information

Article

Using Generative AI to Improve the Performance and Interpretability of Rule-Based Diagnosis of Type 2 Diabetes Mellitus

Leon Kopitar [1,2,*], Iztok Fister, Jr. [2] and Gregor Stiglic [1,2,3]

[1] Faculty of Health Sciences, University of Maribor, Zitna Ulica 15, 2000 Maribor, Slovenia; gregor.stiglic@um.si
[2] Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroska Cesta 46, 2000 Maribor, Slovenia
[3] Usher Institute, University of Edinburgh, Teviot Place, Edinburgh 0131, UK
* Correspondence: leon.kopitar1@um.si

Abstract: Introduction: Type 2 diabetes mellitus is a major global health concern, but interpreting machine learning models for diagnosis remains challenging. This study investigates combining association rule mining with advanced natural language processing to improve both diagnostic accuracy and interpretability. This novel approach has not been explored before in using pretrained transformers for diabetes classification on tabular data. Methods: The study used the Pima Indians Diabetes dataset to investigate Type 2 diabetes mellitus. Python and Jupyter Notebook were employed for analysis, with the NiaARM framework for association rule mining. LightGBM and the dalex package were used for performance comparison and feature importance analysis, respectively. SHAP was used for local interpretability. OpenAI GPT version 3.5 was utilized for outcome prediction and interpretation. The source code is available on GitHub. Results: NiaARM generated 350 rules to predict diabetes. LightGBM performed better than the GPT-based model. A comparison of GPT and NiaARM rules showed disparities, prompting a similarity score analysis. LightGBM's decision making leaned heavily on glucose, age, and BMI, as highlighted in feature importance rankings. Beeswarm plots demonstrated how feature values correlate with their influence on diagnosis outcomes. Discussion: Combining association rule mining with GPT for Type 2 diabetes mellitus classification yields limited effectiveness. Enhancements like preprocessing and hyperparameter tuning are required. Interpretation challenges and GPT's dependency on provided rules indicate the necessity for prompt engineering and similarity score methods. Variations in feature importance rankings underscore the complexity of T2DM. Concerns regarding GPT's reliability emphasize the importance of iterative approaches for improving prediction accuracy.

Keywords: GPT; association rule mining; classification; interpretability; diagnostics

check for updates

Citation: Kopitar, L.; Fister, I., Jr.; Stiglic, G. Using Generative AI to Improve the Performance and Interpretability of Rule-Based Diagnosis of Type 2 Diabetes Mellitus. *Information* **2024**, *15*, 162. https://doi.org/10.3390/info15030162

Academic Editors: Vikas Palakonda and Jae-Mo Kang

Received: 9 February 2024
Revised: 23 February 2024
Accepted: 5 March 2024
Published: 12 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Type 2 diabetes mellitus (T2DM), also known as non-insulin-dependent diabetes, is a prevalent and serious health condition, causing significant illness and death. In 2021, it was reported that around 537 million adults live with diabetes. Projections suggest that the prevalence of diabetes will continue to grow, reaching an estimated 643 million people by 2030 and 783 million by 2045 [1]. It can lead to significant complications such as cardiovascular problems, kidney dysfunction, vision impairment, or even death. Every five seconds, a life is lost due to diabetes. Only in 2021, approximately 6.7 million deaths were caused by diabetes [1]. Early detection through diagnostic methods allows for timely intervention and healthcare, potentially mitigating or postponing these negative outcomes.

Precision medicine, which uses patient-specific data in conjunction with AI to develop personalized treatment recommendations for hypertensive patients with T2DM based on their health characteristics, has the potential to reduce healthcare costs, alleviate patient stress, and enhance overall quality of life [2]. Advancements in diagnostic research have evolved to the point where AI and portable handheld devices, like smartphone-based retinal cameras, are used for the detection of diabetic retinopathy, a diabetes-related eye disease [3]. Furthermore, in a recent study, Tao et al. explored the possibility of employing continuous glucose monitoring profiles and a deep learning nomogram as a tool to predict diabetic retinopathy in individuals with T2DM [4]. Additionally, the rise of large language models (LLMs), such as generative pretrained transformer (GPT), has introduced endless opportunities, since they are extensively trained on vast numbers of data.

Large language models (LLMs), such as GPT, have the capability to revolutionize various domains, from summarizing patient data to patient management and all the way to assisting in making a diagnosis [5]. Numerous studies have taken on the challenge of diagnosing T2DM in recent years with various models such as glmnet, lightGBM, random forest, XGBoost, logistic regression, support vector machine, naive Bayes, soft voting, different decision trees, K-means clustering, association rule mining, and others [6–8]. None of these studies utilized the association rule mining approach in conjunction with GPT to make predictions.

Association rule mining is an approach to finding patterns between variables in large datasets, and its result is the formation of several IF-THEN rules. These rules, by themselves, do not make any predictions and yet have to be summarized. Recognizing the challenge of interpretability and prediction based on complex association rules, GPT possesses the capability of processing vast amounts of rules, interpreting, and recognizing patterns. Lacking consciousness or awareness, GPT relies solely on public data and provided rules, integrating information to simplify the process of making straightforward classifications. Concurrently, it effortlessly provides the rules that form the basis of decision making, thus clarifying the particular rules utilized to arrive at a decision.

Logistic regression and XGBoost are classification models that rely on tabular data for prediction, whereas GPT is designed to work with textual data. The given text can be processed and stored in a dictionary format that resembles tabular data. When used together with association rules, we believe it has the potential to achieve high performance.

This manuscript aims to explore the synergies between association rule mining and state-of-the-art LLPs, particularly GPT, in improving the classification and interpretability of T2DM diagnosis. Through a proof-of-concept study on a prominent diabetes classification dataset, we seek to demonstrate how this integrated approach can enhance diagnostic accuracy and provide actionable insights for healthcare practitioners.

In this proof-of-concept study, we explore the interactions between association rule mining and state-of-the-art NLP techniques to enhance the classification and interpretability of the results demonstrated on a well-known diabetes classification dataset.

To date, we have not encountered any studies conducting GPT-based diabetes classifications on textual data based on the association rules, let alone examining the interpretation of GPT.

2. Materials and Methods

2.1. Dataset and Features

This study was conducted on the Pima Indians Diabetes [9] dataset. The dataset comprises 768 samples, eight predictors, and one outcome class. The following attributes can be found in the dataset: number of times pregnant (Pregnancies), 2 h plasma glucose concentration (Glucose), diastolic blood pressure [mmHg] (BloodPressure), triceps skin fold thickness [mm] (SkinThickness), 2 h serum insulin level [mu U/mL] (Insulin), body mass index (BMI), diabetes pedigree function (DiabetesPedigreeFunction), age [years], and the outcome as diabetic or non-diabetic status (see Table 1).

Table 1. Feature descriptions.

Features	Meaning
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration (2 h in an oral glucose tolerance test)
BloodPressure	Diastolic blood pressure (mmHg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2 h serum insulin (mu U/mL)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Diabetes

The outcome class is slightly imbalanced, with 35% of patients diagnosed with T2DM. The descriptive statistics for all predictors according to the outcome are displayed in Table 2.

Table 2. Descriptive statistics of the dataset.

Features	Non-diabetic—0 [1]	Diabetic—1 [1]	*p*-Value [2]
Pregnancies	3.3 (±3.0)	4.9 (±3.7)	<0.0001
Glucose	110.0 (±26.1)	141.3 (±31.9)	<0.0001
BloodPressure	68.2 (±18.1)	70.8 (±21.5)	<0.0001
SkinThickness	19.7 (±14.9)	22.2 (±17.7)	0.013
Insulin	68.8 (±98.9)	100.3 (±138.7)	0.066
BMI	30.3 (±7.7)	35.1 (±7.3)	<0.0001
DiabetesPedigreeFunction	0.4 (±0.3)	0.6 (±0.4)	<0.0001
Age	31.2 (±11.7)	37.1 (±11.0)	<0.0001

[1] mean (±SD). [2] Wilcoxon rank sum test for continuous values.

2.2. Research Tools

We conducted all analyses in the programming language Python (version 3.12.1), using a Jupyter Notebook. A minimalistic framework for Numerical Association Rule Mining (NiaARM) [10] was used to extract the association rules. NiaARM was used since it enables the management of numerical attributes without the need for discretization [11,12], unlike in other traditional approaches [13,14], where discretization results in the loss of information and its bias can cause inaccuracies in the outcomes [15]. LightGBM [16] was employed as one of the advanced machine learning models for performance comparison against GPT's decisions. It is based on gradient boosting machines with gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) which reduces the computational cost and improves the efficiency of the algorithm. We utilized the dalex package in order to obtain feature importance on the level of local and global interpretability [17]. Shapley additive explanations (SHAP) denote a local interpretability technique utilized to offer insights into how models make decisions, with an emphasis on comprehending decisions on the individual level [18].

We conducted the research on a personal computer with the following specifications: AMD Ryzen 9 3900X 12-core processor 3.80 GHz with 64 GB of RAM.

OpenAI GPT version 3.5, gpt-3.5-turbo-16k-0613, was used to predict the outcome and obtain the interpretation of results. It is limited to 180,000 tokens per minute and 3500 requests per minute using the application programming interface (API) [19]. It is important to note that GPT was not fine-tuned, but only instructed to make decisions based on the provided set of association rules.

The source code is provided on GitHub: https://github.com/lkopitar/GenAI_NiaARM/ (accessed on 20 February 2024).

2.3. Study Design

We aim to enhance the capabilities of GPT by incorporating association rules to improve classification metrics (F1-score, AUC, specificity, sensitivity). While acknowledging the irrelevance of accuracy due to class imbalance, we will include it for comparison with similar problems in other studies.

The dataset did not undergo any specific preprocessing. The dataset was stratified into a training and test set in a ratio of 80/20, resulting in 614 training samples. Since we dealt with an imbalanced dataset (0.65/0.35), where the minority class represents diabetic patients (diabetic status (0)), we oversampled the training set with the oversampling method SMOTE [8,20]. Consequently, oversampling resulted in 800 training samples. The whole preprocessing, rule generation, and GPT testing process is displayed with a pseudocode in Algorithm 1.

Algorithm 1: Pseudocode of the whole process.

```
$\#Preprocessing:$
$\mathbf{D}_{train}, \mathbf{D}_{test} \leftarrow \mathbf{DataStratification}(data, 80\%, 20\%)$
$\mathbf{D}_{balanced\_train} \leftarrow \mathbf{SMOTE}(\mathbf{D}_{train})$
$\#Rulegeneration$
while $TRUE$ do
    $\mathbf{Res}_{rules}, \mathbf{Res}_{perf}, \mathbf{Ratio}_{outcome\_0}, \mathbf{Cov}_{outcomes} \leftarrow NULL$
    $\mathbf{Res}_{rules}, \mathbf{Res}_{perf} \leftarrow \mathbf{NiaARM}(\mathbf{D}_{balanced\_train})$
    $\mathbf{Ratio}_{outcome\_0}, \_ \leftarrow \mathbf{GetOutcomeRatio}(\mathbf{Res}_{rules})$
    $\mathbf{Cov}_{outcomes} \leftarrow \mathbf{GetOutcomesCoverage}(\mathbf{Res}_{rules})$
    if $400 \geq len(\mathbf{Res}_{rules}) \land (0.3 \leq \mathbf{Ratio}_{outcome\_0} \leq 0.7) \land \mathbf{Cov}_{outcomes} \geq \frac{1}{10}$ then
        if $\mathbf{Res}_{perf}[fitness, support, confidence] \geq ALL(0.7)$ then
            break
        end
    end
end
$\#GPTtesting$
$Outcome\_RuleList\_Map \leftarrow newdataframe()$
for $idx \leftarrow 1, 2, \ldots, len(\mathbf{D}_{test})$ do
    $test\_sample \leftarrow \mathbf{toDictionary}(\mathbf{D}_{test}[idx])$
    $test\_outcomes, GPT\_rulesets \leftarrow [\,]$
    for $n \leftarrow 1, 2, 3, 4, 5$ do
        $\mathbf{Pred}_{outcome}, \mathbf{Pred}_{ruleset} \leftarrow \mathbf{GPT}(test\_sample, \mathbf{Res}_{rules})$
        $test\_outcomes, GPT\_rulesets \leftarrow AppendToLists(\mathbf{Pred}_{outcome}, \mathbf{Pred}_{ruleset})$
    end
    $outcome, rulelist \leftarrow \mathbf{extractionMajorityVoting}(test\_outcomes, GPT\_rulesets)$
    $Outcome\_RuleList\_Map \leftarrow \mathbf{addRow}(outcome, rulelist)$
end
return $Outcome\_RuleList\_Map$
```

After the preprocessing step, we performed association rule mining on the training set to obtain the most relevant rules, aiming to comprehensively represent and cover the training set.

We utilized the NiaARM framework (method *NiaARM*()) to conduct association rule mining. NiaARM yields a rule list (variable *Res_rules*), a set of rules composed of antecedents and consequents. It accepts parameters such as optimization algorithm, metrics, and maximum number of iterations or function evaluations. Differential evolution (DE) [21] was used as an optimization algorithm with the following default parameter values: population size (value: 50), scaling factor (value: 0.5), and crossover probability (value: 0.9).

Given that these are default parameters, they serve as a solid starting point. For this research, we did not plan to conduct parameter optimization due to limited resources. Support and confidence were used as evaluation metrics in this step.

We established the requirement that a rule list should not exceed 400 rules, primarily considering the limitations of GPT. Additional criteria were implemented to prevent extreme imbalance in the representation of outcome classes by the following rules:

- The proportion of an outcome class should always be between 0.3 and 0.7, ensuring that rule representation is not highly imbalanced.
- The number of rules containing the outcome must be at least 1/10 of the total generated rules.

As a final criterion, we selected the first set of rules that exceeded the average performance of more than 0.7 for all of the following metrics: average fitness, average support, and average confidence.

In the subsequent step, the provided prompt was input into GPT five times. The prompt is comprised of three messages:

- Message 1: You will be provided with a set of rules presented in a RuleList format, each containing an antecedent and a consequent. For example, a rule might be given as $[A([19, 67])] \Rightarrow [B([59, 98]), C([2.56, 8.3])]$, signifying that if 'A' falls within the range of 19 and 67, it implies that 'B' must fall within the range of 59 and 98, and 'C' must fall within the range of 2.56 and 8.3. The rule list is as follows:
- Message 2: The rule list is provided as a list of rules.
- Message 3: For the test sample: $< TEST\_SAMPLE >$ determine the 'Outcome' value. Respond with an answer in the following format 'Outcome = x' where x can be either '0' or '1'. Additionally, provide a set of rules that significantly contributed to the obtained result in highest frequencies. Provide them in the following format 'Rules = [Rule1; Rule2; Rule3; Rule4]'. Rules should be written as an [antecedent] ⇒ [consequent] and should be separated by a semicolon.

The test sample ($< TEST\_SAMPLE >$) is provided in a dictionary format as follows: {'Pregnancies': 2.0, 'Glucose': 56.0, 'BloodPressure': 56.0, 'SkinThickness': 28.0, 'Insulin': 45.0, 'BMI': 24.2, 'DiabetesPedigreeFunction': 0.332, 'Age': 22.0}.

For each test sample, GPT generated responses five times, providing predictions along with a set of rules that contributed to the obtained result with the highest frequency. These predictions and the associated set of rules for a particular test sample were temporarily stored (method *extractionMajorityVoting*()). The set of rules for a particular test sample is expected to be selected from the NiaARM pool of generated rules. Regarding predicted values, there are instances where GPT may be uncertain and responds in a way that makes it impossible to distinguish the predicted outcome value. This issue was addressed by re-feeding the GPT with the same prompt until the appropriate prediction was obtained (method *GPT*()). The final prediction was determined by applying the majority voting rule to the five predictions (method *extractionMajorityVoting*()).

For example, GPT predicted the following outcomes [1, 1, 1, 0, 0] for the test sample 'x' with the corresponding sets of the most frequent rules [rule list 1, rule list 2, rule list 3, rule list 4, rule list 5]. Therefore, the decision for sample 'x' indicates a diabetic status (value 1), determined by the rules present in rule lists 1, 2, and 3, respectively.

The architecture of the proposed approach is presented in Figure 1.

In the next subsection, we present an example prompt with its corresponding response, along with an accompanying explanation. The rule list is kept simplified.

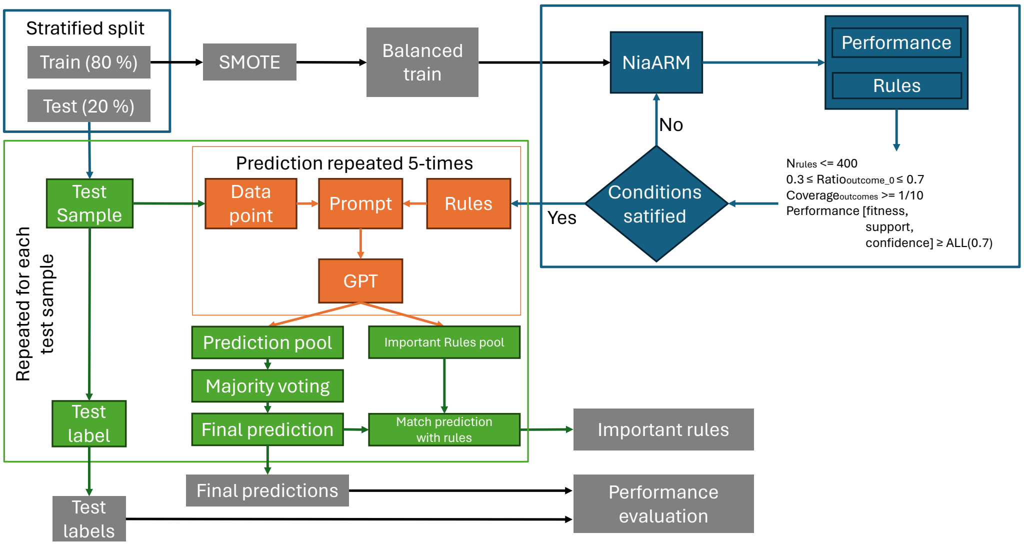

Figure 1. Architecture of the proposed approach.

2.4. Demonstration Example

Prompt: You will be provided with a set of rules presented in a RuleList format, each containing an antecedent and a consequent. For example, a rule might be given as [A([19, 67])] ⇒ [B([59, 98]), C([2.56, 8.3])], signifying that if 'A' falls within the range of 19 and 67, it implies that 'B' must fall within the range of 59 and 98, and 'C' must fall within the range of 2.56 and 8.3. The rule list is as follows: *RuleList([[Insulin([79, 595]), BMI([10.81, 63.30]), BloodPressure([0, 99])] ⇒ [DiabetesPedigreeFunction([1.22, 2.08]), Outcome(1), Pregnancies([6, 17])], … [BloodPressure([73, 119])] ⇒ [Insulin([0, 639]), Outcome(0), DiabetesPedigreeFunction([1.55, 2.42]), Pregnancies([0, 12])]])*. For the test sample: *{'Pregnancies': 2.0, 'Glucose': 56.0, 'BloodPressure': 56.0, 'SkinThickness': 28.0, 'Insulin': 45.0, 'BMI': 24.2, 'DiabetesPedigreeFunction': 0.332, 'Age': 22.0}* determine the 'Outcome' value. Respond with an answer in the following format 'Outcome = x' where x can be either '0' or '1'. Additionally, provide a set of rules that significantly contributed to the obtained result. Provide them in the following format 'Rules = [Rule1; Rule2; Rule3; Rule4]'. Rules should be written as an [antecedent] ⇒ [consequent] and should be separated by a semicolon.

Response: 'Outcome = 0. Rules = *[[Insulin([0, 315]) ⇒ Pregnancy([2, 8]), Outcome(0)], [BMI([5.33, 60.51]) ⇒ Pregnancy([2, 6])], [BloodPressure([83, 112]) ⇒ Glucose([95, 142]), Outcome(0)], [BloodPressure([23, 59]) ⇒ Pregnancy([0, 2]), Outcome(1)], [BloodPressure([87, 122]) ⇒ Glucose([95, 199]), Outcome(0)], [Insulin([0, 427]), BMI([8.15, 64.29]) ⇒ Outcome(0), Pregnancy([2, 12])], [Insulin([63, 518]), DiabetesPedigreeFunction([1.08, 2.08]), BloodPressure([11, 114]), BMI([19.46, 66.36]) ⇒ Outcome(1), Pregnancy([0, 17])], [BloodPressure([61, 85]) ⇒ Glucose([95, 197]), Outcome(0)], [BloodPressure([11, 24]) ⇒ Glucose([95, 148]), Outcome(0)], [BMI([3.83, 58.95]) ⇒ Pregnancy([0, 2]), Outcome(1)], [BMI([2.54, 59.88]) ⇒ Pregnancy([0, 17]), Outcome(0)], [BloodPressure([0, 79]) ⇒ Glucose([95, 142]), Outcome(0)]]'*.

Explanation: Outcome 0 and the provided set of rules are then extracted and stored in the list of responses and rules (variables *test_outcomes* and *GPT_rulesets* in Algorithm 1). The same prompt is run five times. Majority voting then determines the final prediction. Only rules corresponding to the final prediction are kept in the pool. For instance, if GPT predicts four times that the outcome is 0, and that becomes the final decision, four different set of rules are stored in the pool for that prediction.

In this demonstration example, a prediction yields an outcome of 0 with 12 rules being important for the decision in favor of outcome 0. Assuming a similar number of rules will be generated in the remaining three sets, this will bring up to 3 × 12 rules in favor of the predicted outcome 0.

Real-world scenario: We consider the most frequently reported rules (see Table 3) and the following patient:

- Gender: female.
- Age: 35.
- Insulin: 250 (mu U/mL).
- BMI: 21.5.
- BloodPressure: 100/70 mmHg.
- Glucose: 100 mg/dL.
- DiabetesPedigreeFunction: 0.2.
- Pregnancies: 2.

Table 3. Comparison of NiaARM and GPT-recognized rules with frequency values of at least 0.1.

Predicted	NiaARM Rules	GPT-Recognized Rules	Similarity	Frequency
0	[Insulin([0, 427]), BMI([8.148, 64.294]), BloodPressure([11, 114])] => [Outcome(0), Pregnancies([2, 12])]	[Insulin(0, 427), BMI(8.148, 64.294), BloodPressure(11, 114)] => [Outcome(0), Pregnancies(2, 12)]	0.965	0.012
0	[Insulin([281, 386])] => [Outcome(0)]	Insulin(0, 744)] => [Outcome(0)]	0.813	0.012
1	[Insulin([157, 209])] => [Outcome(1)]	[Insulin(0, 846)] => [Outcome(1)]	0.813	0.058
1	[Insulin([0, 601])] => [BloodPressure([71, 76]), Pregnancies([0, 7]), Outcome(0), DiabetesPedigreeFunction ([1.568, 2.42])]	[Insulin(0, 601)] => [BloodPressure(71, 76), Pregnancies(0, 7), Outcome(0), DiabetesPedigreeFunction (1.568, 2.42)]	0.967	0.023
1	[Insulin([0, 427]), BMI([8.148, 64.294]), BloodPressure([11, 114])] => [Outcome(0), Pregnancies([2, 12])]	[Insulin(0, 427), BMI(8.148, 64.294), BloodPressure(11, 114)] => [Outcome(0), Pregnancies(2, 12)]	0.965	0.016

The rules dictate the following:

- Rule 1: [Insulin([0, 427]), BMI([8.148, 64.294]), BloodPressure([11, 114])] ⇒ [Outcome(0), Pregnancies([2, 12])]. The patient's insulin level (250 pg/mL) falls within the range of the first rule, and the patient's BMI (21.5) and blood pressure (100/70 mmHg) also fall within the ranges specified in the rule. Therefore, we can apply this rule to the patient. The rule predicts an outcome of 0 (non-diabetic) and a range of 2–12 pregnancies, which corresponds to the patient; even if the pregnancy level was absent or none, it still would. Predicted: 0.
- Rule 2: [Insulin([281, 386])] ⇒ [Outcome(0)]. The patient's insulin level (250 pg/mL) does not fall within the range of the second rule. Therefore, we cannot apply this rule to the patient. Predicted: None.

- Rule 3: [Insulin([157, 209])] ⇒ [Outcome(1)]. The patient's insulin level (250 pg/mL) falls outside the range of the third rule. Therefore, we cannot apply this rule to the patient. Predicted: None.
- Rule 4: [Insulin([0, 601])] ⇒ [BloodPressure([71, 76]), Pregnancies([0, 7]), Outcome(0), DiabetesPedigreeFunction ([1.568, 2.42])]. The patient's insulin level (250 pg/mL) falls within the range of the fourth rule. Therefore, we can apply this rule to the patient. Predicted: 0.
- Rule 5: [Insulin([0, 427]), BMI([8.148, 64.294]), BloodPressure([11, 114])] ⇒ [Outcome(0), Pregnancies([2, 12])]. The patient's insulin level, BMI, and blood pressure fall within specified ranges of the fifth rule. Therefore, we can apply the rule. Predicted: 0.

Three rules were applicable for the given patient, all of which favored non-diabetic status. This resulted in a score of [0, None, None, 0, 0], with a confidence of 1, calculated as 3/3. Naturally, this is relevant under the assumption that the prediction model achieves at least good performance with an AUC higher than or equal to 80 [22].

3. Results

3.1. Niaarm Rules

The final rule list generated by NiaARM consisted of 350 rules, with an average fitness of 0.81, average support of 0.72, and average confidence of 0.90. In total, 58 rules included the outcome feature, with 36 rules having the outcome as a consequent and 22 as an antecedent. Among these 58 rules, both outcome classes were equally represented at 50% (29 rules each). Outcome 1 was predominantly present as an antecedent (82%, 18 of 22 rules), whereas outcome 0 appeared in 69% of rules (25 of 36 rules) as one of the consequents.

3.2. Model Performance

In terms of model performance, lightGBM (F1-score: 0.714, AUC: 0.780, specificity: 0.82, sensitivity: 0.741, accuracy: 0.792) performed better than the GPT association rules-based model (F1-score: 0.656, AUC: 0.73, specificity: 0.7, sensitivity: 0.759, accuracy: 0.721) across all five evaluation metrics.

3.3. GPT Predictions and Reported Rules

In total, GPT predicted a non-diabetic status (value 0) 349 times, while in 293 decisions, GPT predicted a diabetic status (value 1). According to GPT, non-diabetic status was determined with 2952 unique rules (out of 7514 rules reported in total) and diabetic status with 1904 unique rules (out of 2663 rules reported in total).

Each test sample contributes to at least three and at most five sets of rules. In theory, for all 154 test samples, we can obtain from 462 (=3 × 154) to 770 (=5 × 154) rule lists, where each rule list can hold a different number of rules. Theoretically, the most dominant scenario, where every prediction for a test sample was made in favor of a diabetic status, would result in 380 unique rules (approximately 1904/5). Comparing the count of available rules in NiaARM (350) with the reported 2952 unique rules for non-diabetic status and 1904 for diabetic status reveals a disparity.

Given the substantial variance between the rules reported by GPT and NiaARM, we will restore rule interpretability by employing a similarity score method to compare NiaARM rules with those reported by GPT. Results will be reported later.

Figure 2 displays the ten most frequently reported rules that contributed to the predicted non-diabetic status with the highest frequency according to GPT.

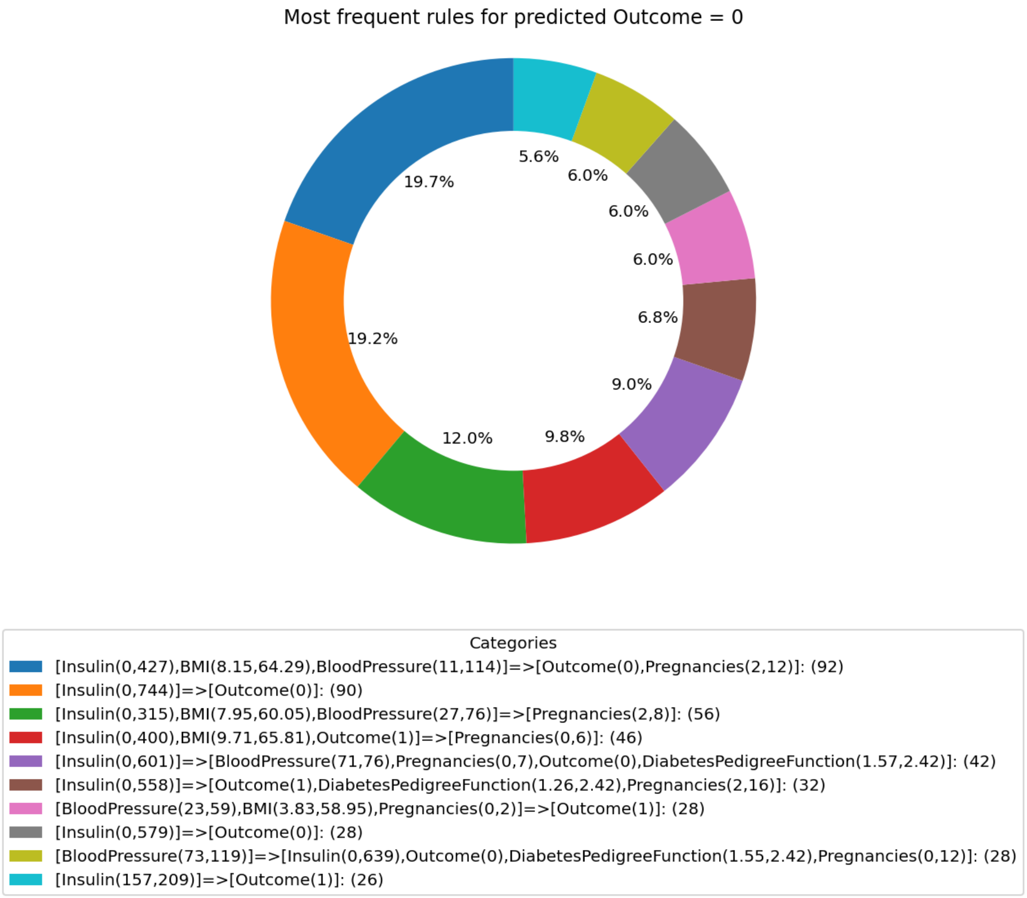

Figure 2. Ten GPT-recognized rules that contributed to predicted non-diabetic status with the highest frequency. Percentages in the circular pie chart represent the proportions among the ten most impactful rules.

GPT identified two rules that have a higher frequency of contributing to predicting non-diabetic status (see Figure 2 and in detail Table 3). Both were, on average, reported in approximately 25% of decisions favoring non-diabetic status (92/349 and 90/349), covering 1.21% (92/7514 and 90/7514) of the rules reported as impactful (see Table 3). Among the ten most impactful rules, these two rules constitute approximately 19% of the overall contribution (see Figure 2). Both rules include insulin levels as an antecedent and non-diabetic status as a consequent. However, one of them incorporates additional features as antecedents (BMI, blood pressure), resulting in the prediction of non-diabetic status and the number of pregnancies. Interestingly, the third most recognized rule does not even consist of the outcome as a feature (see Table 3). Apart from that observation, we have noticed that four other rules (4th, 6th, 7th, 10th) included only diabetic status (Outcome 1) either as an antecedent or consequent (see Figure 2). These rules were as follows:

- '[Insulin(0, 400), BMI(9.709, 65.806), Outcome(1)]⇒ [Pregnancies(0, 6)]'.
- '[Insulin(0, 558)] ⇒ [Outcome(1), DiabetesPedigreeFunction(1.255, 2.42), Pregnancies(2, 16)]'.
- 'BloodPressure(23, 59), BMI(3.832, 58.949), Pregnancies(0, 2)] ⇒ [Outcome(1)]'.
- '[Insulin(157, 209)]⇒[Outcome(1)]'.

Among the ten most impactful rules, the top three represent half (50.9%) of all generated rules that were reported by GPT (see Figure 2), which cover around 3.2% of all rules being reported as impactful for predicting non-diabetic status throughout 349 predictions (see Table 3; the third rule contributed 0.75% (56/7514)).

Regarding GPT-recognized rules that contributed to predicting diabetic status, Figure 3 displays the ten most frequently reported rules. Similar to the previous example, it should be noted that the percentages in the graph only represent the proportion among the ten most impactful rules according to the GPT.

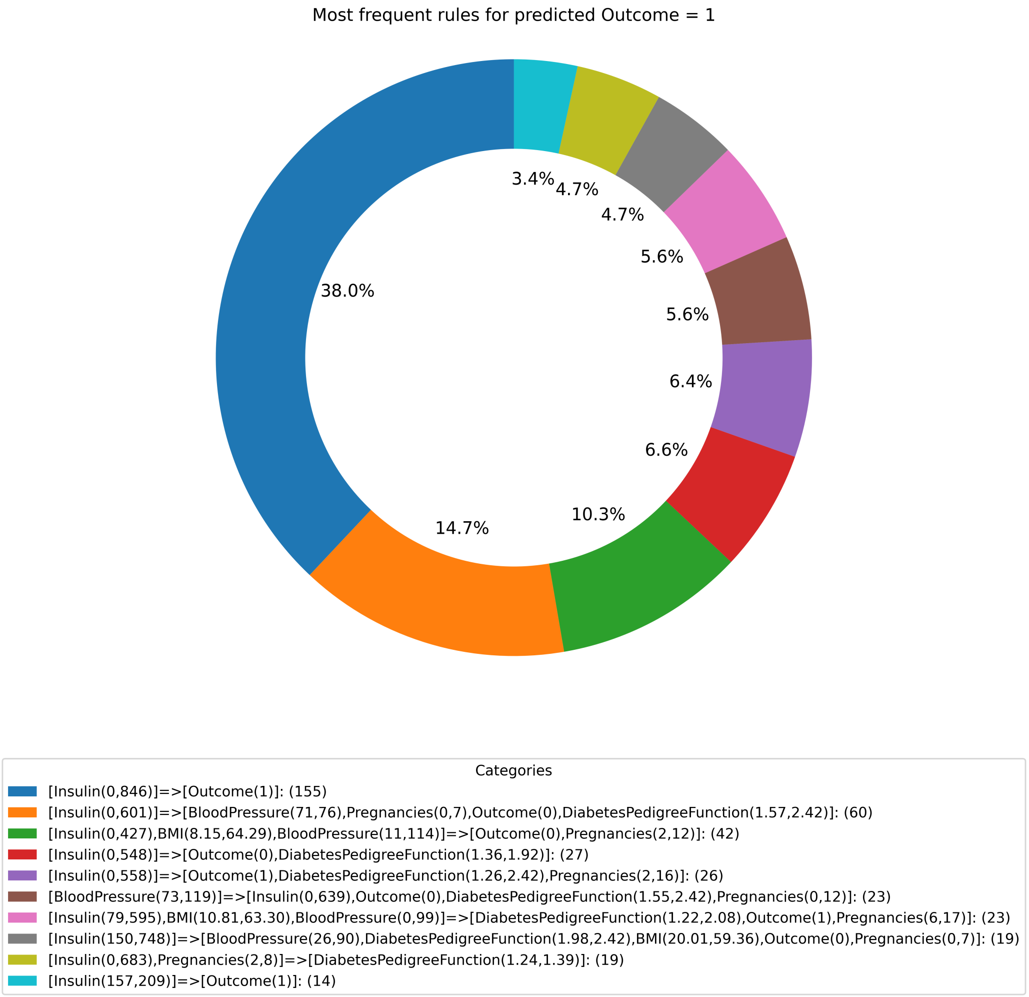

Figure 3. Ten GPT-recognized rules that contributed to predicted diabetic status with the highest frequency. Percentages in the circular pie chart represent the proportions among the ten most impactful rules.

It is evident that the most impactful reported rule dictated that all values fell within the insulin level range [0, 846]. On the dataset level, the insulin levels range from 0 to 846 mu U/mL. This rule covers around 5.8% of rules reported as impactful for predicting diabetic status throughout 293 predictions (see Table 3), meaning that this rule is reported in more than every second prediction in favor of diabetic status (52.9 %). Among the ten most impactful rules, this rule showed up to represent 38% (155/408) of decisions (see Figure 3). On average, this rule is reported in slightly over one-third (38%) of all decisions made in favor of non-diabetic status.

3.4. Computational Time Analysis

The process of searching for association rules with sufficient fitness, support, and coverage is time-consuming. On average, it took 2.57 min (95% CI: 0.42–4.72) for a dataset of size 800 × 9. The process of predicting and retrieving important rules for a single instance takes approximately 3.23 s (±0.6), repeated five times, therefore around 15 s.

3.5. Comparing and Mapping GPT-Reported Rules with NiaARM-Generated Rules

Comparing the number of rules provided by NiaARM (350 rules) with rules reported and recognized by GPT (4856 unique rules), it was apparent that GPT modifies and possibly generates new rules. That led us to an interesting challenge of comparing the rules. The exact comparison of rules returned zero matches. Then, we decided to compare exact matches between antecedents and consequents (for example, $[A(0,1), B(0,1)] \Rightarrow [C(0,1)]$ has a match with $[B(0,1), A(0,1)] \Rightarrow [C(0,1)]$). This comparison also resulted in zero matches.

Next, we looked for matches between antecedents and consequents, neglecting all values. For each match, we calculated a similarity score. In total, we obtained 57,611 matches with various similarity scores. We focused only on the ten rules that had the most frequent impact on the decision according to GPT. Brief results are displayed for both outcomes (see Table 3). Most rules have a very high similarity score (above 0.975) where the only difference was in an extra square bracket that GPT 'forgot' to add.

On the other hand, decisions of the outperforming prediction model lightGBM relied predominantly on features such as glucose, age, and BMI, followed by diabetes pedigree function, blood pressure, skin thickness, insulin, and number of pregnancies (see Figure 4).

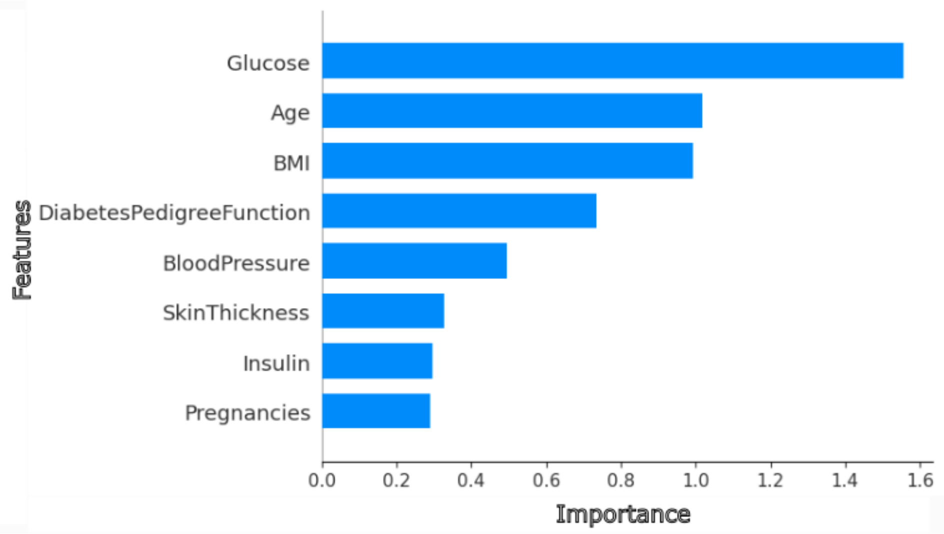

Figure 4. A ranking diagram of feature importance displayed as the average impact on model output magnitude. The bigger the average impact, the higher the importance of the feature.

Beeswarm plots display a correlation of the feature value with its contribution (Shapley additive explanations (SHAP) value) to the diagnosis outcome across all samples (see Figure 5). Glucose, the feature with the highest feature importance, displayed that high levels of glucose contribute to diabetic status (positive SHAP value), while low values contribute to non-diabetic status (negative SHAP value). Similar observations can be made in the cases of age and BMI. Low and high values of the diabetes pedigree function are distinguishable, although it is important to note that the majority of them have minimal influence on the model's output (see Figure 5). While previously mentioned features carry varying degrees of importance, blood pressure can be considered as unstable and dependent on other features.

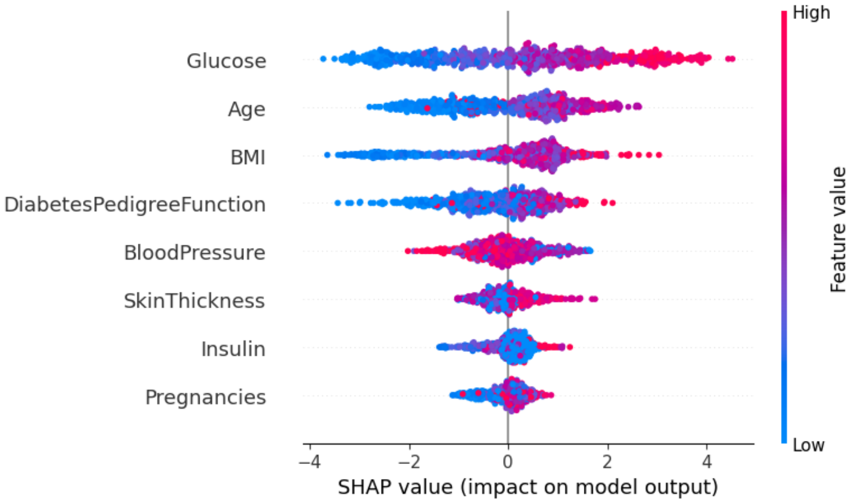

Figure 5. Feature importance displayed as an average impact on model output magnitude.

4. Discussion

Non-insulin-dependent diabetes mellitus is a chronic medical condition where high blood glucose levels occur due to the body's inability to effectively produce insulin.

Our results show that utilizing association rules for GPT-based prediction has limited efficacy in enhancing the classification of diabetic patients (accuracy: 0.721, AUC: 0.73, F1-score: 0.656, specificity: 0.7, sensitivity: 0.759). Comparing the results of previous studies, our approach of GPT rule-based predictions achieves similar scores to the K-Nearest Neighbors Algorithm (accuracy: 71.9%) and classification and regression tree (CART) (accuracy: 72.8%) [23]. It does not lag behind when compared to other methods addressing a similar problem, such as Classification Based on Associations (CBA: 72.9–76.17%) [24–26], Classification based on Multiple Association Rules (CMAR: 72.85–75.1%) [24,25], and HARMONY (72.34%) [25].

In determining the metrics to evaluate the performance of our approach, we favored AUC (area under the ROC curve) and F1-score. These metrics were selected due to their appropriateness for assessing the effectiveness of models in medical classification tasks, particularly in scenarios characterized by imbalanced datasets typical of diabetic patient classification. While accuracy is commonly used [23–26], it may not adequately capture the performance of models when classes are imbalanced. AUC provides a holistic measure of the model's ability to discriminate between diabetic and non-diabetic patients across various threshold levels [27], making it robust to class imbalances. Additionally, the F1-score, which balances precision and recall, is crucial in medical scenarios where both false positives [28] and false negatives [29] carry significant clinical implications. By utilizing AUC and F1-score alongside accuracy, we ensure a comprehensive evaluation of our approach's performance, considering different aspects of classification effectiveness.

Despite not-so-promising results, we believe that with some preprocessing steps, such as feature discretization [6], we could achieve better prediction performance. Additionally, hyperparameter tuning of NiaARM and GPT could be implemented and potentially further improve the classification performance.

The integration of GPT with association rule mining improves the clarity of the diabetes classification model by generating a collection of rules that depict the connections between input features and diabetic status. This enhances the comprehensibility of the model, facilitating easier understanding.

In our study, GPT was constrained by specific instructions, limiting its access to general knowledge and freedom to make decisions based on external sources. Improving instruction through prompt engineering could potentially enhance its diagnostic performance. During the testing phase, each test sample was provided individually to ensure that GPT produced only one answer. Although we attempted to test dozens of test samples at once, it rarely resulted in an equal number of outputs.

Apart from that, there are other limitations to this study that need to be acknowledged. Firstly, exploring the application of GPT-4 could potentially yield even more accurate results, albeit with the challenge of rule summarization due to token limitations. Secondly, the study was conducted using a specific dataset, and generalization to other domains and datasets may require further investigation. Thirdly, the dataset used in this study offers limited representativeness, as it comprises exclusively women over the age of 21, all residing in close proximity to Phoenix, Arizona, and of Pima Indian heritage [9]. This homogeneity may lead to skewed results and biased decision making. Consequently, it emphasizes the necessity for future research dedicated to mitigating dataset biases and devising strategies to improve the applicability of the approach to diverse datasets. Further, the research could extend its scope to encompass other medical conditions, evaluating the applicability of the methodology and its efficacy across diverse medical scenarios. Moreover, by incorporating laboratory results, this might improve the accuracy and comprehensiveness of association rules and GPT's decision making. Regarding association rule mining, improvements could be made to the association rule mining process, such as optimizing the parameters of the NiaARM framework.

In terms of the association rule mining approach, our findings suggest several strengths and weaknesses worth considering. The approach demonstrates its capability to generate a large number of rules, thereby providing a comprehensive representation of the training set. By leveraging support and confidence metrics, we ensured the inclusion of only relevant rules, thereby enhancing the overall quality of the generated rules. The implementation of specific criteria aimed to achieve a balance in rule selection. However, it is essential to acknowledge the weaknesses present in this approach. The complexity of the rules generated can pose challenges in interpretation due to their considerable number, making it difficult to identify the most relevant ones. Additionally, reliance on default parameters may not always yield optimal results for specific datasets, potentially leading to a suboptimal rule list.

Moreover, our study highlights the time-consuming nature of the association rule mining process. On average, it took 2.57 min for the algorithm to search for performance-wise ideal association rules within our dataset. Additionally, the process of predicting and retrieving important rules for a single instance took approximately 3.23 s, repeated five times, totaling around 15 s. These findings underscore the need for efficient optimization strategies in future research endeavors.

Regarding the integration of association rules into the GPT part, our expectations align with previous research indicating potential improvements in classification metrics with the provision of an optimal rule list. Furthermore, the ability to make predictions based on rule sets offers a promising avenue for implementing a rule-based decision-making process, which could enhance the overall predictive capabilities of the model.

Association rules, generated by NiaARM, covered both outcome classes equally as desired. This requirement ensured that our model was not biased towards any particular class, fostering balanced and fair predictions. The outcome representing diabetic status was mainly observed as an antecedent, while the outcome representing non-diabetic status was mainly observed as a consequent. As a result, we obtained lower predictive sensitivity than specificity, which does not align with our expectations. It appears that rules with an outcome of diabetic status as a consequent still outweighed the rules with a non-diabetic status as a consequent. The reason for this might lie in a rule where a broad range of insulin levels automatically determines diabetic status. Moreover, it is worth noting that this rule was reported in more than half (52.9%) of the predictions made by the GPT-based model.

Additionally, upon checking, it was observed that none of the predictions simultaneously reported both GPT-generated rules ($[Insulin(0, 744)] \Rightarrow [Outcome(0)]$ and $[Insulin(0, 846)] \Rightarrow [Outcome(1)]$) as rules that contributed to the decision. Due to the high cost of conducting such analysis, in the future, we would suggest tracking insulin and other parameters along with local interpretations at the individual level [30].

While GPT's assigned task was to utilize NiaARM rules and identify the rules it relied on, it struggled to correctly identify the rules that had the greatest impact. Moreover, it presented rules that closely resembled the actual rules generated by NiaARM. The exact reason for providing rules that resemble the actual rules could be attributed to the extensive number of provided rules, which confused the model. To address this issue, it would be advisable to present rules multiple times within various sub-prompts, utilizing a chain-of-thought prompting approach. It is a technique that enhances the capacity of LLMs to engage in intricate reasoning to a considerable extent [31]. Despite these challenges, the utilization of a similarity score method proved effective in extracting the authentic rules.

Glucose, age, and body mass index were identified as the most impactful features by lightGBM, which corresponds to the previous studies that include them in their models [8,32,33]. The same features were neglected from GPT, except for the BMI. The World Health Organization reports that diabetes is diagnosed if the two-hour plasma glucose level is at least 200 mg/dL or 11.1 mmol/L [34]. In contrast, lightGBM considered insulin to be non-essential for predicting the status, whereas GPT recognized it as one of the most influential features.

It is interesting that GPT recognized two rules that simultaneously determine both diabetic and non-diabetic status based on insulin levels, where insulin levels between 0 and 744 mu U/mL lead to a decision of diabetic and non-diabetic status. The rules are as follows: $[Insulin(0, 744)] \Rightarrow [Outcome(0)]$ for non-diabetic status and $[Insulin(0, 846)] \Rightarrow [Outcome(1)]$ for diabetic status. However, we later discovered that the first GPT-recognized rule is linked to the rule $[Insulin([281, 386])] \Rightarrow [Outcome(0)]$ and the second rule is linked to the rule $[Insulin([157, 209])] \Rightarrow [Outcome(1)]$. Patients suffering from T2DM typically exhibit increased insulin levels unless beta cell dysfunction has occurred [35–37].

Furthermore, when comparing interpretability with our baseline model, XGBoost predominantly relies on glucose levels for predicting diabetic status, with insulin levels ranking barely above the second-to-last place in terms of feature importance. This is in contrast to our approach, where insulin plays a crucial role. Regarding explainability, in XGBoost, higher values of glucose are distinctly separated from lower values, with their frequent contribution to the decision in favor of diabetic status evident. In contrast, for insulin levels, we cannot make a similar assertion.

Elevated blood sugar levels (hyperglycemia), reduced responsiveness to insulin (insulin resistance), and insufficient insulin production are the defining features of T2DM [38]. Furthermore, it is known that patients who exhibit insulin resistance and hyperinsulinemia (an elevated amount of insulin in the blood, considered as an indicator of insulin resistance) face an increased risk of developing T2DM, obesity, cardiovascular disease, cancer, and premature mortality [38,39]. The issue with our dataset is that it lacks information, such as the presence of hyperinsulinemia, and we are unsure in which stage the T2DM patients are; therefore, we cannot rely heavily on insulin levels alone.

Another observation is that GPT recognized the same rule to be impactful for predicting diabetic as well as non-diabetic status. This rule indicates that insulin levels, body mass index, and blood pressure determine non-diabetic status in females who have been pregnant more than once. One of the potential reasons for including a rule for predicting diabetic status that does not directly determine diabetic outcome is that the diabetic outcome was present in only 29% of cases when the outcome was included as a consequent. GPT's decisions may have relied more on the rules containing an outcome of 0, particularly in scenarios when feature values fell outside the range. Diabetes increases the prevalence of hypertension and elevated blood pressure [40]. Elevated blood pressure and hyperglycemia, high blood sugar levels, frequently coexist [41,42]. It was also reported that elevated blood pressure is significantly and independently associated with the development of diabetes [43,44]. Tian et al. went further and demonstrated that diabetes is correlated with arterial stiffness [45], of which the risk factors are age, elevated blood pressure, diabetes, and others [46,47].

We consider the interpretability of the GPT association rule-based model to be of utmost importance, as it helps medical professionals understand the underlying factors contributing to the classification of diabetic patients [48]. However, the concern of leaving interpretability solely to GPT is reasonable, given GPT's tendency to produce hallucinatory outputs and potentially tailor results to fit the user's context. The generated outcomes may lack real-world reliability. For that reason, the iterative steps involving the use of GPT as a prediction tool and majority voting based on multiple responses enhanced the robustness of our predictions, reducing the likelihood of uncertainties in the final outcome.

5. Conclusions

Our research explores the integration of association rules with GPT-based prediction to classify non-insulin-dependent diabetes mellitus, producing outcomes similar to those of established techniques. However, obstacles in interpretation and dataset inclusivity emphasize the need for ongoing improvement. Future efforts should prioritize refining pre-processing methods, fine-tuning hyperparameters, and improving computational efficiency to enhance clinical applicability. Collaboration between domain experts and technological advancements is crucial for advancing diagnostic precision in diabetes classification.

Author Contributions: Conceptualization, L.K. and G.S.; methodology, L.K.; software, L.K. and I.F.J.; validation, L.K.; formal analysis, L.K.; investigation, L.K.; resources, I.F.J.; writing—original draft preparation, L.K.; writing—review and editing, I.F.J. and G.S.; visualization, L.K.; supervision, I.F.J. and G.S.; project administration, G.S.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Slovenian Research Agency under Grants ARRS P2-0057 and ARRS N3-0307.

Institutional Review Board Statement: Ethical review and approval were waived for this study, because the study uses existing data.

Informed Consent Statement: Informed consent not needed, since the data used in this study are publicly available and previously published.

Data Availability Statement: The data used in our study are publicly accessible through the Kaggle repository at https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed on 18 January 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
API	application programming interface
AUC	area under the curve
BMI	body mass index
CART	classification and regression tree
EFB	exclusive feature bundling
GOSS	gradient-based one-side sampling
GPT	generative pretrained transformer
NLP	natural language processing
SD	standard deviation
SHAP	Shapley additive explanations
T2DM	Type 2 diabetes mellitus

References

1. Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.; Mbanya, J.C.; et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119. [CrossRef]
2. Oh, S.H.; Lee, S.J.; Park, J. Precision medicine for hypertension patients with type 2 diabetes via reinforcement learning. *J. Pers. Med.* **2022**, *12*, 87. [CrossRef]
3. Malerbi, F.K.; Andrade, R.E.; Morales, P.H.; Stuchi, J.A.; Lencione, D.; de Paulo, J.V.; Carvalho, M.P.; Nunes, F.S.; Rocha, R.M.; Ferraz, D.A.; et al. Diabetic retinopathy screening using artificial intelligence and handheld smartphone-based retinal camera. *J. Diabetes Sci. Technol.* **2022**, *16*, 716–723. [CrossRef]
4. Tao, R.; Yu, X.; Lu, J.; Wang, Y.; Lu, W.; Zhang, Z.; Li, H.; Zhou, J. A deep learning nomogram of continuous glucose monitoring data for the risk prediction of diabetic retinopathy in type 2 diabetes. *Phys. Eng. Sci. Med.* **2023**, *46*, 813–825. [CrossRef]
5. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef]
6. Patil, B.; Joshi, R.; Toshniwal, D. Association rule for classification of type-2 diabetic patients. In Proceedings of the 2010 Second International Conference on Machine Learning and Computing, Bangalore, India, 9–11 February 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 330–334.
7. Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **2020**, *10*, 11981. [CrossRef]
8. Deberneh, H.M.; Kim, I. Prediction of type 2 diabetes based on machine learning algorithm. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3317. [CrossRef]
9. Smith, J.W.; Everhart, J.E.; Dickson, W.; Knowler, W.C.; Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*; American Medical Informatics Association: Washington, DC, USA, 1988; p. 261.
10. Stupan, Ž.; Fister, I. NiaARM: A minimalistic framework for Numerical Association Rule Mining. *J. Open Source Softw.* **2022**, *7*, 4448. [CrossRef]
11. Fister, I., Jr.; Podgorelec, V.; Fister, I. Improved nature-inspired algorithms for numeric association rule mining. In *Intelligent Computing and Optimization, Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020), Hua Hin, Thailand, 8–9 October 2020*; Springer: Cham, Switzerland, 2021; pp. 187–195.
12. Kaushik, M.; Sharma, R.; Peious, S.A.; Shahin, M.; Ben Yahia, S.; Draheim, D. On the potential of numerical association rule mining. In Proceedings of the International Conference on Future Data and Security Engineering, Quy Nhon, Vietnam, 25–27 November 2020; Springer: Singapore, 2020; pp. 3–20.
13. Zaki, M.J. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* **2000**, *12*, 372–390. [CrossRef]
14. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago de Chile, Chile, 12–15 September 1994; Volume 1215, pp. 487–499.
15. Kaushik, M.; Sharma, R.; Fister, I., Jr.; Draheim, D. Numerical association rule mining: A systematic literature review. *arXiv* **2023**, arXiv:2307.00662.
16. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, 3146–3154.
17. Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.; Biecek, P. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *J. Mach. Learn. Res.* **2021**, *22*, 1–7.
18. Molnar, C. *Interpretable Machine Learning*; Lulu.com: Morrisville, NC, USA, 2020.
19. OpenAI Platform—platform.openai.com. Available online: https://platform.openai.com/account/rate-limits (accessed on 28 September 2023).
20. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
21. Storn, R.; Price, K. Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **1997**, *11*, 341–359. [CrossRef]
22. Safari, S.; Baratloo, A.; Elfil, M.; Negida, A. Evidence based emergency medicine; part 5 receiver operating curve and area under the curve. *Emergency* **2016**, *4*, 111.
23. Šter, B.; Dobnikar, A. Neural networks in medical diagnosis: Comparison with other methods. In Proceedings of the International Conference on Engineering Applications of Neural Networks, London, UK, 17–19 June 1996; pp. 427–430.
24. Li, W.; Han, J.; Pei, J. CMAR: Accurate and efficient classification based on multiple class-association rules. In Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001; IEEE: Piscataway, NJ, USA, 2001; pp. 369–376.
25. Wang, J.; Karypis, G. HARMONY: Efficiently mining the best rules for classification. In Proceedings of the 2005 SIAM International Conference on Data Mining, Newport Beach, CA, USA, 21–23 April 2005; pp. 205–216.
26. Ma, B.; Liu, B.; Hsu, Y. Integrating classification and association rule mining. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 August 1998; Volume 50, pp. 80–86.
27. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]

28. Drew, B.J.; Harris, P.; Zègre-Hemsey, J.K.; Mammone, T.; Schindler, D.; Salas-Boni, R.; Bai, Y.; Tinoco, A.; Ding, Q.; Hu, X. Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PLoS ONE* **2014**, *9*, e110274. [CrossRef] [PubMed]
29. Kaukonen, K.M.; Bailey, M.; Pilcher, D.; Cooper, D.J.; Bellomo, R. Systemic inflammatory response syndrome criteria in defining severe sepsis. *N. Engl. J. Med.* **2015**, *372*, 1629–1638. [CrossRef] [PubMed]
30. Kopitar, L.; Cilar, L.; Kocbek, P.; Stiglic, G. Local vs. global interpretability of machine learning models in type 2 diabetes mellitus screening. In *International Workshop on Knowledge Representation for Health Care*; Springer: Cham, Switzerland, 2019; pp. 108–119.
31. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
32. Joshi, R.D.; Dhakal, C.K. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7346. [CrossRef] [PubMed]
33. Seah, J.Y.H.; Yao, J.; Hong, Y.; Lim, C.G.Y.; Sabanayagam, C.; Nusinovici, S.; Gardner, D.S.L.; Loh, M.; Müller-Riemenschneider, F.; Tan, C.S.; et al. Risk prediction models for type 2 diabetes using either fasting plasma glucose or HbA1c in Chinese, Malay, and Indians: Results from three multi-ethnic Singapore cohorts. *Diabetes Res. Clin. Pract.* **2023**, *203*, 110878. [CrossRef]
34. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed.; International Diabetes Federation: Brussels, Belgium, 2021.
35. Bogardus, C.; Lillioja, S.; Howard, B.; Reaven, G.; Mott, D. Relationships between insulin secretion, insulin action, and fasting plasma glucose concentration in nondiabetic and noninsulin-dependent diabetic subjects. *J. Clin. Investig.* **1984**, *74*, 1238–1246. [CrossRef]
36. Reaven, G.M. Compensatory hyperinsulinemia and the development of an atherogenic lipoprotein profile: The price paid to maintain glucose homeostasis in insulin-resistant individuals. *Endocrinol. Metab. Clin.* **2005**, *34*, 49–62. [CrossRef]
37. Clark, A.; Jones, L.C.; de Koning, E.; Hansen, B.C.; Matthews, D.R. Decreased insulin secretion in type 2 diabetes: A problem of cellular mass or function? *Diabetes* **2001**, *50*, S169. [CrossRef] [PubMed]
38. Santoleri, D.; Titchenell, P.M. Resolving the paradox of hepatic insulin resistance. *Cell. Mol. Gastroenterol. Hepatol.* **2019**, *7*, 447–456. [CrossRef]
39. Tsujimoto, T.; Kajio, H.; Sugiyama, T. Association between hyperinsulinemia and increased risk of cancer death in nonobese and obese people: A population-based observational study. *Int. J. Cancer* **2017**, *141*, 102–111. [CrossRef]
40. Jia, G.; Sowers, J.R. Hypertension in diabetes: An update of basic mechanisms and clinical disease. *Hypertension* **2021**, *78*, 1197–1205. [CrossRef]
41. Przezak, A.; Bielka, W.; Pawlik, A. Hypertension and Type 2 Diabetes—The Novel Treatment Possibilities. *Int. J. Mol. Sci.* **2022**, *23*, 6500. [CrossRef] [PubMed]
42. Kim, H.J.; Kim, K.i. Blood pressure target in type 2 diabetes mellitus. *Diabetes Metab. J.* **2022**, *46*, 667–674. [CrossRef]
43. Kim, M.J.; Lim, N.K.; Choi, S.J.; Park, H.Y. Hypertension is an independent risk factor for type 2 diabetes: The Korean genome and epidemiology study. *Hypertens. Res.* **2015**, *38*, 783–789. [CrossRef]
44. Lee, W.Y.; Kwon, C.H.; Rhee, E.J.; Park, J.B.; Kim, Y.K.; Woo, S.Y.; Kim, S.; Sung, K.C. The effect of body mass index and fasting glucose on the relationship between blood pressure and incident diabetes mellitus: A 5-year follow-up study. *Hypertens. Res.* **2011**, *34*, 1093–1097. [CrossRef] [PubMed]
45. Tian, X.; Zuo, Y.; Chen, S.; Zhang, Y.; Zhang, X.; Xu, Q.; Wu, S.; Wang, A. Hypertension, arterial stiffness, and diabetes: A prospective cohort study. *Hypertension* **2022**, *79*, 1487–1496. [CrossRef] [PubMed]
46. Boutouyrie, P.; Chowienczyk, P.; Humphrey, J.D.; Mitchell, G.F. Arterial Stiffness and Cardiovascular Risk in Hypertension. *Circ. Res.* **2021**, *128*, 864–886. [CrossRef] [PubMed]
47. Laurent, S.; Boutouyrie, P. Arterial Stiffness and Hypertension in the Elderly. *Front. Cardiovasc. Med.* **2020**, *7*, 544302. [CrossRef]
48. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1379. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.