis of diabetes mellitus, being one of the main metrics used to certify the condition, where results over 200mg/dL mean an absolute diagnosis[1] (Aftab et al., 2021). It is therefore assumed that the glucose concentrations will be one of the strongest influences of the outcome, and that it will also correlate heavily with insulin levels.
BloodPressure	Diastolic blood pressure (DBP) does influence the diagnosis of diabetes mellitus, as 56.2% of recently diagnosed patients presented with elevated DBP in Nelaj et al. (2023)'s limited study of 126 patients, but it is not a decisive factor by itself. Therefore, it is assumed that there will be some correlation between DBP and the outcome, but not as major as other factors like plasma glucose levels.
SkinThickness	It is a frequent assumption even non-academically that people who weigh more, and by consequence have higher skin thickness in certain areas such as the triceps, have a higher risk of developing conditions like type 2 diabetes. This is backed by a study by Ruiz-Alejos et al. (2020), which found strong associations between skin thickness and diabetes mellitus, as well as high blood pressure. Therefore, it is assumed that there will be a strong correlation between tricep skin thickness and the outcome, as well as an expectation of strong correlations between thickness, BMI and blood pressure.
Insulin	Diabetes mellitus is directly associated with insulin deficiency, and as such, it is assumed that this factor will be the strongest influence in the outcome. This is because 2-hour serum insulin tests, as used in this dataset, are frequently part of HOMA-IR[2] assessments.

[1] The other main metric is insulin deficiency, meaning that the patient could have glucose levels lower than 200mg/dL and still be diagnosed if they are instead insulin deficient. (Aftab et al., 2021).

[2] Homeostasis Model Assessment of Insulin Resistance, used to measure insulin resistance (Tahapary et al., 2022), which can be used in both type 1 and type 2 diabetes diagnosis (Khalili et al., 2023).

BMI	BMI is likely to be a significant factor in the outcome, which is backed by previous academic studies indicating that 71% of studied individuals showed increases in BMI prior to diagnosis (Donnelly, McCrimmon, and Pearson, 2024). Additionally, BMI is used in insulin resistance measurement assessments, which are key assessments in diabetes diagnosis, meaning that it is a safe assumption that BMI will be a large factor in the outcome.
DiabetesPedigreeFunction	People are more likely to develop diabetes mellitus if there is a family genetic history of the condition, though it is not directly caused by any one particular gene (Diabetes UK, 2024b). With the pedigree function aiming to quantify the inheritance probability, it can be assumed that it will likely correlate heavily with the outcome.
Age	Suastika et al. (2012) studied the effects of age as a risk factor for diabetes mellitus, finding that many natural associated factors of ageing including increases in body fat and decreases in lipid metabolism had considerable influence on the development of insulin resistance and diabetes mellitus by consequence. Therefore, it is likely that there will be a noticeable correlation between a patient's age and the outcome.

Table 2.1: Research-based assumptions prior to any EDA.

Based on these assumptions, the questions that this EDA process aims to answer are:

ID	Research-based assumptions
1	Are there any missing values or values that are not physically possible?
2	Are there any significant outliers?
3	Is the dataset evenly balanced in terms of the outcome? If not, what should be done?
4	Does the rate of diabetes positively correlate with the amount of pregnancies a woman has had?
5	Does the amount of pregnancies influence any of the other features?
6	What is the distribution of blood glucose levels in patients with and without diabetes?
7	Does BMI influence glucose levels?
8	Is diastolic blood pressure a worthwhile diagnosis method in this dataset?
9	Is the average skin thickness of those with diabetes actually higher than those without?
10	How does the relationship between insulin and glucose change between those with and without diabetes?

Table 2.2: The questions that this EDA process aims to answer.

2.3 Splitting the dataset

It is good practice to first split the data into training and testing sets before performing exploratory data analysis to avoid conceptual overfitting, also known as data leakage. Conceptual overfitting occurs when insights gained from the entire dataset influence model development decisions, which may eventually lead to actual overfitting. By excluding the training data from the analysis, it effectively simulates a real-world environment where the data being given to the model is not known, even to its developers.

Splitting the data is a mandatory process when developing supervised learning models, primarily for the prevention of overfitting. Overfitting is a significant challenge in machine learning, where models can perform exceptionally well on their original training data but are unable to generalize to unseen data, making them unsuitable for deployed use. The training set must consist of a large portion of the data so that the model has enough information to analyse and discover trends within, whereas the testing set is a smaller, unseen remainder of the data that the model's predictions can be evaluated against using various metrics. A key point of determination is the proportions of the dataset that should go in each set - there is no 'one-size-fits-all' percentage that can provide the best results for every possible dataset (Sivakumar, Parthasarathy, and Padmapriya, 2024), and factors such as the size of the dataset play a large part in this. Most commonly, splits are either 70:30 or 80:20 for training and testing sets respectively.

The integrated dataset for this project is 2,768 rows. This is considered to be a small dataset, and as such, it will be best to maximize the size of the training split, so a split of 80% training and 20% testing was used, visualized in Figure 2.2.

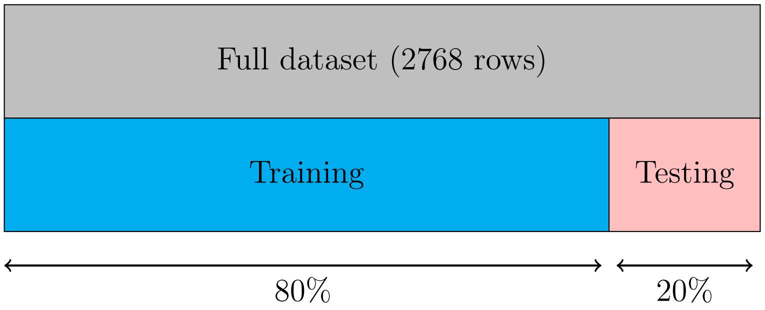
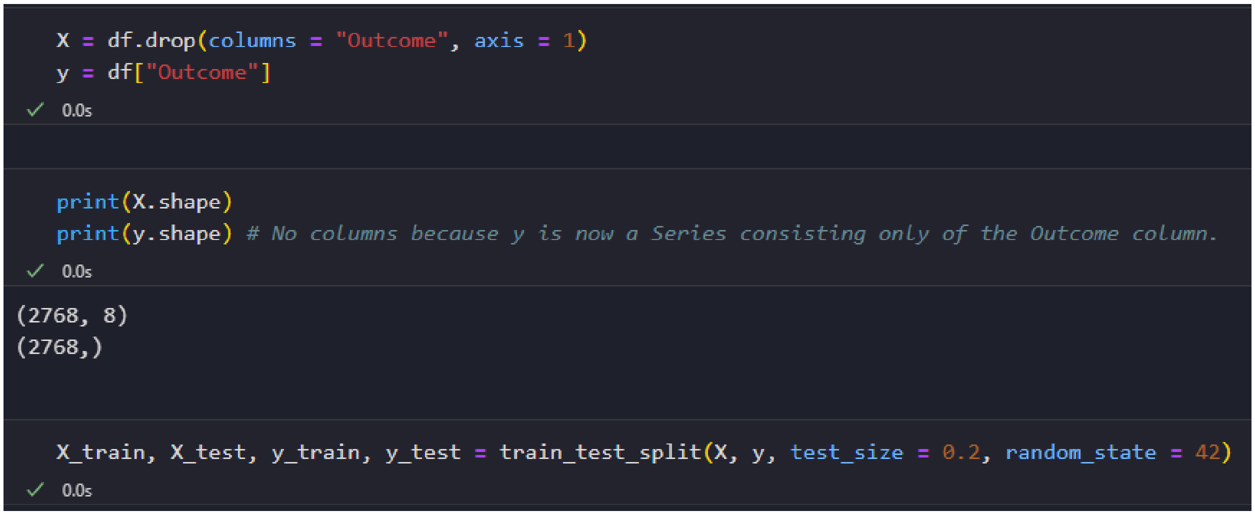


Figure 2.2: A visual representation of the train/test split.

To accomplish this, the data must first be split into X and y tables, where X consists of the eight features, and y is the target variable. After the data is split to X and y, it can be split into training and testing sets through Scikit-Learn's "train_test_split" method, as depicted in Figure 2.3



```
X = df.drop(columns = "Outcome", axis = 1)
y = df["Outcome"]
```
✓ 0.0s

```
print(X.shape)
print(y.shape) # No columns because y is now a Series consisting only of the Outcome column.
```
✓ 0.0s

```
(2768, 8)
(2768,)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```
✓ 0.0s

Figure 2.3: Splitting the data at an 80:20 ratio.

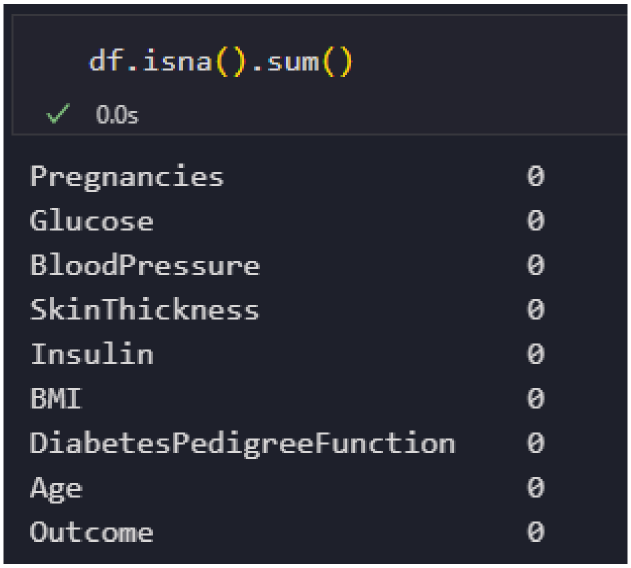
By default, this method will first shuffle all rows in the dataset before splitting it, which introduces an element of randomness which can damage reproducibility. To combat this, the "random_state" parameter can be set to ensure that the same shuffle will occur every time.

2.4 EDA process and results

These subsections correspond in order to each of the questions posed in Table 2.2. Problems identified here are not solved within this section, and are instead solved in Section 3.3.

2.4.1 Identification of missing or impossible values

An initial glance at the dataset would make it appear as though as there are not any missing values in the dataset, as shown in Figure 2.4.

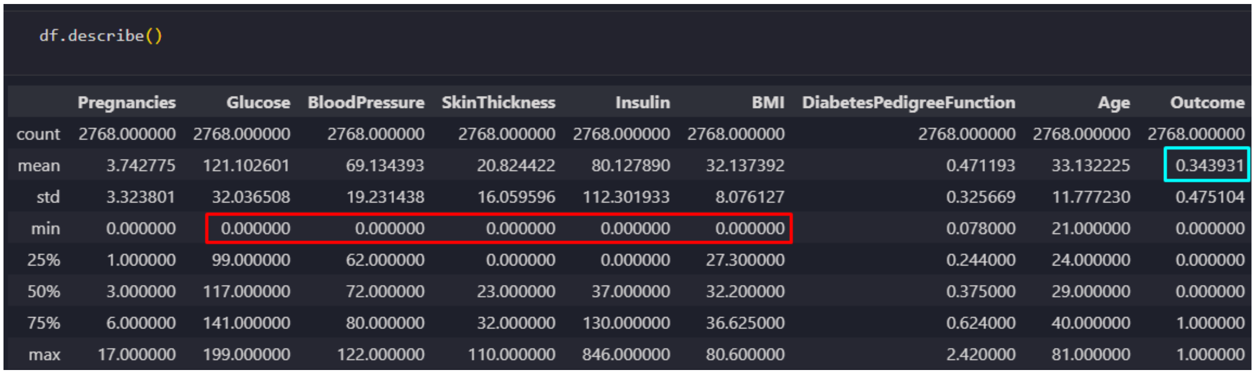


```
df.isna().sum()
```
✓ 0.0s

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
```

Figure 2.4: An initial count of missing values before any further analysis.

This would be very good - if it were true. Instead of missing values, this dataset contains impossible values in five columns, highlighted in red in Figure 2.5.[3]

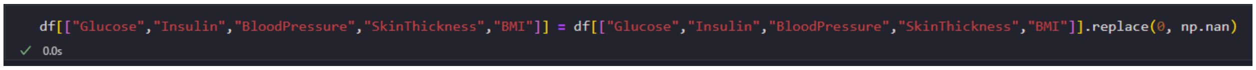


```
df.describe()
```

```
       Pregnancies      Glucose  BloodPressure  SkinThickness      Insulin          BMI  DiabetesPedigreeFunction          Age      Outcome
count  2768.000000  2768.000000    2768.000000    2768.000000  2768.000000  2768.000000               2768.000000  2768.000000  2768.000000
mean      3.742775   121.102601      69.134393      20.824422    80.127890    32.137392                  0.471193    33.132225     0.343931
std       3.323801    32.036508      19.231438      16.059596   112.301933     8.076127                  0.325669    11.777230     0.475104
min       0.000000     0.000000       0.000000       0.000000     0.000000     0.000000                  0.078000    21.000000     0.000000
25%       1.000000    99.000000      62.000000       0.000000     0.000000    27.300000                  0.244000    24.000000     0.000000
50%       3.000000   117.000000      72.000000      23.000000    37.000000    32.200000                  0.375000    29.000000     0.000000
75%       6.000000   141.000000      80.000000      32.000000   130.000000    36.625000                  0.624000    40.000000     1.000000
max      17.000000   199.000000     122.000000     110.000000   846.000000    80.600000                  2.420000    81.000000     1.000000
```

Figure 2.5: The overall dataset description. Physically impossible values are in red.

It is not possible for a living person to have a 0 in any of the five highlighted columns, and these datasets do not contain information of the deceased. Therefore, it can be deduced that these values are erroneous, and should be treated as though they are missing. It is theorised by some in Kaggle discussions of these datasets that the zeroes in the Insulin column actually meant "imperceptible levels", which would have been useful data. However, a further review of literature around the datasets, especially Hayashi and Yukita (2016)'s paper, led to the discovery that these zeroes truly are missing values that were missing due to experimental invalidity at the time of their collection.



```
df[["Glucose","Insulin","BloodPressure","SkinThickness","BMI"]] = df[["Glucose","Insulin","BloodPressure","SkinThickness","BMI"]].replace(0, np.nan)
✓ 0.0s
```

Figure 2.6: Converting the impossible values to NaNs which are recognised by Pandas.

Now that there are officially recognised missing values, they can be visualised using the Missingno package, which can produce a matrix of missing values by column, seen in Figure 2.7.

[3] The mean of the outcome is highlighted in blue, as the fact that it is under 0.5 indicates that there are more cases of outcome 0 (no diabetes) than 1 (diabetes). This will be further analysed in Section 2.4.3.

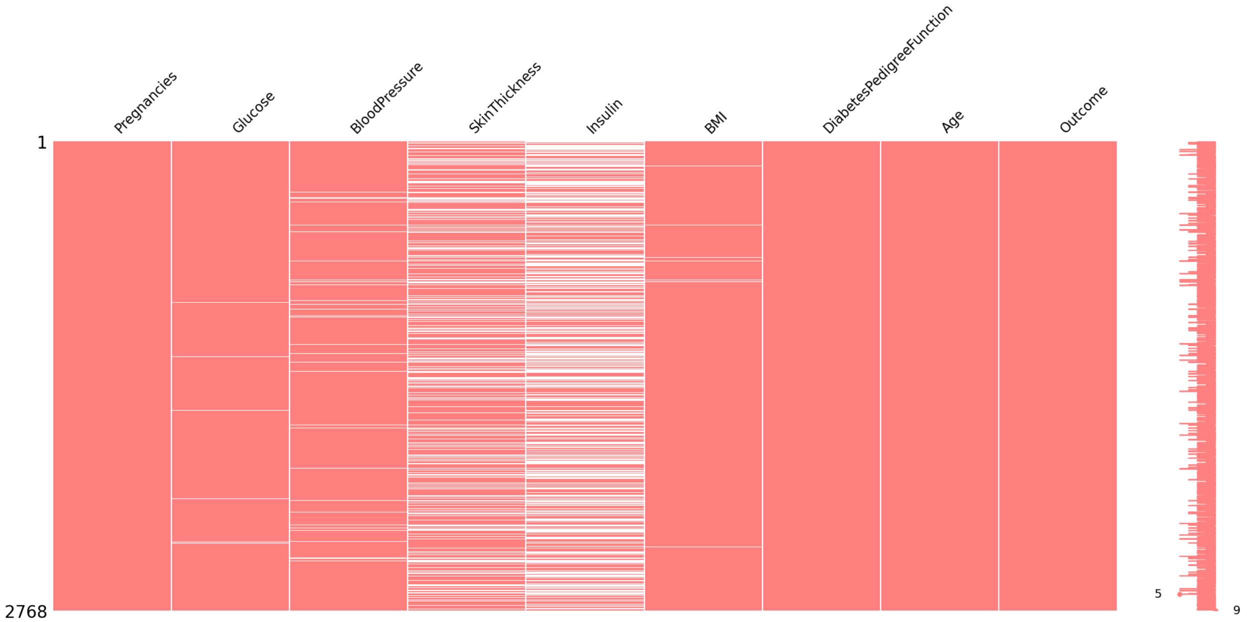


Figure 2.7: A matrix of the missing data per column.

This matrix reveals that the Insulin and SkinThickness columns contain considerable amounts of missing data alongside many missing rows of BloodPressure, and some missing rows of BMI and Glucose, which will need to be fixed before any model can be trained. This will be accomplished through imputation in Section 3.3.

2.4.2 Identification of outliers

2.4.3 Dataset imbalance

2.4.4 Influence of pregnancies

On the diabetes outcome

On other factors

2.4.5 Glucose levels with and without diabetes

2.4.6 Influence of BMI

2.4.7 Relevance of diastolic blood pressure

2.4.8 Skin thickness with and without diabetes

2.4.9 Insulin and glucose with and without diabetes

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

Adam, Sumaiya, Harold David McIntyre, Kit Ying Tsoi, Anil Kapur, Ronald C. Ma, Stephanie Dias, Pius Okong, Moshe Hod, Liona C. Poon, Graeme N. Smith, Lina Bergman, Esraa Algurjia, Patrick O'Brien, Virna P. Medina, Cynthia V. Maxwell, Lesley Regan, Mary L. Rosser, Bo Jacobsson, Mark A. Hanson, Sharleen L. O'Reilly, Fionnuala M. McAuliffe, and the FIGO Committee on the Impact of Pregnancy on Long-term Health and the FIGO Division of Maternal and Newborn Health (2023). "Pregnancy as an opportunity to prevent type 2 diabetes mellitus: FIGO Best Practice Advice". In: *International Journal of Gynecology & Obstetrics* 160 (S1), pp. 56–67. ISSN: 1879-3479. DOI: 10.1002/ijgo.14537.

Aftab, Shabib, Saad Alanazi, Munir Ahmad, Muhammad Adnan Khan, Areej Fatima, and Nouh Sabri Elmitwally (2021). "Cloud-Based Diabetes Decision Support System Using Machine Learning Fusion". In: *Computers, Materials & Continua* 68 (1), pp. 1341–1357. ISSN: 1546-2226. DOI: 10.32604/cmc.2021.016814.

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).

Dennison, Rebecca A., Eileen S. Chen, Madeline E. Green, Chloe Legard, Deeya Kotecha, George Farmer, Stephen J. Sharp, Rebecca J. Ward, Juliet A. Usher-Smith, and Simon J. Griffin (Jan. 2021). "The absolute and relative risk of type 2 diabetes after gestational diabetes: A systematic review and meta-analysis of 129 studies". In: *Diabetes Research and Clinical Practice* 171, p. 108625. ISSN: 01688227. DOI: 10.1016/j.diabres.2020.108625.

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 (2024a). *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).

Diabetes UK (2024b). *What causes type 2 diabetes?* Diabetes UK. URL: https://www.diabetes.org.uk/about-diabetes/type-2-diabetes/causes (visited on 12/14/2024).

Donnelly, Louise A., Rory J. McCrimmon, and Ewan R. Pearson (2024). "Trajectories of BMI before and after diagnosis of type 2 diabetes in a real-world population". In: *Diabetologia* 67 (10), pp. 2236–2245. ISSN: 0012-186X. DOI: 10.1007/s00125-024-06217-1.

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.

Khalili, Davood, Marjan Khayamzadeh, Karim Kohansal, Noushin Sadat Ahanchi, Mitra Hasheminia, Farzad Hadaegh, Maryam Tohidi, Fereidoun Azizi, and Ali Siamak Habibi-Moeini (Feb. 14, 2023). “Are HOMA-IR and HOMA-B good predictors for diabetes and pre-diabetes subtypes?” In: *BMC Endocrine Disorders* 23, p. 39. ISSN: 1472-6823. DOI: 10.1186/s12902-023-01291-9.

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.

Nelaj, Ergita, Margarita Gjata, Irida Kecaj, Ilir Gjermeni, and Mihal Tase (June 2023). “HIGH BLOOD PRESSURE IN THE NEWLY DIAGNOSED TYPE 2 DIABETES PATIENTS”. In: *Journal of Hypertension* 41 (Suppl 3), e172. ISSN: 0263-6352. DOI: 10.1097/01.hjh.0000940640.80128.7a.

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).

Ruiz-Alejos, Andrea, Rodrigo M Carrillo-Larco, J Jaime Miranda, Robert H Gilman, Liam Smeeth, and Antonio Bernabé-Ortiz (Jan. 2020). “Skinfold thickness and the incidence of type 2 diabetes mellitus and hypertension: an analysis of the PERU MIGRANT study”. In: *Public Health Nutrition* 23 (1), pp. 63–71. ISSN: 1368-9800. DOI: 10.1017/S1368980019001307.

Sivakumar, Muthuramalingam, Sudhaman Parthasarathy, and Thiyagarajan Padmapriya (Sept. 6, 2024). “Trade-off between training and testing ratio in machine learning for medical image processing”. In: *PeerJ Computer Science* 10, e2245. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.2245.

Suastika, Ketut, Pande Dwipayana, Made Siswadi, and R.A. Tuty (Dec. 12, 2012). “Age is an Important Risk Factor for Type 2 Diabetes Mellitus and Cardiovascular Diseases”. In: *Glucose Tolerance*. Ed. by Sureka Chackrewarthy. InTech. ISBN: 978-953-51-0891-7. DOI: 10.5772/52397.

Tahapary, Dicky Levenus, Livy Bonita Pratisthita, Nissha Audina Fitri, Cicilia Marcella, Syahidatul Wafa, Farid Kurniawan, Aulia Rizka, Tri Juli Edi Tarigan, Dante Saksono Harbuwono, Dyah Purnamasari, and Pradana Soewondo (Aug. 1, 2022). “Challenges in the diagnosis of insulin resistance: Focusing on the role of HOMA-IR and Tryglyceride/glucose index”. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 16 (8), p. 102581. ISSN: 1871-4021. DOI: 10.1016/j.dsx.2022.102581.

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.