Bridging Clinical Guidelines and Machine Learning: A Decision Tree Analysis of Diabetes Diagnosis

Amogh Kadam, Lamya Al Alawi

Faculty of Sciences, Vrije Universiteit Amsterdam

XB_0021: AI for Health Project

Annette ten Teije and Judith Bosmans

June 30, 2024

Bridging Clinical Guidelines and Machine Learning: A Decision Tree Analysis of Diabetes Diagnosis

Diabetes is a long-term medical condition marked by high blood glucose levels and abnormal protein and fat metabolism (Roglic, 2016). Blood glucose level rises due to the lack of ability to produce insulin or an inability to use insulin efficiently (Roglic, 2016). According to the initial WHO Global Report on Diabetes, there are 422 million people with diabetes, which nearly quadruples the number in 1980 (*Global Report on Diabetes,* 2016). With heart attack, stroke, blindness, renal failure, and amputation of a lower limb being among its consequences makes it a significant global health challenge (*Global Report on Diabetes* 2016).

Conventional techniques for diagnosing diabetes mostly rely on clinical guidelines that have been created by consensus among experts after substantial study. Although these recommendations are very helpful, it may be based on rigid standards that do not take into consideration the complicated, dynamic nature of unique patient data. On the other hand, machine learning algorithms have become very effective tools for improving medical diagnosis and decision-making in recent times. Of these methods, decision tree learning stands out in that it may provide clear, comprehensible models and it may provide explicit rules based on patient data. Hence, it is especially well-suited for medical applications where understanding the reasoning behind a forecast is just as crucial as the prediction itself.

This paper will employ decision tree learning to extract knowledge rules from patient data and compare them with recognised clinical guidelines for diabetes diagnosis. The goal is to ascertain how much machine-learned rules align with or depart from accepted clinical guideline approaches. The study focuses on the following research question, “How do the rules derived from decision tree learning on diabetes datasets compare to the clinical guidelines for diabetes diagnosis?”. The intended endpoint is that this study will provide

insights that can inspire the improvement of clinical recommendations as well as the creation of more efficient, data-driven diagnostic instruments. Therefore, it aims to close the gap between data-driven insights and clinical competence.

Related Work

Dudkina et al. (2021) created a decision tree model using an open dataset of 786 real-life patients, finding that more data allocated to testing leads to more accurate results. While Pei et al. (2020) used 14 dependent variables and two outcome variables, the result was high accuracy and recall in diabetes management. Tigga & Garg (2020) found Random Forest Classifier as the most accurate model for predicting Type 2 diabetes risk when evaluating diabetes risk based on risk factors. Chen et al. (2017) proposed a hybrid model using K-means and J48 decision tree for Type 2 diabetes classification, and Azad et al. (2022) introduced a novel prediction model using Decision trees, Synthetic Minority Oversampling, and Genetic Algorithm.

Methodology

Datasets and Clinical Guideline

The first dataset used in this report was collected by Neha Prerna Tigga and Dr. Shruti Garg of the Department of Computer Science and Engineering. It includes data on various health indicators collected from an online and offline questionnaire of 18 questions (Tigga & Garg, 2020). This dataset mainly had categorical data and not cleaned, hence in this experiment the dataset was cleaned before use. The second dataset used in this analysis originates from the Biostatistics program at Vanderbilt University. It includes real patient data

intended to predict diabetes using demographic and laboratory variables. The Diabetes_Classification file contains cleaned and manipulated data, excluding patients without haemoglobin A1c measurements so this was chosen instead of other files.

For the clinical guideline, the credibility and reputation of the National Institute for Health and Care Excellence (NICE) was noted as a diabetes guideline from a reliable and reputable source was required. The target population were adults with type 1 diabetes and the guideline represented that. This targeted approach ensures that the recommendations are directly applicable and precise, ultimately supporting the delivery of high-quality patient-centred care.

Experimental Setup

Tools

The first tool which was heavily used was Jupyter Notebook. Jupyter Notebook is an open-source platform which allows users to write code in multiple languages. In this experiment, the Jupyter Notebook was accessed via Visual Studio Code, and was coded in Python as Scikit Learn was only available using python.

Scikit-Learn is an open-source library for Python, which allows users to perform machine learning and data modelling. This made using the dataset and training a model easier and with no complications. As for cleaning the data, Pandas was used. Graphviz was another tool which helped visualise the decision trees. Finally, Miro was used to visualise the clinical guidelines.

Data Pre-processing

Different processes were used, which were dependent on the datasets. Since two datasets were utilised in the experiment, the data preprocessing process will be discussed in

relation to each of them. Firstly, for the first dataset there were some errors in the dataset such as variations in values (Low and low). Pandas does not recognise those errors, so the variations were reduced. There were also spelling mistakes (such as o, which meant no) which were corrected. Next, there were columns which were difficult to normalise. One example was age as ranges were used instead of an integer. To fix this issue, Pandas' get_dummies function was used to convert it into columns which have boolean values. Another step was to convert 'Yes' and 'No' into boolean values, using the replace function found in Pandas.

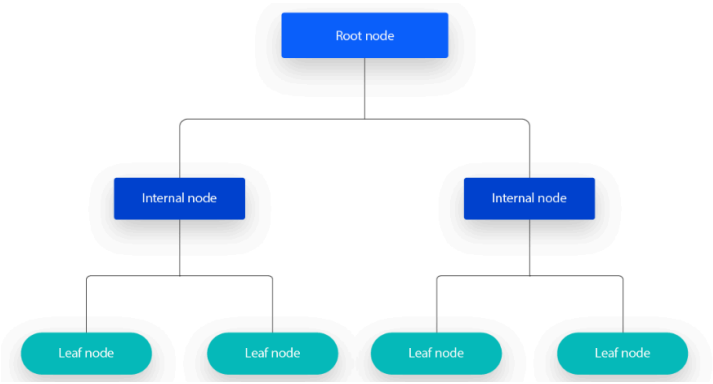
The second dataset was found to have some redundant columns which were removed. For example, since there was a column for Waist/Hip ratio, it was found to be redundant if there were also columns for Hip and Waist measurements. As for the target column Diabetes, it had values 'Diabetes' and 'No Diabetes'. These values were changed to boolean values aloowing it to be read by the model.

Decision Trees and its Algorithms

Decision trees are a popular supervised machine learning technique that works well for both classification and regression applications. The objective is to build a model that can predict the value of a target variable using basic decision rules derived from data attributes (*1.10. Decision Trees*, n.d.). It has three major components: the root node, internal (decision) nodes, and leaf nodes. The root node represents the full dataset, which is then partitioned into more homogenous groupings by decision nodes. Leaf nodes reflect the final choice or conclusion.

Figure 1

A view on how a decision tree is structured



Note. From What is a Decision Tree? by IBM, n.d. (https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.)

ID3 (Iterative Dichotomiser 3), created by Ross Quinlan, uses entropy and information gain as measures to assess potential splits (*What Is a Decision Tree? | IBM*, n.d.). When data gain reaches zero or all instances have a single goal, the tree stops growing (Shamrat et al., 2022). This approach is used for categorical data and cannot handle missing values, making it unsuitable for the datasets chosen.

CART (Classification and Regression Trees) is a decision tree algorithm that can handle both classification and regression problems. For classification, CART evaluates the quality of splits using the Gini impurity measure, with the goal of minimising impurity in the resultant nodes (*What Is a Decision Tree? | IBM*, n.d.). CART's use of binary splits simplifies the tree structure and implementation (Shamrat et al., 2022). In the paper comparing ID3, C4.5 (another decision tree algorithm), and CART by Shamrat et al. (2022) it was said that in terms of accurate identification rates CART was better than ID3 and C4.5. So, with that said, CART was chosen as the algorithm for the experiment.

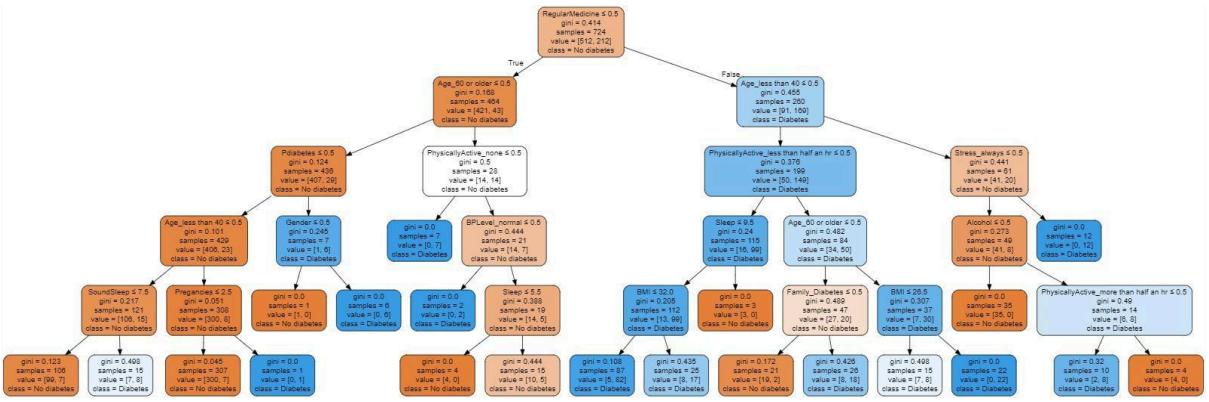
Experiment Method

In order to do a comparison between the decision trees and the clinical guidelines, the guidelines were converted into a flowchart to help visualise the guideline. The guideline split into two flowcharts, one that gave an overview of diabetes diagnosis and management while the other looked into a detailed section on diagnosing diabetes (section 1.1 in the guideline). Since decision trees can be seen as a type of visualised guideline with multiple decision points present, it was possible to make the comparison between the two. The detailed flowchart is then compared to the two decision trees respectively, looking into the differences and similarities.

Results

Figure 2

Decision tree based on the first dataset

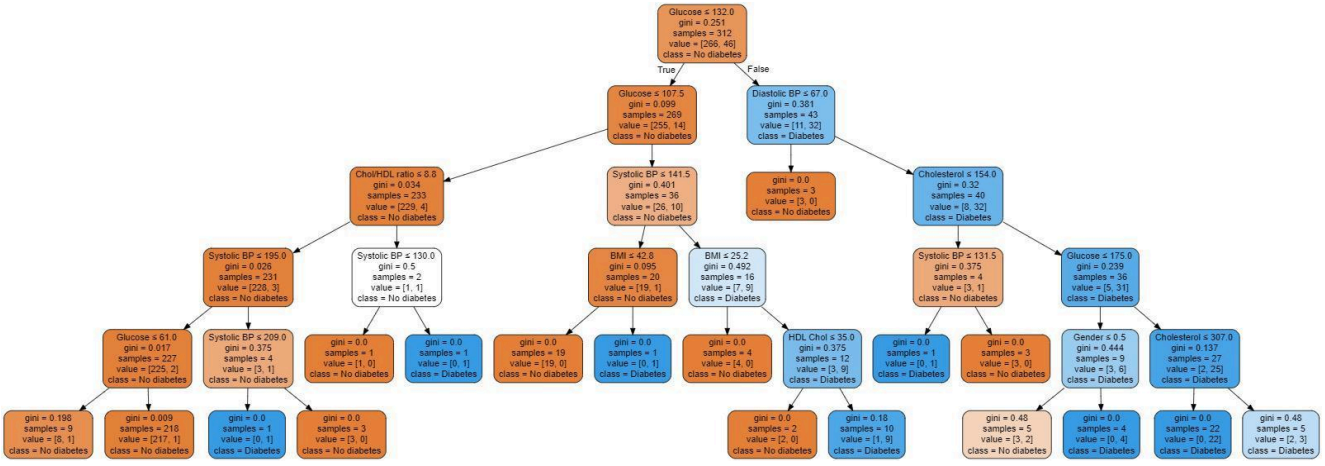


There are key decision points in the decision tree, which helps make the classifications. The root node (or the first decision point) of the decision tree is whether or not the patient takes medicine regularly. Another decision point is the Age (based on if they are 60 years or older), as age can be considered as a significant factor for diabetes (older

patients being more prone to it). The final example of a decision point in the decision tree is gestational diabetes, if the patient has it, it makes it more likely for the patient to develop diabetes. Each decision point in this decision tree narrows down the group into smaller groups, deciding if the groups have diabetes or not.

Figure 3

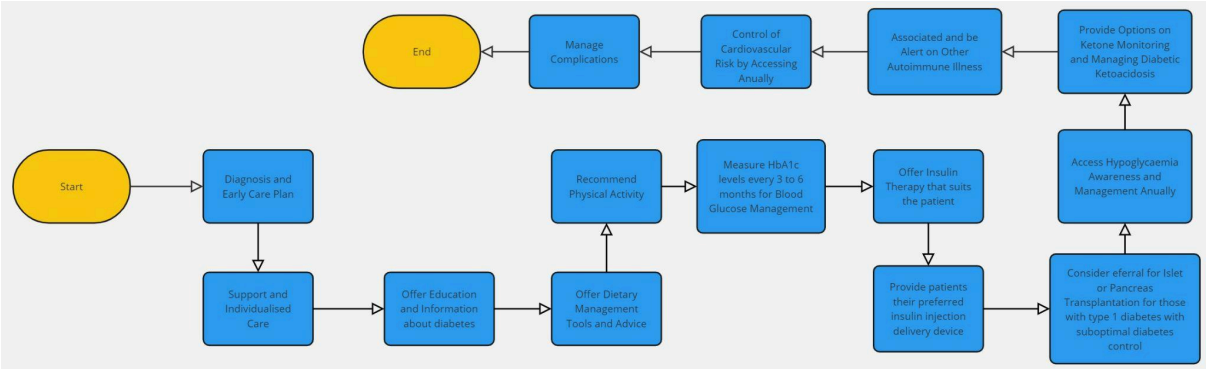
Decision Tree based on the second dataset



This second decision tree was based on the second dataset. There are now new decision points which are found in this decision tree. The root node of this decision tree is Glucose level (which was not considered in the first decision tree), and this node is once again found deeper in the decision tree to further narrow down the samples. Another decision point is BMI, which can be considered as a measurement for possible obesity (as it takes height and weight into consideration). The final example of a decision point is the cholesterol to HDL ratio, as it shows that lower cholesterol to HDL ratio indicates better heart health, and lower risk of diabetes.

Figure 4

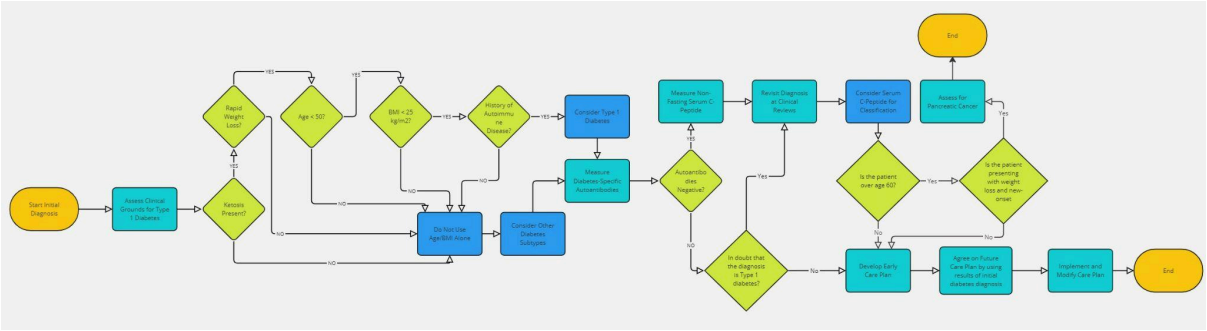
An overview of the guideline



The flowchart above was based on the NICE guideline's main points. This resulted in a fairly simple diagram showing the process from diagnosing type 1 diabetes in adults (those age 18 and older) to educating those diagnosed with relevant information to help manage their condition. After that the overview of the treatment plan can be seen and it then goes to show doctors also need to be aware and alert on other possible autoimmune illnesses. Finally, the rest of the care plan is to manage complications of type 1 diabetes.

Figure 5

Detailed flowchart on diagnosis and early care plan of diabetes



Here it starts with accessing diabetes by checking if the patient has one or more symptoms (ketosis present, rapid weight loss, etc). If the answer is no to any of them it is reminded to not use age or BMI alone as a diagnosis for diabetes. Then it goes to suggest looking into other diabetes subtypes. If the answer was yes, then it suggests considering type 1 diabetes. The two suggestions lead to measuring antibodies. If the result was negative then the measurement of non-fasting serum c-peptide and re-evaluation of diagnosis is needed. Otherwise if positive and not in doubt, then it goes to the overview of care plans.

Re-evaluation is also needed if the result was positive and in doubt. If the patient was above 60 and having weight loss and new-onset then they are to be assessed for pancreatic cancer. Lastly, if the patient was under 60 or did not have the symptoms of weight loss and new-onset then they are presented with their care plans.

Discussion

The similarities between the first dataset and NICE guideline on type 1 diabetes in adults is that it both uses BMI and age as an indicator to diabetes. It can be seen that in the guideline it asks to not use BMI or age alone as a definite indicator to type 1 diabetes, this is followed in the decision tree as there are other nodes that define whether or not the patient has diabetes or not. There is also a similarity between the decision tree and the clinical guidelines, which is using BMI as a decision point. In the decision tree, BMI is used two times as a decision point, which have the same parent node. If Systolic BP is smaller than 141.5, then it would lead to one of the BMI child nodes. If it wasn't smaller than 141.5, it would once again lead to a BMI child node. If Systolic BP requisites have not been met, the next decision point would be if the BMI of the patient is < 25.2 kg/m². The clinical guidelines has a similar decision point where it asks if the patient has a BMI of 25 kg/m², which has a similar prerequisite as the decision tree.

However, this is where similarities end, for example, the second decision tree doesn't consider the age of the patient when classifying if he/she has diabetes. Furthermore in both decision trees, it does not consider if there has been any rapid weight loss, if ketosis is present or if there is a history of Autoimmune disease. This clearly shows that clinical guidelines (which are supported by scientific methods) consider different and considerably less factors when diagnosing a patient with type 1 diabetes, in comparison to the decision trees. This could hint to the fact that the decision tree may have redundant decision points

when performing classifications, which may lead to misclassification or an overall slower process. On the other hand, this may otherwise hint that the dataset include other subtypes of diabetes and not only type 1.

Despite the discrepancies between the datasets and guidelines, both datasets included in this study are useful since they include a wide range of demographic and health characteristics important for diabetes diagnosis and management. The use of genuine patient data from several sources assures that the conclusions are applicable to real-world circumstances. The datasets conformity with NICE emphasises its usefulness even more. Using decision tree algorithms to analyse this information results in unambiguous, interpretable principles that physicians can easily understand is an important aspect. This transparency is critical in clinical settings, as knowing the reasoning behind a diagnosis is equally significant as the diagnosis itself.

Limitation

This study has various limitations that should be noted. The emphasis is solely on comparing diagnostic data, which restricts the capacity to draw broad conclusions about the whole healthcare process or patient outcomes by excluding treatment efficacy, patient management techniques, and quality of life after diagnosis. Furthermore, the research approach is observational rather than experimental, making it harder to demonstrate causation or rule out confounding variables. Furthermore, the study only included adults with Type 1 diabetes, limiting the findings' applicability to other groups such as children, adolescents, and those with Type 2 diabetes. These limitations emphasise the unique circumstances and limits under which the study was done, influencing the findings and their broader applicability.

Conclusion

In conclusion, this work shows that decision tree learning has the potential to improve the accuracy and interpretability of diabetes diagnosis when compared to established clinical criteria. By comparing machine-learned rules obtained from patient data to existing clinical standards, this resulted in discovering both similarities and differences that indicate potential for incorporating data-driven insights into clinical practice. While the decision trees had some redundant decision points, it provided a more thorough approach by taking into account a wider variety of variables. This implies that, while clinical standards are important for providing consistent care, machine learning models can deliver personalised and thorough diagnostic insights that may enhance patient outcomes.

Future work

Given the limitations found in this work, various directions for further research are suggested. First, broadening the scope beyond diagnostic data to include treatment efficacy, patient management approaches, and quality of life after diagnosis will offer a more complete picture of the healthcare process and patient outcomes. Second, using a controlled experimental design might assist establish causality and limit the effect of confounding factors, hence increasing the findings' validity. Furthermore, expanding the research population to include children, adolescents, and people with Type 2 diabetes will improve the generalizability of the findings. Future study should also look into how other demographic characteristics, such as socioeconomic position, ethnicity, and geographic location, influence diabetes diagnosis and management. These guidelines will aid in establishing a more comprehensive understanding of diabetes treatment and enhancing healthcare methods for a wide range of populations.

Self-Reflection

Amogh Kadam

In this report, exploring the realm of data-cleaning methods via Pandas to convert the patient dataset into an asset for the machine learning model, was quite insightful, since it made me realise how not all datasets are perfect. It could be possible that datasets may contain redundant columns or missing/incorrect values. This made me think about how easily influenced a model can be by uncleaned/unprocessed datasets. If this model was applied in healthcare, there would be serious detriments to the model's predictions, and the health of the patients. One particular difficulty that arose during the report was understanding the clinical terms found in the dataset columns and their implications. Some columns could be redundant and were still included in the decision tree model. In the future, more research should be done on each column and their influence on the diagnosis of diabetes.

Lamya Al Alawi

This paper investigated the feasibility of bridging the gap between machine learning and healthcare. The inspiration for adopting this issue originated from the need to strengthen healthcare recommendations, which are crucial for both patients and healthcare professionals. Throughout this work, I improved my data analysis and visualisation, and machine learning abilities. I also developed a greater awareness of the value of data-driven evidence-guideline techniques. One important problem was determining the significance of specific medical words in the dataset and whether or not they should be included. In the future, investing more time to grasp what each column represents is critical for a successful start in data pre-processing.

References

1.10. Decision Trees. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/tree.html

Azad, Chandrashekhar, et al. "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus." Multimedia Systems (2022): 1-19. https://link.springer.com/article/10.1007/s00530-021-00817-2

Chen, Wenqian, et al. "A hybrid prediction model for type 2 diabetes using K-means and decision tree." 2017 8th IEEE international conference on software engineering and service science (ICSESS). IEEE, 2017. https://ieeexplore.ieee.org/abstract/document/8342938/

Dudkina, Tetiana, et al. "Classification and Prediction of Diabetes Disease using Decision Tree Method." IT&AS. 2021. https://ceur-ws.org/Vol-2824/paper16.pdf

Pei, Dongmei, Tengfei Yang, and Chengpu Zhang. "Estimation of diabetes in a high-risk adult Chinese population using J48 decision tree model." Diabetes, Metabolic Syndrome and Obesity (2020): 4621-4630. https://www.tandfonline.com/doi/abs/10.2147/DMSO.S279329

Roglic, G. (2016). WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), 3. https://doi.org/10.4103/2468-8827.184853

Shamrat, F. M. J. M., Ranjan, R., Hasib, K. M., Yadav, A., & Siddique, A. H. (2022). Performance evaluation among ID3, C4.5, and CART Decision Tree algorithm. In Lecture notes in networks and systems (pp. 127–142). https://doi.org/10.1007/978-981-16-5640-8_11

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

What is a Decision Tree? | IBM. (n.d.). https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

World Health Organization. (2016). Global Report on Diabetes.