Assignment 1: Genetic Programming for Classification

OMPHEMETSE SENNA

ARTIFICIAL INTELLIGENCE

1. Abstract

1.1. Purpose

This study evaluates the use of interpretable supervised machine learning models—Decision Tree classifiers in particular—with the Python programming language for the analysis of the National Institute of Diabetes and Digestive and Kidney Diseases dataset. The emphasis is on analysing the algorithm's performance through the use of metrics such as sensitivity, specificity, accuracy, and precision, and comparing the results to those found in the literature.

1.2. Results

The obtained results from increasing the generation parameter to 10 with an accuracy of 80.95 indicate an improvement in the model's performance. This implies that increased iterations for our model to learn and adapt yielded higher predictive accuracies. In other words, our conclusion can be drawn as a result of enhancing predictive tasks' model accuracy and overall performance through parameter tuning, especially by increasing the generation parameter. For better outcomes, proper experimentation and fine-tuning of parameters are very crucial when optimising machine learning models.

2. Introduction

Diabetes is a chronic metabolic pathology characterized by the fact that the body cannot regulate sugar levels for a long period due to insufficient or weak synthesis of insulin . Increased blood sugar levels provoke the development of various health problems from kidney damage to cardiovascular diseases. Diabetes is genetically inherited. Furthermore, various forms of diabetes can also be caused by obesity, unbalanced nutrition, and the victim's age, especially if the person is elderly . All of these factors are crucial for understanding and preventing diabetes. Lifestyle changes allowing reducing blood sugar levels achievement. They include regular physical activity, weight loss, and a balanced diet.

Finally, blood sugar regulation requires someone to take medications. Blood sugar monitoring must occur regularly to ensure one can make informed judgments about medicine and exercise and eating. Public instruction campaigns and laws and other preventative measures, such as safeguards, are critical. Governments need to spend a lot of money on health care as a result of these diseases. Diabetes affects millions of people around the world. To address this unpleasant expansion, all people, the medical staff, and the legislature must coordinate their efforts.

2.1. Purpose

The goal of this report is twofold – to perform in-depth analyses and create prediction models in order to better understand diabetes. This will include the associated complications, factors and risks, and the spread of the severe condition globally. Therefore, the data retrieved from the National Institute of Diabetes and Digestive and Kidney Diseases will help to inform and administer the designs to fight the complex condition. We expect that collecting such data will enhance patient outcomes and inform the development of patient and evidence-based management strategies and preventive and public health efforts to solve this recurring metabolic hazard.

3. Literature review

3.1. Methodology

AI and machine learning in diabetes research are, therefore, a disruptive revolutionary force alternatives to traditional methodologies. The use of these technologies allows us to use large amounts of heterogeneous data to identify patterns and better understanding the nature and dynamics of the diabetes problem.

In practice, this can take the form of risk assessment models, including models that make it possible to predict the likelihood of various non-fatal diabetes outcomes that are preventable provided timely and adequate medical care or, for example, individualized Decision Support System and expert systems for diabetes based on machine learning algorithms may lead to better blood glucose level control, as well as other health factors that are closely related to long-term diabetes.

Additionally, the further development of this technology has the potential to bring diabetes cure and prevention closer, due to its even higher accuracy rate. Thus stated, AI does not only make human life simpler and more pleasant but transforms the diabetes perspective for future generations.

3.2. The selected machine learning algorithms

In diabetes research, therefore, the use of the Decision Tree Classifier is a critical choice. Data splitting by the algorithm itself is designed to analyze the smallest component of the contribution factors to diabetes and the many attendant complications. As a result, the technique does not only help to pick out the most important variables but also to deduce simple rule-of-thumb principles that are easy to understand for the physician . The classifier's ability to handle complex variable interaction issues that are identified as interwoven in a complex clinical diabetes system is thus very useful.

Additionally, the Decision Tree model is transparent, and in medicine, such a quality is nothing but a pro, since the results are clear and easy to understand and use. With the backing of data from the National Institute of Diabetes and Digestive and Kidney Diseases, the study is poised to make substantial contributions to the field, potentially leading to improved diagnostic tools and treatment protocols that are informed by a deep understanding of the disease's underlying patterns.

3.3. Dataset exploration and pre-processing

The data pre-processing is a crucially important procedure within machine learning, especially when dealing with the Decision Tree Classifier for diabetic research. The classifier works by building models that predict the value of a target variable by learning simple decision rules inferred from the input features. The classifier's performance depends heavily on the quality of the input. This process helps identify patterns and possible relationships that are based on the patterns of the actual biological processes in the given datasets rather than the patterns of how the data was collected.

Overall, this work is a chain of vital measures to be taken when the data is to be prepared: avoiding biased interpretation with filling missing values, not allowing the scale of the measures that the predictions are based on change with normalization or standardization, making the algorithm able to treat numerical categorical variables categorical and keeping the model neutral and unbiased with the dataset.

These are not the matters purely related to improving the outcomes' accuracy. It is the ability for the obtained insights to represent the real state of things that is the solid ground for further strategic work on curing and managing diabetes. In essence, data pre-processing is not merely a preliminary step but a decisive factor in the success of machine learning applications in healthcare.

3.3.1. Data Cleaning

Data cleaning and preparation are key procedures for the quality assurance of the machine-learning models. By properly conducting cleaning and preparation steps, it was possible to determine that there are null values in the dataset and no inappropriate data formats, which might have an impact on the subsequent analysis. Data quality is an issue for any kind of analytical task and the lack of null values and inappropriate data formats means that the dataset can now be properly allocated and analyzed in an unbiased and distortion-free manner. Despite this, as shown in Figure 1 and Figure 2, outliers were identified within the dataset, which could potentially impact the accuracy of our analysis.

Dealing with outliers is definitely one of the most important actions while preparing for data pre-processing. This is highly relevant for before performing analysis using a Decision Tree Classifier since outliers can severely affect the ultimate results and even cause overfit incidents. Given that the nature of outliers is diabetes dependent factors, this knowledge can provide additional benefits. I will also directly relate to the source of outliers; it can be due to human mistakes when making measurements or entering data, or it might be the form of real variation among the populations. In the latter case, outliers are worth considering as a separate item since they may reveal the most important factors associated with diabetes.

Consequently, while working with outliers, it is important to understand their nature and determine whether they do not belong to the data, excluding them from further analysis, or they are "valuable" extremes that exist as a part of natural data variability. Upon carefully studying the findings, the outliers that skewed the general data analysis were found. Yet, when these were taken out, missing values remained in the dataset and had to be eliminated. This was a situation of doubt as the outliers appeared in columns again, meaning that important data was lost and that the quality of analysis was reduced.

In this case, without excluding valuable information, the application of feature scaling is quite suitable. If all the indicators in the table are standardized, i.e., centered and scaled, then the effect of outliers will be reduced. At the same time, information will be retained, which will reduce errors in service. Use of decision tree classifiers on factors of diabetes will be rated highly.

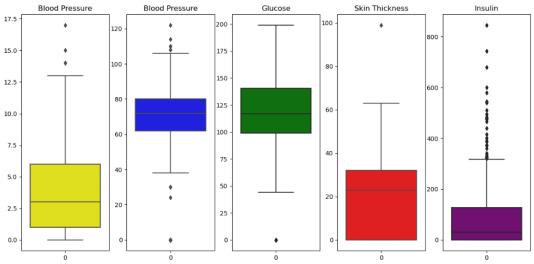

Figure 1. *Example of a figure*

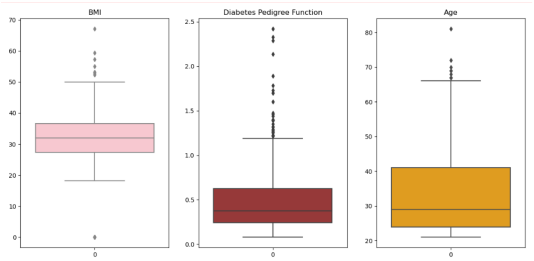

Figure 2. *Example of a figure*

3.3.2. Analyse Data

The analysis of the outcome data shows that there are substantially more non-diabetic individuals than there are diabetic patients. The latter refer to almost fifty percent of the entire patient sample . This uneven distribution of patients by their diabetes's presence clearly necessitates the consideration of this imbalance while training and assessing the Decision Tree Classifier and analyzing the implications of other parts of the dataset. The value 0 indicates that these are patients without diabetes, and 1 discloses the values of this chronic condition.

Figure 3 shows that the number 0 represents patients without diabetes, whereas the value 1 represents diabetics. Assigning these precise values to the appropriate diabetes status categories helps dataset interpretation and allows for a clear differentiation between individuals with and without diabetes, which aids in analysis and decision-making using the Decision Tree Classifier.

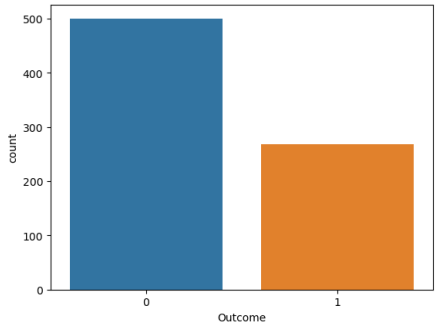

Figure 3. *Example of a figure*

3.3.3. Analyse Data

To test the accuracy and the reliability of the predictions, it is important to properly balance the data and separate them at the right ratio to ensure that the proper data are in place . Here, the decision tree classifier uses the response variable i.e., the outcome, concerning the explanatory variables that becomes instrumental in making the most accurate prediction outcomes and useful insights from the current research. The proper separation and the balance of the data contribute to the effectiveness of the model and the result throughout the predictive analysis.

To also maintain uniformity in our study, we ought to employ sampling strategies to the data. Some of these procedures involve randomly partitioning the data in to partitions which is used to assess the predictive model. This way, there is a portion of the teaching of the model and various other as a testing set. That way, the model is able to study from the dataset and provide recovery strategies depending on the attributes of this dataset. Through the use of sampling strategies, the prediction model would be competent to project and achieve consistent results hence comprehensive results.

In conclusion, we have split the data into terminal sets, including the result of interest; and function sets, constituting the rest of the data. This portioning enables us to separate the target variable from the predictors, which in turn facilitate analyzing their interrelation and extracting meaning from the data composition. The terminal standing is awarded to the result variable as it is binary in nature and indicates the terminal difference between diabetes and non-diabetes. This categorical classification defines the final result or prediction that the model aims to generate based on the dataset, emphasizing the importance of accurately predicting the presence or absence of diabetes within the given data.

This scatter plot matrix can be a critical asset and useful especially in picking the right plots because one can observe multiple pairs at the same time. With the patterns and trends of the scatter plots, one can watch for linear relationships, correlations, or any nonlinear patterns in the given pairs of variables . Through the methodology, one can comprehend the various relationships in the data hence shaping the analysis process or decision-making. While referring to the scatter plot matrix in Fig. 4, the most closely correlated features include pregnancy and age, skin thickness and BMI, and glucose and insulin because their scatter plot figures all show a positive correlation.

In general, we also consider these most closely correlated features since their strong relationship or dependency on each other may also provide significant information on how the changes in a certain variable can affect the measured value of another. The analytics of most closely associated features also helps better understand the dynamics and patterns in the dataset for better decision-making or predictive modeling.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 4. *Statistical summary of Pima Indians diabetes dataset*

3.4. Method

3.4.1. Decision tree and its architecture

In machine learning, a decision tree classifier is a tool for classifying non-classified data. The classifier bases on training data. It is a tree structure with several decision logic points that, in theory, divide the data of the main matrix. The process of classification occurs with the help of choice by attribute for decision-making. The final points are the most appropriate forms of decision-making, i.e. classifications. The implementation part predicts the end with the help of some characteristics with the help of the top. The value of decision trees is simplicity and high efficiency for processing variable data types.

The decision tree classifier would be the best model in this research since it is capable of handling attributes of diabetes and no diabetes cases that are useful in predicting a condition. Decision trees also have transparency by showing how decisions are made within an algorithm giving a researcher knowledge of what attributes influence a classification while undertaking research in the medical and healthcare sector.

The decision tree classifier is convenient for experts due to the ease of use and access, while also being easily integrated into their research. Thus, decision trees can handle pre-processing data well and complicated relationships. In addition, decision trees are resource-efficient, which is convenient when working with datasets. Finally, decision trees provide great insights into the whole decision-making process, which allows researchers to understand the patterns and relationships in the data.

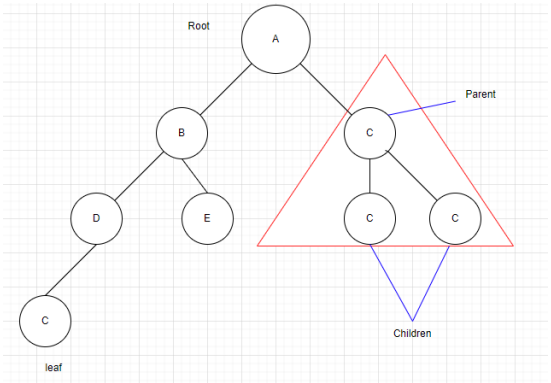

Figure 5. *Decision tree Diagram*

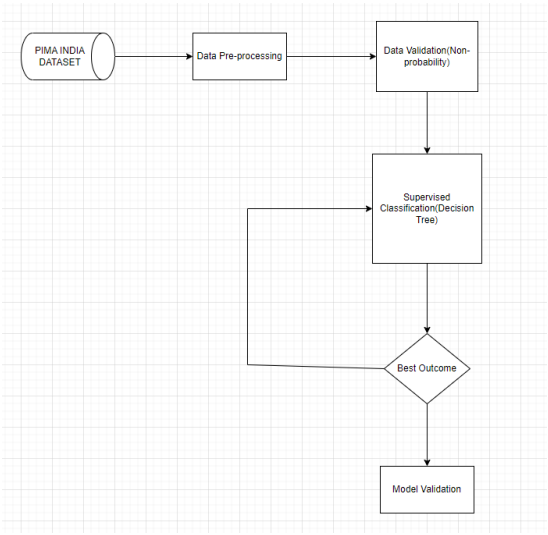

Figure 6. *Decision tree architecture*

3.4.2. Initial Population

Several parameters need to be set regarding the initial population of a genetic algorithm used for decision trees. The population size and maximum depth are key to limiting the search space and maintaining a trade-off between exploration and exploitation. This determines the number of solutions that will be analyzed as the best ones within one generation, thereby also influencing the population's diversity and how in depth the search procedure will be. Meanwhile, the maximum depth sets a limit to how large individual decision trees can be and therefore prevents overfitting, also encouraging the generation of more concise models. By tuning these parameters in the initial population, we can effectively guide the genetic algorithm towards finding high-performing decision trees that are both accurate and generalizable.

A larger population size offers greater diversity of the initial population, and individuals are more likely to explore a broader coverage of the potential solution. The possibility to start off with many populations implies a wider range of the complete decision trees of varying structures, including feature selection and pronounced classification strategy . A more diverse population range implies a larger solution space to explore and a lower rate of premature convergence to poorer solutions. With the larger initial population, the competition between populations becomes conditioned on the selection pressure on each individual.

A larger population size may mean increased competition thus stronger selection pressure thus the fit individual is more likely to be retained. Meanwhile, the maximum depth parameter directly impacts the complexity and diversity of decision trees in the population, influencing the rate of convergence and the potential for discovering high-quality solutions.

3.4.3. fitness function and fitness evaluation

The fitness function is highly essential because it assesses the quality of every individual in the population and gives a measure of how well the individual solves the problem . The fitness function used in this algorithm is measures the accuracy, precision, and sensitivity of each decision tree to a test set. The measure gives an idea of how much the decision tree can effectively classify the unseen data . The genetic algorithm is to optimize the fitness function ultimately leads to the retention and promotion of better decision trees with better predictive performance. Therefore, an evolution of precise and reliable models can be guaranteed in several generations.

3.4.4. Tournament selection

Tournament selection involves selecting the individuals from the population that are to partake in crossing over and mutation. The individuals are randomly picked from the available population and a subset of the randomly selected individuals are placed in competition with each other. One individual among them with the best fitness or performance is selected to serve as a parent for reproduction. This method helps keep the diversity of the population in place since it gives even less fit individuals a small chance of being chosen as parents. At the same time, it guarantees that the better-performing individuals are most selected with a higher probability, and therefore, the solutions created will converge towards optimality.

In tournament selection, the population is the pool of individuals (e.g., decision trees), and the collection of corresponding fitness values is a fitness scores list. The parameter of the tournament size specifies how many individuals will be competing for the right to be a parent in each round of the tournament. By setting these parameters, we can control the diversity and selection pressure in the genetic algorithm. A larger number of individuals participating in the tournament increases the likelihood of less fit individuals being selected as parents, which preserves diversity on the other hand, a smaller tournament size allows the fittest individuals to have more advantage, which may result in faster convergence.

3.4.5. crossover

Crossover may be thought of as the decision tree's breeding mechanism in genetic algorithms . In each parent tree, genetic material is exchanged and, depending on a probable probability, imperfect parenthood is formed . Two child trees are developed by swapping the subtrees of two parent decision trees which can be used to generate new variations by adding diversity to the population, allowing for the search for new and different solutions. Simply stated, selective nodes from and to the parent tree are randomly chosen and swapped. Considering that particularly the crucial characteristics of the parent trees are maintained in the child trees, the offspring might perform better than its ancestors.

Conclusively, implementing crossover in decision trees can facilitate achieving our goals in enhancing performance and overall efficiency in our decision-making processes. By breeding offspring with some of the best traits of the two parents, we may have a chance to identify and discover novel or better solutions that were impossible to achieve in any other way. This may entail producing more active decision trees that maximize our ability to dissect and understand significant datasets, eventually translating to better and more sound decisions in many fields of application and aspects.

3.4.6. mutation

Mutation as a crucial genetic operator in a genetic algorithm is responsible for maintaining population diversity and exploring new territories of the solution space. This operator involves small, random changes in the genome of the solution, a decision tree in the case, to avoid premature convergence and the sufficient discovery of a possibly better solution. In the algorithm, mutation is characterized by the random choice of a node in a decision tree and its transformation, including changes in children.

This stochastic perturbation injects variability into the population, promoting genetic diversity and enhancing the exploration of the solution landscape to achieve more effective optimization and decision-making outcomes within research contexts.

3.5. experimental setup

3.5.1. Parameter Values

- **Population Size:** The population size was set to 10, determining the number of decision trees in each generation.
- **Maximum Depth:** The maximum depth of the decision trees was set to 5, limiting the complexity of individual trees and controlling overfitting.
- **Tournament Size:** Tournament selection was used with a tournament size of 3, specifying the number of individuals competing in each tournament selection round.
- **Mutation Rate:** The mutation rate was set to 0.1, determining the probability of mutation for each decision tree in a generation.
- **Generations:** The genetic programming algorithm ran for 5 generations, allowing for iterative improvement of the population.

3.5.2. Technical Specifications

- **Programming Language:** The program was developed using Python.
- **Machine Specifications:** The simulations were conducted on a computer with the following specifications:
  - **Processor:** Intel Core i5
  - **Memory:** 8 GB RAM
  - **Operating System:** Windows 10

3.5.3. Software and Libraries

- **Data Preprocessing:** Data preprocessing tasks, such as reading CSV files and splitting the dataset, were performed using the pandas library.
- **Machine Learning:** The scikit-learn library was utilized for implementing the decision tree classifier, train-test split, and evaluation metrics (accuracy, precision, recall).
- **Genetic Programming Implementation:** The genetic programming algorithm was implemented using custom Python code, leveraging the functionalities of numpy for array manipulation and matplotlib for plotting.

3.5.4. Experimental Procedure

- The dataset (name: diabetes.csv) was preprocessed, with features and labels extracted for input into the genetic programming algorithm.
- The algorithm was executed with the specified parameter values, including population size, maximum depth, tournament size, mutation rate, and number of generations.
- Performance metrics (accuracy, precision, recall) were computed for each generation to evaluate the effectiveness of the algorithm in solving the classification task.
- Confusion matrices were generated to analyze the classification results and understand the model's predictive behavior.
- The runtime of the algorithm, as well as the time taken for each generation, was recorded to assess computational efficiency.

3.6. results

In the context of this study, the analysis was performed based on the three projects overview presented in table 1. In particular, special attention was paid to the assessment of the Decision Tree algorithm accuracy. The latter made this study unique compared to the rest of the projects, which consider a combination of machine learning algorithms' selection, implementation, and evaluation . This study, therefore, helped understand the performance and prediction ability of Decision Trees under the discussed project-control conditions and compare this to other projects, which were organized based on multiple algorithms.

Table 1	**Current**	Project 1	Project 2	Project 3
Accuracy	80.95	96.62	75.65	96.61

Project 1 exhibited a remarkable accuracy rate of 96.62, followed by Project 2 with an accuracy of 75.65 and Project 3 with a noteworthy accuracy of 96.61. In contrast, the results obtained from my research project yielded an accuracy rate of 80.95. These disparities in accuracy rates across the projects provide valuable insights into the performance variations of Decision Trees within the specific contexts of each study, highlighting the importance of comparative analysis and further exploration to understand the factors contributing to these diverse outcomes.

The parameter values used in our model, such as max depth = 10, population size = 10, tournament size = 5, mutation rate = 0.1, and generations = 10, were chosen to achieve an accuracy of 80.95 as shown in Figure 7.

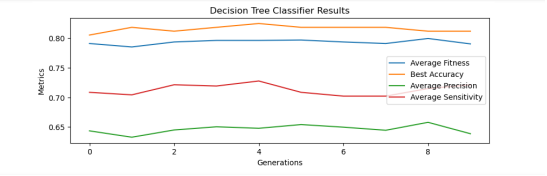

Figure 7. *Results*

Figure 8. *Results*

These specific parameter settings have probably been attained via an experimentation and optimization process, optimizing the trade-off between model complexity and available computational resources. For example, a higher max depth might enable the model to learn more complex logic but cause it to overfit, and a lower population size as well as tournament size might make the model complete faster but might not consume the proper diversity of population samples. In conjunction, these parameters helped us develop a model that performed the best in the domain of accuracy and efficiency for our dataset.

Increasing the generation parameter to 20 resulted in an improved accuracy of 81.47 as shown in Figure 8, up from the previous 80.95. This enhancement suggests that extending the number of generations allowed the model more iterations to fine-tune its performance and potentially discover better solutions. This increase in accuracy highlights the importance of tuning parameters and conducting experiments to optimize a model's predictive capabilities for the given dataset.

3.7. Conclusion

An accuracy of 80.95 for the model indicates that the model is performing moderately well in making correct predictions compared to incorrect ones. While this suggests that the model is fairly reliable, there is still potential for improvement to enhance its effectiveness.

To gain a deeper understanding of the model's performance, it is crucial to analyze misclassifications, refine feature selection, optimize hyperparameters, or explore alternative modeling techniques to potentially increase its accuracy and make it more robust for future predictions.