JULY 2023

ENHANCING CLINICAL GUIDELINES THROUGH THE INTEGRATION OF DECISION TREES FOR THE DIAGNOSIS OF DM-2

PROJECT AI IN HEALTH

By:

Kirandeep Gill (2695941)
Márk Bartos (2724195)

Enhancing Clinical Guidelines through the Integration of Decision Trees for the Diagnosis of DM-2

Group 6: Kirandeep Gill (2695941) and Márk Bartos (2724195)

Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands

Abstract

Introduction Diabetes mellitus (DM) is a common chronic disease, approximately 463 million people are living with diabetes globally. The prevalence of diabetes is about 1.1 million in the Netherlands, of which 90% are patients with diabetes mellitus type 2 (DM-2). Tools that could be crucial for optimizing the diagnosis are decision tree learning and clinical guidelines. The aim of this paper is to explore whether decision trees provide similar predictions compared to established clinical guidelines by NHG for DM-2 diagnosis in adults.

Experimental set-up Two different datasets have been used from India and Germany. One focuses on lifestyle, whereas the second primarily centres around clinical measurements. The pre-processing included imputing missing values, One-Hot Encoding and removing the 'Age attribute', as only adults were relevant to this study. Next, a SKLearn's Decision tree Classifier with K-Fold Cross Validation was employed.

Results In Decision Tree 1 regular medicine and high blood pressure were found to be the most contributing factor to DM-2, with a slight contribution of family diabetes to the onset of DM-2. Glucose levels and BMI were found to be the most significant factor in raising the risk of diabetes in Decision Tree 2. The features of sleep and prediabetes remained missing from the clinical guidelines during the analysis of the results, and relevant literature adds to this.

Conclusion In conclusion, the decision trees and clinical guidelines shared many similarities in BMI, glucose, insulin, pregnancies, and blood pressure, but differ in sleep and prediabetes, and completely overlook symptoms such as thirst, weight loss, neurogenic pains and sensory disorders.

Discussion Future work in this field could include creating more detailed models and focusing on implementing the results of decision tree models into automated healthcare systems, such as HiX, used in The Netherlands.

Keywords: *Machine Learning, Decision Tree Learning, Guidelines, Diabetes Type-2, Health Care, Model, Diagnosis, SciKit Learn*

1 Introduction

In the field of healthcare, clinical guidelines function as descriptions of actions that need to be carried out for the diagnosis or management of patients under particular clinical conditions, such as diabetes mellitus (DM). These guidelines are based on available literature and databases and therefore provide valid and mostly up-to-date recommendations. The clinical guidelines can also function as a tool to compare them to the output generated by decision trees. In this paper, two decision trees are generated based on existing datasets and are compared to established clinical guidelines from Het Nederlands Huisartsen Genootschap (NHG)[1] using scikit-learn (SKLearn). This paper aims to explore whether decision trees provide similar predictions compared to established clinical guidelines by NHG for type 2 diabetes diagnosis in adults.

1.1 Diabetes in numbers: prevalence, incidence and costs

Diabetes mellitus (DM) is a common chronic disease worldwide. According to the Diabetes Federation Diabetes Atlas (9th Edition), approximately 463 million people are living with diabetes, and this number continues to expand[2]. The Netherlands is also grappling with diabetes among its inhabitants. The prevalence of diabetes was about 1.1 million in the Netherlands in 2021, of which 618 300 are men and 538 500 are women[3]. Roughly 90% (1,028,700) of all diabetes patients have type 2 diabetes mellitus (DM-2), of which 53% (546,600) are men and 47% (482,100) are women[3]. Diabetes is more frequent in men than women in almost all age groups[4]. When looking at the demographic development of the Dutch population, the incidence of DM-2 is expected to fall by around 117 000 to 131 000 new patients per year until 2040[3]. In 2019, the estimated healthcare expenditure for DM-1 and DM-2 was 1.308 million euros, accounting for 1.4% of the total healthcare expenditure in the Netherlands[5].

1.2 Diabetes mellitus

Diabetes mellitus is a medical term used to describe a group of metabolic diseases characterized by hyperglycemia, which is an elevation of blood glucose levels. Chronic hyperglycemia can lead to several problems, such as the malfunctioning of various tissues and organs including the eyes, kidneys, nerves, blood vessels and heart. DM is most predominantly associated with defects in insulin secretion or insulin action. DM can be divided into four types: diabetes mellitus type 1 (DM-1), diabetes mellitus type 2 (DM-2), other specific types, and diabetes gravidarum (gestational diabetes).[6-8] Type 2 diabetes, also referred to as non-insulin-dependent diabetes, is a medical condition concerning a reduction in insulin action with progressive loss of functioning of beta cells, often accompanied initially by relative insulin deficiency[8]. The risk of developing type 2 rises with advancing age, obesity, and insufficient physical activity and is frequently linked to a significant genetic predisposition[8].

1.3 Machine Learning and clinical guidelines

Machine learning (ML) has promoted itself as a powerful tool in healthcare, changing the way medical data is used and analysed[9]. One specific area where machine learning has been utilized is in diagnosing, treating, and managing diabetes mellitus. Decision tree learning, frequently used in machine learning, is widely used in machine learning for the purpose of classification and prediction. It functions as a framework for analysing patient data and comparing it with established clinical guidelines[9,10]. One example from previous research would be the clinical practice guideline of Singapore's Ministry of Health on DM[11] (see Figure 1). This flowchart is created to function as diagnostic recommendations in clinical practice when fasting is complemented with certain target values or parameters, such as blood glucose level (input) and a diagnosis of DM (output). Another example is the study conducted by Alghamdi et al. (2017) on predicting DM with the aid of ML[12]. Decision tree learning played a significant role when identifying and signalling important predictors for diabetes, along with providing interpretable rules for the diagnosis process of DM.

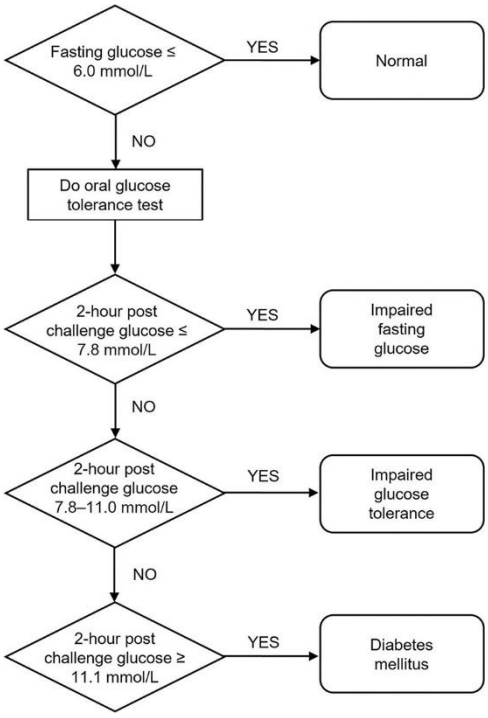

Figure 1. A diagnostic flowchart for diabetes mellitus during fasting glucose is < 7.0 mmol/L or casual or 2-hour post-challenge glucose is < 11.1 mmol/L.

2 Experimental Set-up

In preparation for answering the research question, two different datasets have been collected. "Prediction of Type 2 Diabetes using Machine Learning Classification Methods"[13] (further referenced as Database 1) and "Predicting diabetes diseases using mixed data and supervised machine learning algorithms"[14] (further referenced as Database 2). Both datasets provide a set of patient details and corresponding targets, indicating the presence or absence of diabetes in a patient. These two datasets have been imported into a Juptyer notebook with the help of the Pandas Dataframe library, where a Decision Tree learner using the SKLearn framework is intended.

2.1 Imputation

Imputation[15] is required if missing values are present in a dataset. By employing appropriate techniques, the imputation method estimates the missing values and subsequently fills them in with this estimate. These techniques utilize statistical measures, namely mode and mean. To impute a categorical value, the mode of that feature column has to be used, and to impute a numerical value, the mean of that feature column has to be used.

As missing values were present in both datasets, imputations were used accordingly with the help of Conditional Formatting, =AVERAGE and =IF(A1<>0, A1, "") in Excel. Zero values were considered a mistake, except for pregnancies, as only in that case it is normal to have zero of them. On the contrary, BMI, Glucose, Insulin, Skin Thickness and Age are considered to be missing, if the null value is present. In Dataset 1, missing values for Pdiabetes were imputed with the mode value of "no", and for BMI with the mean value of 26. In Dataset 2, missing values for Glucose have been imputed with the mean of 122, for BloodPressure with 72, for SkinThickness with 30, for Insulin with 154 and for BMI with 32. Both in Dataset 1 and Dataset 2, if the target value was missing for an instance, the whole instance has been removed.

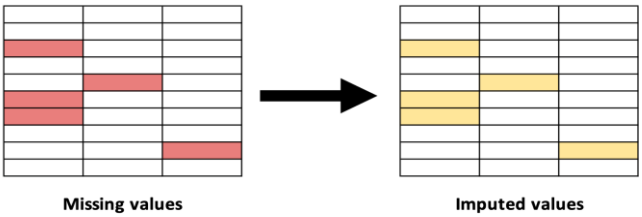

Figure 2. Imputation

2.2 One-Hot Encoding

One-Hot Encoding[16] is a technique, used to transform categorical variables into numerical variables, to make data suitable for machine learning algorithms. It creates new, binary (1 or 0) columns for categorical features, where each column represents if a specific instance belongs to a category, or not. The use of One-Hot Encoding is essential for SKLearn's Decision Tree learner, as it could only work with numerical features. By converting data using One-Hot Encoding instead of numerical encoding, it allows learning algorithms to handle data correctly, minimize bias and effectively learn representative patterns.

Since both datasets contain mixed values - strings for categories and integers for numerical values - One-Hot Encoding had to be used. For Dataset 1 "Age", "Gender", "Family_Diabetes", "highBP", "PhysicallyActive", "Smoking", "Alcohol", "RegularMedicine", "JunkFood", "Stress", "BPLevel", "Pdiabetes" and "UninationFreq" has been nominated as categorical variables, while "BMI",

"Sleep", "SoundSleep" and "Pregnancies" were marked as numerical features. For Dataset 2, only "Outcome" has been marked categorical, all the other values were numerical.

Figure 3. Numerical Encoding

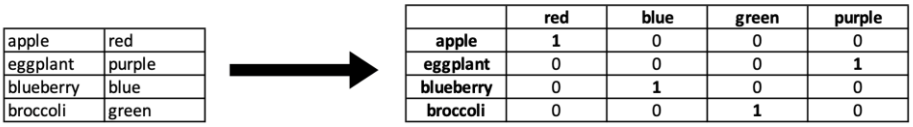

Figure 4. One-Hot Encoding

2.3 Removal of Age

As a last step of preprocessing, the attribute of Age has been removed from both datasets. This is due to the fact that reviewed guidelines focus on the occurrence of diabetes in adults, and both datasets contain instances from patients above the age of 18. No further age grouping is required, as this research does not focus on the additional influence of age, but the other affecting features of Type 2 Diabetes in adults.

2.4 Decision Tree Learner

After pre-processing, the feature-spaces have been nominated as X, and the target values - the occurrence of diabetes -, as Y. The SKLearn Decision Tree Classifier[17] has been trained with a maximum depth of 3, and Entropy has been chosen as a criterion method for determining a split. When deciding on the maximum depth, we have found that 4 creates trees that include too many influencing factors, and anything below Depth 2 is influenced too much by the previous levels. While with a tree of maximum depth 2, not enough information was available to analyse and compare with an existing clinical guideline.

Entropy[18] has been preferred in this project since it provides slightly more nuanced results and uses an information-theoretic measure of uncertainty. Although Entropy is a computationally more expensive algorithm when compared to Gini, such limitations were not present during our work, due to the smaller size of Dataset 1 and Dataset 2 (950 and 2000 instances respectively) and available computing power.

$$H = -\Sigma\, p(x)\, log\, p(x)$$

Figure 5. The formula of Entropy

2.5 Training

The K-Fold Cross Validation[19] technique is used to improve the generalization ability and performance of a machine learning model, by dividing the available dataset into equally-sized K subsets (also known as "folds"). Then the model is trained K-times, each time using a different fold as the validation set to evaluate the model performance. This approach produces better-performing models, prevents under- or overfitting, and makes it possible to train better on smaller datasets.

To achieve the highest performance during training, the K-fold Cross Validation algorithm was implemented to support Decision Tree learning. The use of this method was specifically important, as both Dataset 1 and Dataset 2 were small in size compared to industry standards. Since, it is best practice to choose a higher K-value for a smaller dataset, for Dataset 1, with 950 instances, a K-value of 10 has been chosen, while for Dataset 2, which is double the size with 2000 instances, a K-value of 5 has been chosen. Table 1 illustrates the accuracy scores of each fold present during training and their averages. Decision Tree 1 got a slightly higher average during K-Fold Cross Validation.

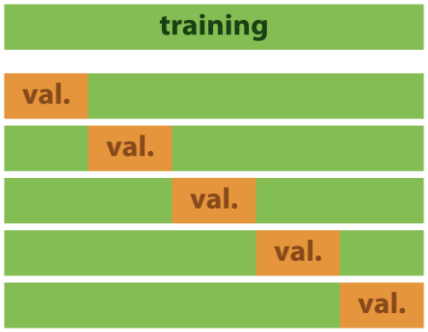

Figure 6. K-Fold Cross-Validation[20]

Table 1. Accuracy Scores

K^{th} Fold	Decision Tree 1	Decision Tree 2
1	0.85263	0.7525
2	0.83158	0.7900
3	0.84211	0.7275
4	0.86316	0.7500
5	0.85263	0.7225
6	0.80000	-
7	0.84211	-
8	0.85263	-
9	0.85263	-
10	0.78947	-
Average	0.8379	0.7485

2.6 Visualization

After training with K-Fold Cross Validation, the Decision Trees were visualized using Graphviz, to provide a clear and interpretable visual overlook on the trained decision tree algorithm. These graphs offer detailed insight into the sample sizes, the entropy values and the predicted class.

2.7 Summary of Pre-Processing

In conclusion, two datasets, Dataset 1 and Dataset 2 have been collected and preprocessed with several methods. These include imputing to fill in missing values with mean and mode, One-Hot Encoding to transform categorical variables to numerical columns and the removal of the Age feature. The SKLearn Decision Tree Classifier was trained using Entropy, the criterion of maximum depth of 3 and using K-Fold Cross Validation. The results suggest that the Decision Trees were able to highlight the features most important in making a Type 2 Diabetes diagnosis.

3 Results

Decision Trees are Machine Learning algorithms used for classification and regression tasks. Each of the nodes in the tree-like structure corresponds to a feature and their outcome in terms of the target value. They are easy to visualize and have great readability[21].Therefore, in areas, such as health care, where understanding the insights is crucial, decision trees offer an intuitive and transparent solution to the decision-making process[22]. One of the most famous decision tree algorithms includes ID3, Gini and C4.5. Entropy is an information-gain-based algorithm, calculating the information gain of each node, and choosing the highest one to split the data. Entropy minimizes uncertainty and creates more nuanced results. Although it's computationally expensive, due to the high amount of log calculations, it has been chosen as our preferred algorithm, as nuanced results were more important than the fastness of the calculation.

3.1 Decision Tree 1

Decision Tree 1 is based on Dataset 1, which is focused on lifestyle features and their influence on Diabetes Type 2. The list of features includes Gender, Family Diabetes, High Blood Pressure, Physical Activity, Smoking, Alcohol, Taking Medicine Regularly, Consumption of Junk Food, Stress, the presence of Prediabetes, Urination Frequency, BMI, Hours of Sleep, Hours of Sound Sleep, and the Number of Pregnancies.

The first split at Depth 0 is at taking regular medicine or not (Regular Medicine_yes <= 0.5), with the highest information gain of 0.853. This shows that the highest difference in having diabetes or not starts with taking regular medicine. Out of 300 patients, who do take regular medicine, 194 do and 106 do not have diabetes, while out of the 555 patients who do not take medicine regularly 44 do and 511 do not have diabetes. This indicates that not taking regular medicine lowers the chances of developing Type 2 Diabetes by almost 11 times, however, taking regular medicine does not increase the chances of Type 2 Diabetes.

The biggest split at Depth 1 is having High Blood Pressure (BPLevel_high <= 0.5). This split has an entropy of 0.937. Out of the 137 patients who have a high blood pressure level, 111 have and 26 do not have diabetes, while out of the 163 patients who do not have a high diabetes level indicated, 83 have and 80 do not have diabetes. This indicates that having a high blood pressure level definitely increases the risk of developing diabetes, while not having a high blood pressure does not necessarily decrease the chances of developing it.

The second-biggest split at Depth 1 is SoundSleep being less than 5.5 hours, with an entropy of 0.4. Out of the 248 patients who have less than 5.5 hours of sound sleep 5 have and 243 do not have diabetes, and out of the 307 patients who have more than 5.5 hours of sound sleep 39 have and 268 do not have diabetes. Since only 5 people have diabetes out of the ones who have less than 5.5 hours of sound sleep, there is not enough data to make a representative conclusion here.

A significant detail presents in Depth 2 including the history of diabetes in the family (Family_Diabetes_yes <= 0.5) with an entropy of 0.701. Out of the 79 people who have diabetes running in their family, 76 have diabetes, and only 3 do not have it. Out of the 58 people who do not have a history of diabetes in their family, 35 are diabetic and 23 are not. Therefore, having family diabetes significantly increases the chance of having diabetes.

Also at Depth 2, not having Prediabetes (Pdiabetic_no <= 0.5) also seems to lower the chances of developing diabetes, however on the contrary, there is not enough data (only 7 samples) to back up this claim. Physical activity at Depth 2 is also hard to investigate, as other influencing factors at Depth 1 and Depth 0 interfere with the results.

Other remarks include those attributes that do not show up at the first 3 levels of Depth. These are medically also considered to be influential, however, their influence might not be as strong as the ones present in the first 3 levels. Further research could investigate the relevance of these factors. These attributes include Gender, Smoking, Alcohol, Consumption of Junk Food, Stress, Urination Frequency, BMI, Hours of Sleep and the Number of Pregnancies.

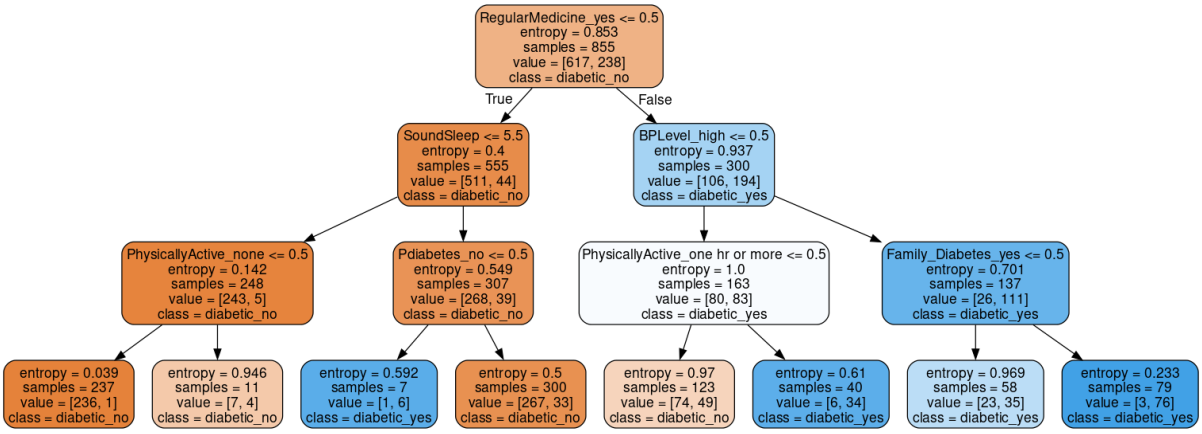

Figure 7. Decision Tree Number 1 (Based on Dataset 1)

3.2 Decision Tree 2

Decision Tree 2 is based on Dataset 2, which focuses more on medical measurements rather than lifestyle choices. This dataset as opposed to the previous one also includes numerical features only, except for the target value. Therefore, the optimal split is not only calculated for features, but for what value that feature produces the highest entropy.

The first split at Depth 0 is determined at Glucose levels being below or equal to 123.5, with an entropy of 0.931. Out of the 912 patients who have it below or equal to 123.5, 154 are diabetic and 758 are not. Out of the 688 patients whose Glucose level is above 123.5, 287 do not have diabetes, while 287 do have it. Therefore, it is concluded that a 123.5 or higher level of glucose can lead to Diabetes.

At Depth 1, the highest split with an entropy of 0.98 is BMI being less or equal to 29.95. If it is satisfied with the precondition of Depth 0, so if the patient has a BMI index of 29.95 or lower and their Glucose levels are higher than 123.5, then out of these 206 patients, 68 do and 138 do not have diabetes. While if the patient's BMI is higher than 29.95 but the precondition of Depth 0 is still falsely satisfied, then out of those 482 patients, 333 do and 149 do not have diabetes. Therefore, it can be concluded that in the case of raised Glucose levels (> 123.5) and a BMI higher than the obese limit (> 29.95) can contribute to Diabetes, while BMI lower than the obese limit (<= 29.95) can lower the chances of developing it.

At the other node of Depth 1, with an entropy of 0.655 also stands BMI, however, now with a condition of being lower than 26.9, slightly higher than the overweight limit. If it is satisfied with the precondition of Depth 0 being true, so if the patient is below the overweight limit (<= 26.9 BMI) and their Glucose levels are below 123.5, then out of those 248 people, 1 do and 247 do not have diabetes. While if the patient's BMI is higher than 26.9, but their Glucose level is 123.5 or below, then out of those 664 patients, 153 do and 511 do not have diabetes. According to the data,

ones with a BMI lower than 26.9 - with the precondition of Depth 0 being true - have a proportionally lower chance of diabetes, than ones with a BMI higher than 26.9.

While at Depth 2 it is hard to investigate single factors due to the influence of Depth 1 and Depth 0, the number of pregnancies seems to have the biggest difference in proportion distribution. Additionally, the following attributes do not show up in the first 3 levels of Depth: SkinThickness and DiabetesPedigreeFunction.

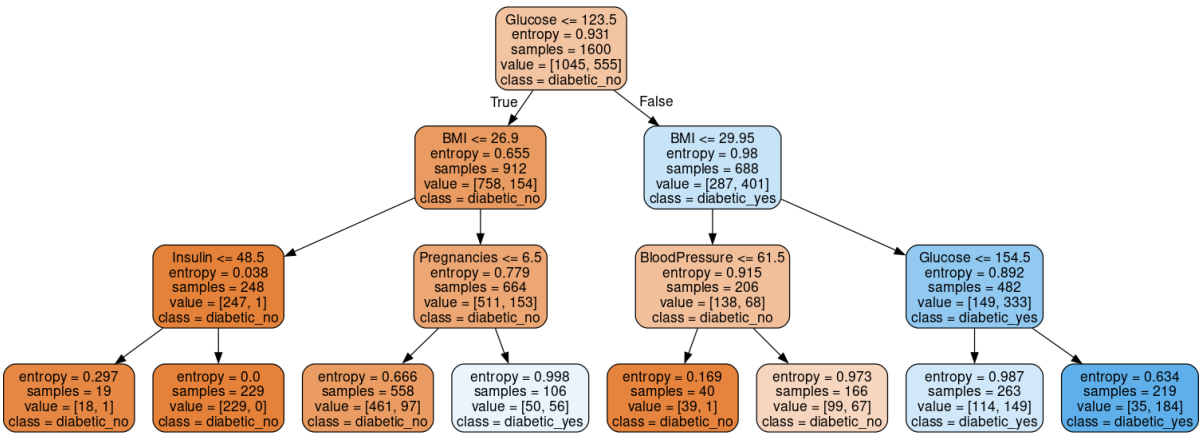

Figure 8. Decision Tree Number 2 (Based on Dataset 2)

4 Analysis of the results: agreements and disagreements with other studies or reviews

During the analysis of the output of the decision trees, several features were identified as key contributors in the diagnosis process of DM-2. First, the focus will be on decision tree number 1 including, "RegularMedicine_yes", "SoundSleep", ''BPLevel_high'', "PhysicallyActive_none", "Pdiabetes_no", ''PhysicallyActive_one hr or more'' and "Family_Diabetes_yes". We observed similarities and differences when comparing these features to the Dutch NHG guidelines[1] on DM-2. The section 'Richtlijnen diagnostiek/Diagnostics guidelines' was mainly applicable to this paper.

To begin with, the feature "SoundSleep" is not specifically mentioned in the Dutch NHG guidelines. No guideline whatsoever is provided by NHG on the importance of sleep regarding the diagnosis of DM-2. In contrast, according to PubMed, the search terms ''diabetes type 2'' ''and'' ''sleep'' have over 1000 hits over the past 5 years exploring this potential relationship and a proportion of them suggest further examination of sleep on DM-2[23,24]. One research article states that there is a clear relationship between sleep and circadian rhythm disruption and DM-2, but only a few clinical guidelines have included this relationship when diagnosing or even treating patients with DM-2[25].

In addition, decision tree number 1 classifies ''Prediabetes'' as a feature in the diagnosis part of DM-2. This is in opposition to the NHG guidelines, in which prediabetes is not covered in the diagnosis process. However, prediabetes is mentioned in the document by NHG, but merely to mention that it exists and to state that approximately 50% of the patients with prediabetes do not necessarily get diagnosed with DM-2. Except for that, prediabetes is excluded from the diagnosis of DM-2 in the NHG guidelines. When looking at the literature, the American Diabetes Association[26] and The American Association of Clinical Endocrinology[27] do recommend

universal screening for prediabetes, suggesting that clinical guidelines should include questions about prediabetes.

On the other hand, during the analysis of the decision trees, the remaining features of decision tree number 1 were present in the NHG guidelines on DM-2. Nevertheless, some guidelines and recommendations were more detailed in the NHG guidelines and not present in the decision tree. For instance, the NHG guidelines include recommendations on the anamnesis of the patient concerning symptoms such as thirst, weight loss, neurogenic pains and sensory disorders and recurrent urinary tract infections, while the decision tree did not cover this. Additionally, the NHG guidelines highlight the significance of risk assessment on cardiovascular health, eye examination, feet examination and dietary modifications, which are, except for cardiovascular health, not mentioned in decision tree number 1.

Second, decision tree number 2 including ''Glucose'', ''BMI'', ''Insulin'', ''Pregnancies'' and ''BloodPressure'' will be analysed. We mostly observed similarities when comparing these features to the Dutch NHG guidelines[1] on DM-2. All the features of decision tree number 2 were mentioned in the NHG guidelines. The feature ''Pregnancies'' is mentioned in the form of gestational diabetes. Attention to gestational diabetes is brought to the screening and diagnosis of diabetes. It is important for clinicians to take gestational diabetes into account when diagnosing, as this requires different follow-up methods[28]. However, this feature is of more relevance in the case of gestational diabetes and not DM-2[8]. Gestational diabetes is discussed in a different guideline of NHG[29]. The remaining features are discussed consistently, as they play a significant role in the diagnosis and treatment of DM-2.

5 Discussion

5.1 Summary of main results

The results from the analysis of Decision Tree 1 and Decision Tree 2 provide us insights into what factors influence the onset of Type 2 Diabetes. Since Decision Tree 1 focuses more on lifestyle, regular medicine and high blood pressure were found to be the most contributing factor to DM-2. However, a history of having diabetes in the family also slightly contributes to the development of the detailed disease. In Decision Tree 2, which concentrates on medical measurements, glucose levels and BMI were found to be the most significant factor in raising the risk of diabetes. The relevance of attributes which did show up in the guideline but not in the used databases should be investigated further. In conclusion, these findings helped us discover influencing factors of DM-2 development and suggest areas for future research.

5.2 Overall completeness and applicability of the findings

The results of this paper are applicable to a certain extent, as the included datasets assessed a varied diabetic and non-diabetic population with only adults, both males and females and a wide range of glucose levels. It should be noted that, however, this paper only focused on patients with type 2 diabetes and the data was collected from hospital settings.

In this research paper, we utilized, as stated before, two datasets in order to investigate whether decision trees provide similar predictions compared to established clinical guidelines by NHG for type 2 diabetes diagnosis in adults. Dataset 1 may include features more applicable to the DM-2 population of India, as this was the country in which the data was collected. The data of dataset 2 was collected from Frankfurt Hospital in Germany and the Pima Indian dataset. However, the NHG guidelines are created by the Dutch and serve as a framework for the diagnosis of individuals

with DM-2 in the Netherlands. Guidelines are often tailored to specific healthcare systems, which can differ between countries. The Netherlands may have different demographic characteristics, cultural factors or healthcare practices and technology compared to the populations represented in both datasets.

Conversely, the diagnosis of DM-2 is in most countries established with a used oral glucose tolerance test (OGTT), fasting blood/plasma glucose, HbA1c, random blood glucose and glycosuria[30]. This implies that most countries diagnose patients with DM-2 with these tests. The screening and management of DM-2 can, however, be different between countries in established guidelines. The utilization of different datasets can broaden the scope of our research. Nonetheless, further validation in these distinct healthcare settings is required to provide more specific decision trees.

5.3 Reflection on the search process of the clinical guidelines

In this paper, we encountered some challenges in finding specific clinical guidelines completely dedicated to the diagnosis of DM-2. Most of the available English clinical guidelines, such as NICE guidelines, focus on the management and treatment of DM-2, with some additional information on diagnostics or screening. The NHG guideline that is used contained more information about diagnostics, but remained incomplete, with also more emphasis on the management of DM-2. This served as a potential limitation to our study, as more specific guidelines on the diagnosis of DM-2 in adults were needed to ensure the accuracy and reliability of our findings. Our search strategy was therefore adapted, and we included several relevant sections from the NHG guideline on DM-2 in adults.

5.4 Reflection on SciKit

SciKit provides a library of highly customizable algorithms, including the Decision Tree Classifier. We were able to fine-tune it to our needs by the change of attributes, including maximum depth and the classification criterion, which was entropy in our case. SKLearn in combination with GraphViz created visualizations that were information-rich, including the criteria, the entropy value, the sample size, the split of values and the class of the nodes. GraphViz also had the opportunity to be customized with class labels, feature labels and colour coding. Altogether, the details we needed to analyse our data were available to us, and we were able to compare the generated decision trees to existing clinical guidelines, and understand why the algorithm decided to make a split and create such a tree. However, due to One-Hot Encoding, SciKit's Decision Tree Learner created split criteria that were hardly human-interpretable. For example, on Decision Tree 1, Depth 0 includes a split "regularmedicine_yes <= 0.5" meaning that if someone does take regular medicine, which comes to a value of 1, it makes the node false (1 not less than or equal to 0.5), which is the opposite truth value of "Does the patient take medicine?" question. In spite of these limitations, SciKit enabled us to understand the correlation between decision trees and clinical guidelines.

5.5 Future work

Future work in this field could include focusing on creating more detailed models and implementing the results of decision tree models into automated healthcare systems, such as HiX, used in The Netherlands. Integrating such classification models, trained on more in-depth data and medical approval, could potentially result in enhanced clinical decision-making processes and therefore improved clinical outcomes. These systems, if pre-conditions persist, such as certain medical conditions, other specific measurements, or known lifestyle choices, could automatically invite patients for Type 2 Diabetes screening. These systems could also include visual interfaces

for healthcare workers, enabling a clear view of the decision tree models and their outputs during diagnosis. With these changes, combining the efforts of healthcare workers and data scientists, both patients and healthcare providers could be benefited.

6 Conclusion

In this paper, two decision trees were generated, the first is based on Dataset 1, the second is based on Dataset 2 and both were compared to the NHG guidelines on the diagnosis of DM-2. In conclusion, the decision trees and clinical guidelines shared many similarities in BMI, glucose, insulin, pregnancies, and blood pressure but differ in sleep and prediabetes, and completely overlook symptoms such as thirst, weight loss, neurogenic pains and sensory disorders. The results indicate that the influencing factors of these lifestyle choices and clinical measurements should be further investigated clinically. This paper may assist clinicians in the early diagnosis and prevention of DM-2 and provides a clear overview of the decision-making process.

References

[1] Diabetes mellitus type 2. (n.d.). NHG-Richtlijnen. Retrieved [22.06.2023], from https://richtlijnen.nhg.org/standaarden/diabetes-mellitus-type-2#samenvatting

[2] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice. 2019 Sep 10; 157(1): 1-5

[3] Nielen M, Poos R, Korevaar J. Diabetes mellitus in Nederland. Prevalentie en incidentie: heden, verleden en toekomst [Internet]. Utrecht: Nivel; date unknown; [update unknown]; Retrieved [16.06.2023], from https://www.nivel.nl/sites/default/files/bestanden/1003898.pdf

[4] Volksgezondheid en Zorg info. Diabetes mellitus | Leeftijd en geslacht [Internet]. Place unknown: Volksgezondheid en Zorg info; 2022 nov 29 Retrieved [23.06.2023], from https://www.vzinfo.nl/diabetesmellitus/leeftijd-en-geslacht

[5] Volksgezondheid en Zorg info. Diabetes mellitus | Zorguitgaven [Internet]. Place unknown: Volksgezondheid en Zorg info; 2022 July 6. Retrieved [16.06.2023], from https://www.vzinfo.nl/diabetesmellitus/zorguitgaven

[6] Lucier J, Weinstock RS. Diabetes Mellitus Type 1. StatPearls. 2022 May 11. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507713/

[7] Harreiter J, Roden M. Diabetes mellitus– Definition, Klassifikation, Diagnose, Screening und Prävention (Update 2019). Wiener Klinische Wochenschrift volume. 2019 Apr 12; 131(1): 6-15.

[8] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. 2014 Jan 1; 37(1): 81-90.

[9] Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. Current Genomics, 22(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

[10] Sim, J., Fong, Q. W., Huang, W., & Tan, C. M. (2021). Machine learning in medicine: what clinicians should know. Singapore Medical Journal. https://doi.org/10.11622/smedj.2021054

[11] Ministry of Health, Singapore. Clinical Practice Guidelines on Diabetes Mellitus. Available from: https://www.moh.gov.sg/docs/librariesprovider4/guidelines/cpg_diabetes-mellitus-booklet---jul-2014.pdf Last accessed on 04 Oct 2019.

[12] AlGhamdi, M., Al-Mallah, M. H., Keteyian, S. J., Brawner, C. A., Ehrman, J. K., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLOS ONE, 12(7), e0179805. https://doi.org/10.1371/journal.pone.0179805

[13] Neha Prerna Tigga and Dr. Shruti Garg of the Department of Computer Science and Engineering, BIT Mesra, Ranchi-835215 for research, non-commercial purposes only. For more information and citation of this dataset please refer: Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716. DOI: https://doi.org/10.1016/j.procs.2020.03.336.

[14] Daanouni, O., Cherradi, B., & Tmiri, A. (2019, October). Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In Proceedings of the 4th International Conference on Smart City Applications (pp. 1-6).

[15] MLVU. (n.d.). Lecture 5: Data Pre-Processing. Retrieved [21.06.2023.], from https://mlvu.github.io/lecture05#slide-014

[16] scikit-learn. (n.d.). sklearn.preprocessing.OneHotEncoder. In scikit-learn: Machine Learning in Python (version 1.2.2). Retrieved [20.06.2023.], from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[17] scikit-learn. (n.d.). Decision Trees. In scikit-learn: Machine Learning in Python (version 1.2.2). Retrieved [20.06.2023.], from https://scikit-learn.org/stable/modules/tree.html

[18] scikit-learn. (n.d.). Tree algorithms: ID3, C4.5, C5.0, and CART. In scikit-learn: Machine Learning in Python (version 1.2.2). Retrieved [20.06.2023.], from https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart

[19] scikit-learn. (n.d.). 3.3.1. Cross-validation: evaluating estimator performance. In scikit-learn: Machine Learning in Python (version 1.2.2). Retrieved [20.06.2023.], from https://scikit-learn.org/stable/modules/cross_validation.html

[20] MLVU. (n.d.). Lecture 3: Model Evaluation. Retrieved [21.06.2023.], from https://mlvu.github.io/lecture03#slide-023

[21] Ceballos, F. (2021, December 8). Scikit-Learn Decision Trees Explained - Towards Data Science. Medium. Retrieved [23.06.2023.], from https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

[22] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706–716. https://doi.org/10.1016/j.procs.2020.03.336

[23] Antza, C., Kostopoulos, G., Mostafa, S. A., Nirantharakumar, K., & Tahrani, A. A. (2022). The links between sleep duration, obesity and type 2 diabetes mellitus. Journal of Endocrinology, 252(2), 125–141. https://doi.org/10.1530/joe-21-0155

[24] Yuan, S., & Larsson, S. C. (2020). An atlas on risk factors for type 2 diabetes: a wide-angled Mendelian randomisation study. Diabetologia, 63(11), 2359–2371. https://doi.org/10.1007/s00125-020-05253-x

[25] Parameswaran, G., & Ray, D. W. (2021). Sleep, circadian rhythms, and type 2 diabetes mellitus. Clinical Endocrinology, 96(1), 12–20. https://doi.org/10.1111/cen.14607

[26] 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. (2017). Diabetes Care, 41(Supplement_1), S13–S27. https://doi.org/10.2337/dc18-s002

[27] Handelsman, Y., Bloomgarden, Z. T., Grunberger, G., Umpierrez, G. E., Zimmerman, R. A., Bailey, T. L., Blonde, L., Bray, G. A., Cohen, A. F., Dagogo-Jack, S., Davidson, J. A., Bloomgarden, Z. T., Ganda, O. P., Garber, A. M., Garvey, W. T., Henry, R. J., Hirsch, I. B., Horton, E. S., Hurley, D. L., . . . Zangeneh, F. (2015). American Association Of Clinical Endocrinologists And American College Of Endocrinology -Clinical Practice Guidelines For Developing A Diabetes Mellitus Comprehensive Care Plan – 2015. Endocrine Practice, 21, 1–87. https://doi.org/10.4158/ep15672.glsuppl

[28] McIntyre, H. D., Catalano, P. M., Zhang, C., Desoye, G., Mathiesen, E. R., & Damm, P. (2019). Gestational diabetes mellitus. Nature Reviews Disease Primers, 5(1). https://doi.org/10.1038/s41572-019-0098-8

[29] Diabetes en zwangerschap. (n.d.). NHG-Richtlijnen. Retrieved [22.06.2023], from https://richtlijnen.nhg.org/multidisciplinaire-richtlijnen/diabetes-en-zwangerschap

[30] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S. J., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice, 183, 109119. https://doi.org/10.1016/j.diabres.2021.109119

Appendix

Tree1.ipynb is shown below. Except for some minor differences, such as the K-value for Cross Validation, this code is identical to Tree2.ipynb. The code-base is available on GitHub: https://github.com/mrkbrts/vu-paih-group6

```
import pandas as pd
df = pd.read_csv('dataset1_imp.csv')
df.head()

from sklearn.preprocessing import OneHotEncoder

# Separate the numerical and categorical columns
numeric_cols = ['BMI', 'Sleep', 'SoundSleep', 'Pregancies']
categorical_cols = ['Age', 'Gender', 'Family_Diabetes', 'highBP', 'PhysicallyActive', 'Smoking',
'Alcohol', 'RegularMedicine', 'JunkFood', 'Stress', 'BPLevel', 'Pdiabetes', 'UriationFreq', 'Diabetic']

# Apply One-Hot Encoding to categorical columns
encoder = OneHotEncoder(handle_unknown='ignore')
encoded_categorical = encoder.fit_transform(df[categorical_cols])

# Create a dataframe for encoded categorical columns
encoded_categorical_df = pd.DataFrame(encoded_categorical.toarray(),
columns=encoder.get_feature_names_out(categorical_cols))

# Combine the numerical and encoded categorical columns
processed_df = pd.concat([df[numeric_cols], encoded_categorical_df], axis=1)

import matplotlib
import numpy as np
from sklearn import tree
from sklearn.model_selection import KFold
from sklearn.metrics import accuracy_score

# Load the Heart.csv dataset
X = processed_df.drop(['Diabetic_no', 'Diabetic_yes', 'Age_less than 40', 'Age_40-49', 'Age_50-
59', 'Age_60 or older'], axis=1)
Y = processed_df['Diabetic_yes']

# Prepare the K-Fold Cross Validation
n_folds = 10  # Set the number of folds
kf = KFold(n_splits=n_folds, shuffle=True, random_state=42)
evaluation_results = []

# Train with (K-fold Cross Validation)
for train_index, test_index in kf.split(X):
    x_train, x_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = Y.iloc[train_index], Y.iloc[test_index]
```

```
    # Train the decision tree classifier
    clf = tree.DecisionTreeClassifier(max_depth=3, criterion="entropy") # Set Maximum Depth
and other criteria
    clf = clf.fit(x_train, y_train)

    # Evaluate the model
    y_pred = clf.predict(x_test)
    accuracy = accuracy_score(y_test, y_pred)
    evaluation_results.append(accuracy)

    # Print evaluation result for the current fold
    print("Accuracy for fold {}: {:.5f}".format(len(evaluation_results), accuracy))

print("Average accuracy across all folds: {:.4f}".format(np.mean(evaluation_results)))

import matplotlib.pyplot as plt
from sklearn import tree
import graphviz

# Export the decision tree graph as Graphviz source code
dot_data = tree.export_graphviz(clf, out_file=None, feature_names=X.columns,
class_names=['diabetic_no','diabetic_yes'],filled=True, rounded=True, proportion=False)
graph = graphviz.Source(dot_data)
graph
```