Diabetes Trait Analysis and Prediction

By

Minyang Yu

MSc Dissertation

UNIVERSITY OF CARDIFF

A MSc dissertation submitted to the University of Cardiff
in accordance with the requirements of the degree of
MASTER OF COMPUTING AND IT MANAGEMENT in the
Faculty of Computer Science and Informatics.

May 6, 2024

Declaration of own work

I declare that the work in this MSc dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Minyang Yu, 5/5/2024

Acknowledgement

I would like to thank my supervisor Luis. He gave me great support in guidelines for writing a dissertation.

Abstract

Diabetes is one of the most common diseases among middle-aged and elderly populations, significantly reducing the quality of life for those affected. Consequently, it is critical to proactively prevent and predict the propensity for this condition. This study explores a range of symptoms commonly associated with diabetes, such as polyuria and obesity, using a dataset from the University of California, Irvine's machine learning repository. Through targeted analysis, we have identified the interrelationships between various traits, their impact on the likelihood of developing diabetes, and the potential for early disease prediction. Additionally, we offer relevant health recommendations based on our findings.

Keywords: diabetes, machine learning, trait analysis

Number of words in the dissertation: .... words.

Contents

List of Tables

1 Introduction

Diabetes mellitus, characterized by the body's inability to produce or respond adequately to insulin, leads to significant metabolic dysfunctions involving carbohydrates, proteins, and fats, contributing to severe long-term complications in the vasculature, nervous system, and various organs. This condition manifests clinically with symptoms such as increased urination, intense thirst, excessive eating, and unexplained weight loss. Given its status as the fifth leading cause of death globally, enhancing public awareness and early intervention for diabetes is crucial.

Diabetes affects individuals of all ages, manifesting differently across the lifespan. Type 1 diabetes, also known as insulin-dependent or juvenile diabetes, is marked by the body's failure to produce insulin and typically presents suddenly in younger individuals. Conversely, Type 2 diabetes, associated with ineffective insulin use, has traditionally been linked to adults but is increasingly diagnosed in children, paralleling rising trends in obesity and sedentary lifestyles among the youth.

The linkage between obesity and the development of Type 2 diabetes is well-established, with excess body weight and physical inactivity being primary risk factors. However, data analysis reveals a weak association between obesity and diabetes in the study population, predominantly middle-aged adults who are more susceptible to Type 2 diabetes. This suggests a prevalence of non-obese Type 2 diabetes, possibly driven by genetic factors or cellular dysfunctions that impede insulin action beyond mere insulin resistance.

Recent statistics from the International Diabetes Federation (IDF) indicate that diabetes affected approximately 9.3% of the global population aged 20 to 79 in 2019, with projections suggesting an increase to 10.2% by 2030 and 10.9% by 2045. Despite numerous studies leveraging machine learning to predict diabetes, few have comprehensively integrated various symptomatic data, which is essential for nuanced understanding and management of the disease.

This research therefore integrates data analysis with machine learning techniques to draw more detailed insights into the multifactorial aspects of diabetes, aiming to bridge gaps in current methodologies and enhance predictive accuracies.

To deepen our comprehension of the association between various symptoms and diabetes, this study aims to address four critical inquiries:

- Does aging increase the susceptibility to diabetes, suggesting a higher incidence in older populations?
- Between genders, which is more predisposed to developing diabetes, and what might this imply about hormonal or biological influences on disease prevalence?
- Among the various clinical and physiological attributes examined, which ones significantly influence the likelihood of a diabetes diagnosis?
- What symptoms should be considered early indicators of diabetes to facilitate timely intervention and management?

These questions are essential for developing targeted prevention strategies and improving diagnostic criteria, thereby enhancing patient outcomes in populations at risk.

2 Literature Review

Recent research has made significant advancements in the application of machine learning to predict medical outcomes, particularly diabetes and its associated conditions. This review synthesizes findings from key studies that leverage various statistical models to enhance predictive accuracy and clinical utility.

One seminal contribution to this field was detailed by authors in a notable paper [1], where decision trees and neural networks were utilized to predict diabetes within an Iranian cohort. Both models demonstrated robust accuracy, marking a foundational advance in diabetes prediction methodologies.

Further investigations have expanded the scope to encompass patients with comorbid conditions. For instance, a pivotal study [2] integrated multiple algorithms—including logistic regression, decision trees, random forest, and support vector machines—to predict diabetes among patients with cardiovascular diseases. Of these, the random forest algorithm was distinguished by its superior predictive accuracy.

In another innovative approach, researchers [3] employed a support vector machine trained with data from oral glucose tolerance tests alongside demographic and clinical information to forecast the onset of type-2 diabetes. This study underscored the necessity of extensive clinical data for refining model performance.

Explorations into specialized applications of machine learning have also yielded promising results. A recent study [4] applied deep learning techniques to optical consistent radiology to classify conditions linked to diabetes, such as polyene chromophore disorders. Tested on SERI-CHUK and A2A SD-OCT datasets, this model effectively differentiated among various diabetes-related ocular conditions with high accuracy.

Moreover, the use of machine learning extends into predictive models for other severe health risks like heart failure (HF). Several studies have utilized survival analysis to predict HF risks in patients with diverse medical backgrounds. One such study [5] used the R package "pec" to

compare the performance of a random forest model against a Cox regression model, achieving replicable predictions using publicly available data. Another [6] implemented random survival forest analysis, integrating a wide range of clinical variables to model HF hospitalization risks among diabetic patients.

The utilization of advanced machine learning techniques was further exemplified by Mamun and Farjana [7], who compared various algorithms to predict patient survival in HF scenarios using data from the UCI heart failure dataset. Their findings highlighted the LightGBM algorithm as particularly effective, achieving an accuracy of 85% and an AUC of 93%.

This review indicates a dynamic and evolving landscape in the use of machine learning for medical prognosis. The integration of diverse algorithms not only enhances predictive accuracy but also facilitates a deeper understanding of the intricate relationships between diabetes, heart failure, and other comorbid conditions. The ability to accurately predict these outcomes is crucial for timely and effective patient management and treatment planning.

3 Research Methodology

In this part, we mainly focus on data analysis, baseline model preparation and model optimization.

3.1 Data Cleaning

The dataset is formatted in CSV, comprising 16 attributes and one dependent variable. The attributes encapsulate various health indicators including age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. The dependent variable represents the diabetic condition of the patients, indicating whether they are diagnosed with diabetes.

For analytical purposes, it is imperative to convert all categorical attributes into binary numerical values. Specifically, attributes with responses "Yes" are encoded as 1, while those with "No" are encoded as 0. Similarly, for the target variable, "Positive" responses are converted to 1 and "Negative" to 0. The attribute 'age' exhibits a broader range compared to other variables, which could disproportionately influence the computational outcomes. To mitigate this, normalization of the 'age' attribute is performed by dividing each value by its maximum, ensuring that this feature aligns with the binary nature of the other variables and does not skew the analytical results.

Subsequently, we assess the dataset for missing values and discover that all entries are complete without any omissions. We also evaluate the balance of the dataset. The class labeled as "Positive" comprises 320 samples, whereas the "Negative" class contains 200 samples, resulting in a ratio of 1.6:1. This indicates a moderate imbalance, which is not considered extreme. However, to enhance the robustness of the model, particularly for training purposes, a data augmentation strategy is employed. We randomly select a number of samples from the "Negative" class equivalent to the discrepancy between the "Positive" and "Negative" classes. These selected samples are then duplicated, but with a modification to the 'age' attribute, where each age value is randomly

adjusted by 1-3 years. This adjustment occurs prior to the standardization process. Given that the remaining attributes are binary, altering them could potentially lead to a shift in the data distribution during subsequent machine learning processes. Therefore, we restrict modifications solely to the 'age' attribute, as this is expected to have a minimal impact on the original distribution of the data following standardization. This method ensures that our data augmentation does not distort the fundamental characteristics of the dataset.

It is evident that various attributes may exert differing influences on the likelihood of diabetes onset among patients. To accurately ascertain the distinct impact of each feature, it is essential to segregate the dataset into positive and negative samples. Subsequently, employing a radar chart enables the visualization of attribute prevalence ratios within both the positive and negative classes. This graphical representation highlights the disparities among attributes, thereby elucidating their relative importance in contributing to diabetes. Moreover, this analytical approach allows for the identification and exclusion of non-critical features, thereby minimizing the potential for misinformation. By filtering out less significant attributes, the analysis can focus more precisely on those factors that genuinely affect diabetes outcomes, enhancing the overall accuracy and reliability of the predictive model.

The radar map effectively illustrates the contribution of each attribute to the incidence of diabetes. To further delve into the interrelationships among the attributes, a heatmap is utilized, which facilitates a visual exploration of the correlations within the dataset. For the 16 attributes and one target variable, we employ the Pearson correlation coefficient to quantify the degree of linear relationships between variables.

We establish a threshold for the correlation coefficient at 0.4. Any correlation pair displaying an absolute value below this threshold is subsequently masked in the heatmap. This approach is adopted to exclude low-correlation pairs, thereby focusing on more significant relationships that potentially have a greater impact on diabetes outcomes. This selective visualization helps to streamline the analysis, highlighting only the most pertinent and influential correlations, and aids in the identification of potential drivers of diabetes within the dataset.

3.2 Model Preparation

In this section, we prepare several machine learning models to explore the connection between attributes and diabetes.

3.2.1 Data Preparation

Given the dataset's composition of 320 positive samples and 200 negative samples, it is imperative to maintain an equivalent ratio in both the training and testing sets to ensure a representative distribution of data. Consequently, we separate the positive and negative samples and allocate 30% of each category to the testing set. To address the imbalance between the two classes, we employ the data augmentation strategy detailed in section 3.1, aimed at equalizing the number of positive and negative samples in the dataset.

In order to identify attributes that may carry non-essential information, we implement a Principal Component Analysis (PCA) decomposition technique. The dimensions for PCA are configured to vary from 1 to 16, allowing for an exploration of dimensionality reduction across the entire set of attributes. The proportion of total variance explained by each principal component is plotted along the y-axis, serving as a metric to evaluate the effectiveness of the decomposition.

By examining the trend in the cumulative explained variance, we can ascertain the point at which additional dimensions cease to contribute significantly to the data's variance, indicating redundancy or non-essential information. This analysis helps in determining whether dimension reduction through PCA is warranted, potentially enhancing the performance of classification models by filtering out unimportant attributes that could otherwise obscure or distort significant patterns in the data.

3.2.2 Baseline Model and other Models

The baseline model serves as an initial reference point in the process of model development for classifying the target variable. Although it might not achieve high accuracy, it provides a foundational benchmark against which more refined models can be evaluated. In this context, we utilize a logistic regression model as our baseline classification approach. This model employs the logistic

function to estimate the probability of the target variable belonging to a specific class. Outcomes with a predicted probability above 0.5 are classified as positive, whereas those below this threshold are classified as negative. We implement the L2 regularization method as the penalty criterion to control for model complexity and prevent over fitting.

For the development of more sophisticated models, we prioritize obtaining the most robust versions through meticulous validation and optimization techniques. Specifically, we employ a 10-fold cross-validation strategy, which provides a reliable estimate of the model's performance by averaging results across ten different partitions of the dataset. This method ensures that our evaluation is both thorough and resistant to the peculiarities of any particular data split.

Additionally, to identify the optimal configuration of model parameters, we utilize a grid search technique. This approach involves exhaustively exploring all plausible combinations of parameters specified in a predefined grid. By evaluating the performance of the model across these combinations, we can pinpoint the set of parameters that yields the best overall results.

- We start by testing a Support Vector Machine (SVM) model with two main parameters: the kernel type and the regularization coefficient. The kernel options include linear, rbf, poly, and Sigmoid. We vary the regularization coefficient from 0.1 to 3, using 100 equal steps for initial broad testing. Based on the performance curve from this initial range, we refine our search by narrowing the range and reducing the interval size. We then employ a grid search to find the optimal combination of these parameters. Finally, we evaluate the model's performance and robustness using 10-fold cross-validation, ensuring we validate across different data subsets and reduce over-fitting risks.

- We next evaluate a decision tree model, favored for its efficacy with datasets that have a large number of attributes and its strong fitting capabilities. However, a significant limitation of decision trees is their propensity to over fit the data. The parameters we manipulate are criterion, maximum depth, minimum samples split and minimum samples leaf. Criterion is to measure the quality of a split, which can be chosen from Gini coefficient, entropy coefficient and log loss coefficient. Maximum depth controls the depth of the tree. Too deep of a tree has good performance on training set but can easily over-fit on testing set. Minimum samples split is the minimum number of samples requited to split an internal node. This parameter

ensures that samples would not be split with slight difference, which can filter the influence of outliers and noise. Minimum samples leaf is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least the value of minimum samples leaf training samples in each of the left and right branches. This can have the effect of smoothing the model. We use grid search to find the best combination and use cross validation to evaluate the model.

- The two models mentioned above are both simple straight-forward ones. The third model we examine here is random forest. It uses ensemble method, which combines the predictions of several base estimators built with a given learning algorithm in order to improve generalization and robustness over a single estimator. Random forest uses a diverse set of classifiers created by introducing randomness in the classifier construction, in this circumstance, which is decision tree classifier. They are then fitted on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The main parameters we adjust here are n_estimators, criterion, max_depth, min_samples_leaf and min_samples_split. n_estimators controls the number of trees in the forest, which is the number of basic classifier (decision tree). This can strongly affect the performance of the model since model basic estimators means better fitting and lower over-fitting. The rest parameters are the same as decision tree. They control the basic estimators' arguments. We use grid search to find the best combination and use cross validation to evaluate the model.

- The last model we choose is neural networks. As traditional machine learning algorithms have limitations when fed with enough data, which means their performance will not continue to increase even if increasing the training data. On the other hand, the performance of neural networks is almost better than any other machine learning algorithms. This is because neural networks present the behavior of large amount of neurons instead of simple sum of individual neuron. Therefore, the system can perform complex non-linear dynamics characteristics, which makes it easy to fit all kinds of data. In this project, we use feed-forward fully connected neural network since there is no image pattern and each attribute is equally important. We deploy a network with 4 layers: 16-32-64-128-1. They are connect with ReLU layer and end with a Sigmoid function to map the result to a binary value. We use 10 fold

cross validation to evaluate the model.

4 Results

In this part, we review the methods used and present the results with detailed analysis.

4.1 Data Cleaning

Following the pre-processing steps, the analysis of the dataset reveals a standard deviation of 0.135 for the age attribute, indicating a narrower variation compared to the other features, whose standard deviations range between 0.41 and 0.51. This suggests that the dataset predominantly represents a middle-aged population and confirms that the feature distributions are relatively balanced. Additionally, the absence of missing values in the dataset eliminates the need for any data imputation processes. This level of data completeness and uniformity enhances the reliability of subsequent analyses and modeling efforts.

In the dataset, the division between patients labeled as positive (320) and negative (200) was analyzed to assess the distribution of various attributes across these classes, as depicted in the referenced radar map (Fig. 4.1). The visualization employs two distinct color schemes: an orange curve representing the positive class and a blue curve for the negative class. The radar map effectively highlights the disparities between the two classes for each attribute. Notably, a greater divergence between the classes in a specific attribute suggests a stronger association of that attribute with the likelihood of diabetes. According to our analysis, attributes such as polyuria, polydipsia, sudden weight loss, irritability, partial paresis, and polyphagia exhibit a more pronounced correlation with the presence of diabetes. These findings indicate that these symptoms are significantly linked to the disease. Conversely, attributes like gender and alopecia display a higher prevalence in the non-diabetic group. This observation suggests that these factors, while present among patients, do not play a decisive role in determining the presence of diabetes. Such insights are crucial for refining diagnostic criteria and improving targeted interventions for diabetes.

To further elucidate the relationships among the attributes and their connection to diabetes, we

employed a heatmap visualization (Fig. 4.2). By setting a correlation threshold of 0.4, we aimed to focus on the most significant relationships and exclude those with low correlation, ensuring clarity in our analysis.

The heatmap analysis revealed that attributes such as polyuria and polydipsia exhibit strong correlations with diabetes, corroborating their significant roles as identified in the radar map analysis. These findings suggest that these symptoms frequently co-occur and are key indicators of the disease. On the other hand, sudden weight loss and partial paresis, while associated with diabetes, show weaker connections. This indicates that while they are relevant, their predictive power might be less pronounced compared to polyuria and polydipsia.

It also highlighted a negative correlation between gender and diabetes, suggesting that females are more likely to develop diabetes compared to males. This gender-specific insight aligns with the broader epidemiological data and could be crucial for targeted health interventions and further studies.

Additionally, the strong association between polyuria and polydipsia, indicated both by their high correlation in the heatmap and their highlighted importance in the radar map, suggests that these symptoms often co-manifest in diabetic patients. This relationship may provide a useful diagnostic criterion, as the co-presence of these symptoms could be a strong predictor of diabetes.

4.2 Models Development

In this section, we mainly focus on different models to fit the dataset.

4.2.1 Data Preparation

The reserved information ratio curve from PCA analysis (Fig. 4.3) indicates that a minimum of 12 features are required to retain 90% of the original dataset's information, suggesting that only 4 features are compressible. Given that our dataset is not particularly large, we have decided to forgo the use of PCA decomposition in subsequent analyses. This decision is based on the limited reduction in dimensionality offered and the potential loss of critical information, which may not justify the complexity reduction in this context.

Figure 4.1: Radar map.

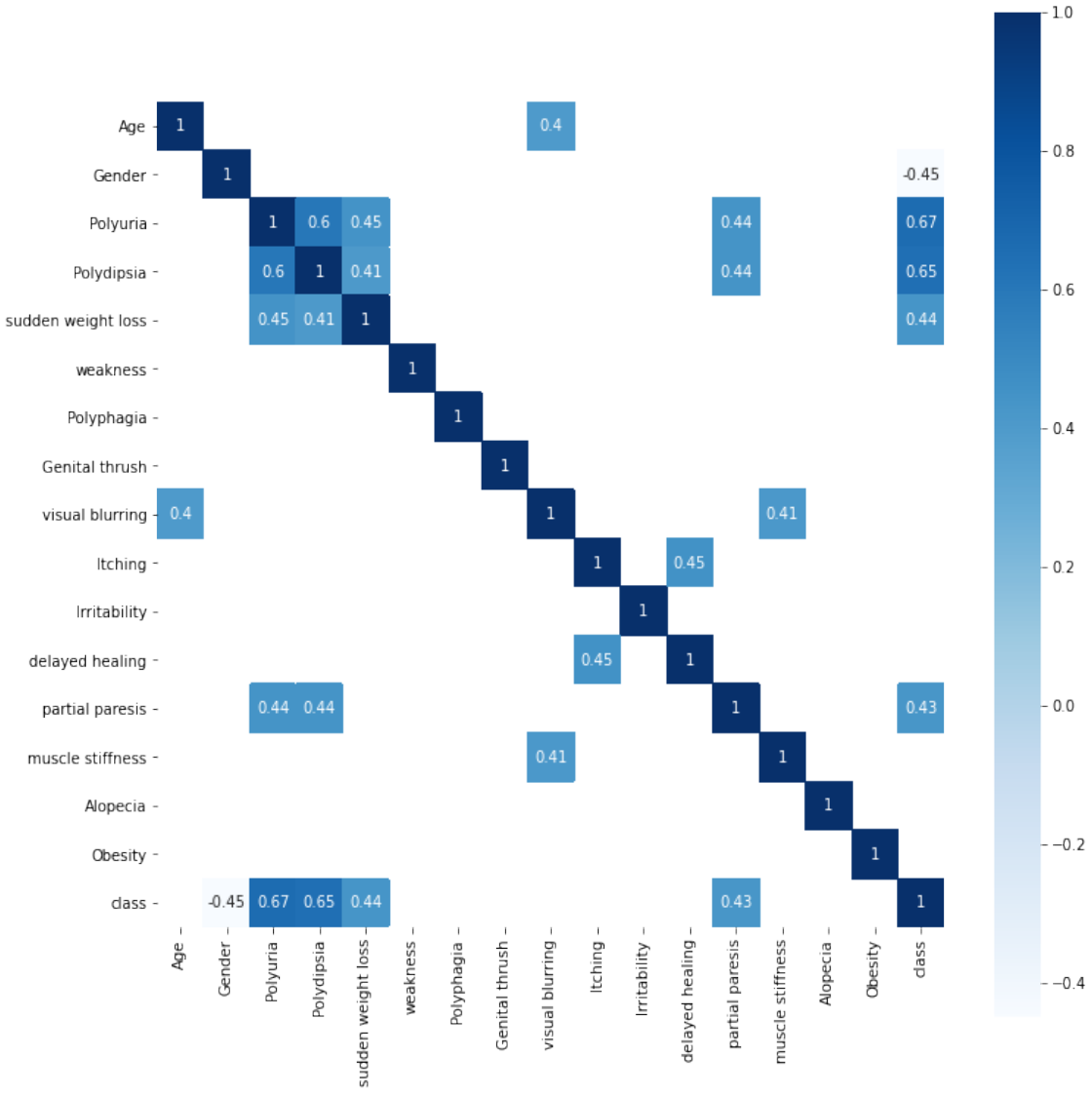

Figure 4.2: Heat map.

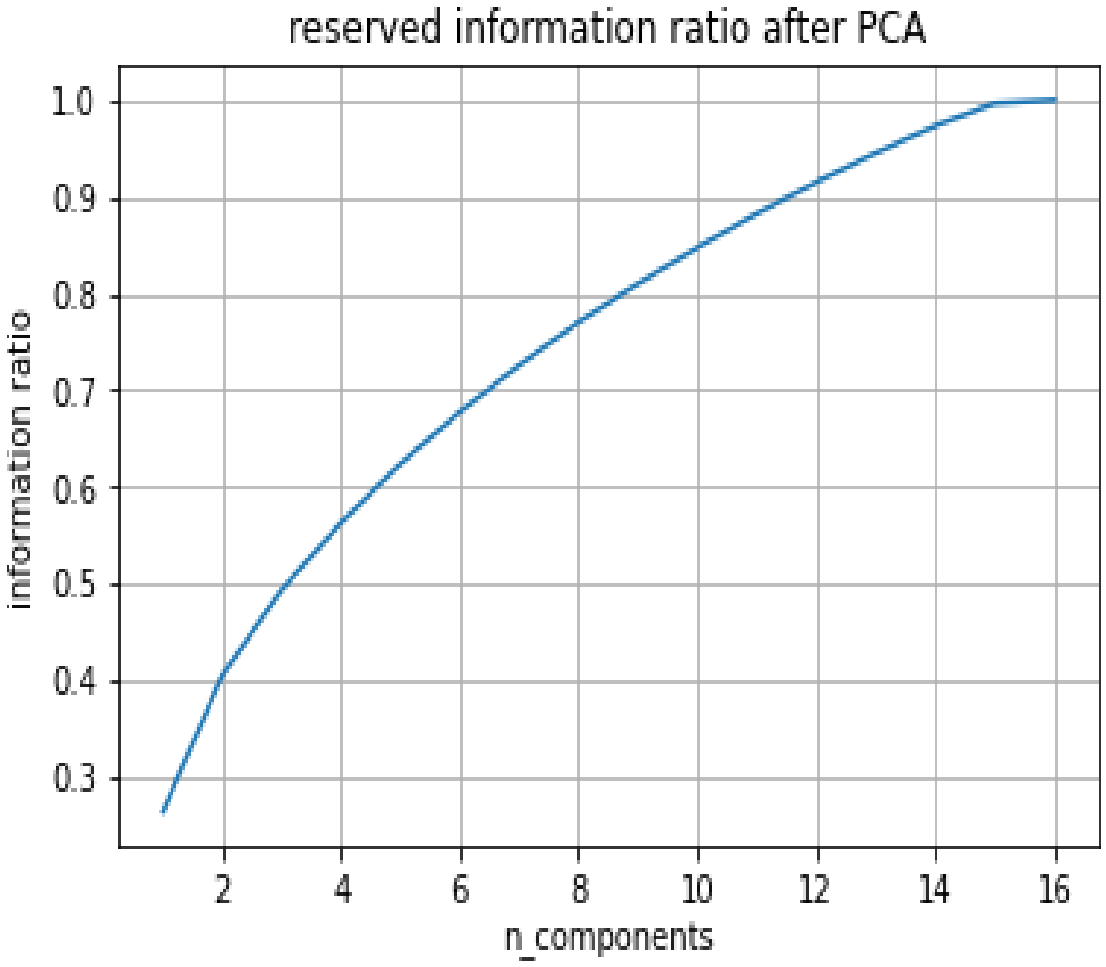

Figure 4.3: PCA.

baseline	training set	testing set
precision	91.25	93.68
recall	91.35	92.71
F1	91.04	93.19

Table 4.1: baseline model evaluation

SVM	training set	testing set
precision	97.79	97.89
recall	98.66	96.88
F1	98.22	97.38

Table 4.2: SVM model evaluation

4.2.2 Baseline Model and other Models

In order to prevent the bias of single criterion, we use precision, recall and F1 scores to evaluate models.

For the baseline model, we choose logistic regression. We present the performance both on training set and testing set to avoid over-fitting (Table: 4.1).

In the SVM model, we varied the kernel type and the regularization coefficient, employing a grid search approach to determine the optimal parameter combination. The best-performing parameters identified were: kernel as rbf and regularization coefficient at 1.184 (Table: 4.2). The evaluation metrics—precision, recall, and F1 score—demonstrate similar performance between the training set and the testing set, indicating that the model does not suffer from over-fitting. Both precision and recall are high, and the F1 score aligns closely with these metrics, suggesting that the model effectively handles the potential data imbalance. Notably, the errors in the testing set show a significant decrease: 66.61% improvement in precision, 57.20% in recall, and 61.53% in F1 score. These improvements confirm that the SVM model is proficient in classifying diabetes, making it a robust tool for this task.

In the decision tree model, we optimized several parameters, including the criterion, maximum depth, minimum samples required to split an internal node, and minimum samples required at a

Decision tree	training set	testing set
precision	94.86	97.87
recall	95.67	95.83
F1	95.13	96.84

Table 4.3: Decision tree model evaluation

leaf node. The optimal settings determined were: criterion as gini, maximum depth of 7, minimum samples split of 1, and minimum samples leaf of 3 (Table: 4.3). This configuration shows no significant signs of over-fitting, indicating a stable model performance. However, when compared to the SVM model, the decision tree model exhibits weaker performance metrics. Specifically, the improvements in the testing set, when compared to the baseline model, are as follows: precision improved by 66.30%, recall by 42.80%, and F1 score by 53.60%. Although these metrics show considerable improvements over the baseline, they are less impressive than those achieved by the SVM model, highlighting a relative deficiency in handling the complexity or variability within the dataset.

To enhance the performance of the decision tree model, we employed an ensemble method, specifically random forest, which leverages the strengths of multiple decision trees to improve generalization and reduce over-fitting. Based on the optimal parameters identified for the decision tree model, we established a suitable range for tuning the random forest model, thereby narrowing the search space for finding the most effective settings. In the random forest configuration, we primarily adjusted the n_estimators parameter, which represents the number of trees in the forest, with each tree functioning as an independent classifier. The optimal parameter combination identified was: n_estimators=160, criterion=gini, max_depth=8, min_samples_split=1, and min_samples_leaf=2 (Table: 4.4). Notably, the performance on the testing set surpassed that on the training set, suggesting an exceptionally well-generalized model. This unusual result may indicate that the available data quantity might be insufficient, potentially leading to high variance in model performance across different data segments. The improvements observed in the testing set were substantial: precision increased by 100.00%, recall by 71.47%, and F1 score by 84.58% compared to the baseline model. These metrics significantly outperformed those of the single decision tree

Random forest	training set	testing set
precision	95.94	100.00
recall	96.07	97.92
F1	95.94	98.95

Table 4.4: Random forest model evaluation

Neural network	training set	testing set
precision	97.50	99.33
recall	98.34	99.58
F1	99.17	99.50

Table 4.5: Neural network model evaluation

model, highlighting the effectiveness of the ensemble approach. The random forest model not only enhanced performance metrics on both training and testing sets but also demonstrated the practical viability of using ensemble methods to boost accuracy and reliability in predictive modeling. This reinforces the value of ensemble strategies in complex predictive tasks, particularly when dealing with heterogeneous or limited datasets.

The last model we explore is neural network. We use fully connect feed forward network as each attribute can have same impact on the prediction. With this simple configuration of the network (16-32-64-128-1), we easily derive a good performance after only 20 epochs (Table: 4.5). Errors in testing set decrease by 89.40% in precision, 94.24% in recall and 92.66% in F1 compared to baseline model.

To gauge the enhancement of models over the baseline, we computed the average reduction in errors for precision, recall, and F1 score, detailed in (Table: 4.6). A higher average error reduction indicates superior performance. The SVM showed significant improvements over the decision tree but is susceptible to unbalanced data, impacting its efficacy. The random forest, using 160 base estimators, outperforms the decision tree by enhancing generalization and reducing over-fitting risks through its ensemble method. The neural network, despite the limited dataset size, delivered the best performance due to its deep architecture and ability to model complex patterns effectively,

	SVM	decision tree	random forest	neural network
error improvement	61.78	54.23	85.35	92.10

Table 4.6: Average error improvement

achieving the highest average error reduction among the tested models. This confirms its robustness in extracting meaningful insights from the data.

$$average\ error\ improvement = \frac{precision + recall + F1}{3}$$

5 Discussion and Conclusion

The conclusion needs to provide

- A short summary (What has been done and what are the main results)
- Limitations of your work, where applicable.
- Discussion of your work in the bigger picture (How does this contribute to the research field?)
- Future work (What could be next steps in this work?). Remember to keep future work realistic. A good approach is to discuss what the next progression of this project would be, and to justify why this would be interesting.

You will find it easier to write your conclusion if you copy-and-paste your *Aims, Objectives*, and any research questions or hypotheses you stated. You can then discuss each of these explicitly in turn, and how you were able to answer them or complete them successfully. When things have not gone as well as you would have hoped, demonstrate your critical thinking and reasoning to analyse the short-comings of your project - to demonstrate that you understand the underlying causes and that you could conduct good futurework from this learning experience.

References

[1] A. Yılmaz, Prediction of type 2 diabetes mellitus using feature selection-based machine learning algorithms przewidywanie cukrzycy typu 2 z wykorzystaniem algorytmów uczenia maszynowego opartych na selekcji cech,

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Machine learning and data mining methods in diabetes research, *Computational and structural biotechnology journal*, vol. 15 2017, pp. 104–116, 2017.

[3] H. T. Abbas, L. Alic, M. Erraguntla, *et al.*, Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test, *Plos one*, vol. 14, no. 12 2019, e0219636, 2019.

[4] O. Perdomo, H. Rios, F. J. Rodríguez, *et al.*, Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography, *Computer methods and programs in biomedicine*, vol. 178 2019, pp. 181–189, 2019.

[5] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, Evaluating random forests for survival analysis using prediction error curves, *Journal of statistical software*, vol. 50, no. 11 2012, p. 1, 2012.

[6] M. W. Segar, M. Vaduganathan, D. K. McGuire, M. Basit, and A. Pandey, Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: The watch-dm risk score. diabetes care 2019; 42: 2298-2306, *Diabetes care*, vol. 43, no. 2 2020, E26–E27, 2020.

[7] M. Mamun, A. Farjana, M. Al Mamun, M. S. Ahammed, and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: Am i safe from heart failure?" In *2022 IEEE world AI IoT congress (AIIoT)*, IEEE, 2022, pp. 194–200.