60th International Scientific Conference on Information, Communication and Energy Systems and Technologies, Ohrid, North Macedonia, June 26–28, 2025

Data Visualization and Prediction for Healthcare Purpose Using Random Forest Algorithm

Evgenija Gjurovska[1], Snezana Savoska[2], Natasha Blazheska-Tabakovska[3] and Kostandina Veljanovska[4]

Abstract –**The paper aims to use data visualization and prediction techniques for diabetes using the Random Forest (RF) algorithm, enhancing the healthcare analytics process. It intends to explain the ways of using data visualization to interpret patterns and explains how RF algorithm helps in creating a model for diabetes prediction.**

Keywords – **Data Visualization, Random Forest model, Diabetes prediction, Machine learning.**

I. Introduction

Data visualization techniques play a crucial role in this study by providing clear insights into the relationships between different medical parameters and diabetes risk. Graphs, heat maps, and other visual representations help in understanding patterns, identifying key risk factors, and improving model interpretability, making the prediction process more transparent and actionable for healthcare professionals. Diabetes is recognized as a major concern in the public health society and as a global epidemic. They treated diabetes as a chronic condition that is caused by the insulin production deficit or the body's inability to use insulin effectively. Prolonged high blood sugar levels associated with diabetes can lead to many complications, including eyes damage as well as kidneys, heart, blood vessels, and nerves damage. This study focuses on utilizing machine-learning algorithms, specifically the RF algorithm, to predict an individual's likelihood of having diabetes based on medical data [1]. The objective is to develop a reliable prediction model that has to predict whether a patient has diabetes by analyzing particular diagnostic parameters measured and saved in the dataset. In addition, various techniques will be explored to enhance the model's accuracy and performance, ensuring its effectiveness in medical decision-making. The RF algorithm, applied to a dataset containing various patient records, demonstrated strong predictive capabilities. The model successfully classified individuals as diabetic or non-diabetic with a high accuracy rate, offering a robust and interpretable tool for healthcare professionals to support early diagnosis and intervention [2].

Recently, the usage of big data in the healthcare industry has become increasingly prevalent. If healthcare professionals use data visualization, they can provide valuable information by analyzing large amounts of data. In this way, they can improve patient information and healthcare workers' knowledge, identify trends, and decide for providing new treatments [2]. Python is used as emerging, easy to use, and most popular programming language for processing big data used in medicine due to its variety, variability, versatility, and powerful visual data analysis capabilities. Python comes with a wide range of libraries and frameworks, designed particularly for data analysis, visualization, and machine learning (ML) [3]. Python's libraries empower healthcare analytics and professionals with powerful tools needed for large amounts of data analysis, including medical images and clinical trial results. The goal of data visualization is quite clear: to make sense of data and use the information for the benefit of the organization. Thus, data gains greater value when visualized. Without visualization, it is challenging to quickly communicate findings from data and identify patterns to extract insights and seamlessly interact with the data [4]. This paper presents the importance and need for data visualization, with the aim of reducing the time spent understanding data by maximizing the use of data to make key decisions and make future predictions as well as identify possible problems. One of the goals of the paper is to highlight the need to derive indicators of the success of visual representation that would enable the reuse of the models that are applied.

According to American Diabetes Association (ADA), standards "Diabetes mellitus refers to a collection of metabolic disorders characterized by elevated blood sugar levels (hyperglycemia) resulting from deficiencies in insulin action, insulin secretion, or both" [5]. International Diabetes Federation (IDF), in 2017, stated that 425 million individuals worldwide were living with diabetes. By 2019, this number increased to 463 million adults aged 20 to 79 years. IDF then became alarmed that diabetes has become most significant global health crisis in the 21st century [5]. Nowadays, the number of people who live with diabetes increases all the time.

Looking at the healthcare system globally, a lot of data is generated on a daily basis, more than at any time in history. Healthcare today is an evidence-based practice and requires that information be used quickly and effectively to improve treatments and outcomes for patients, which requires data on every patient, every visit, every diagnosis, laboratory, imaging, everything that is generated as information for every individual – patient, doctor, healthcare worker [6]. In the healthcare sector, patient data may be scattered across databases from different hospitals, private GPs, family and other doctors,

[1] Evgenija Gjurovska, University St. Kliment Ohridski Bitola, Faculty of ICT, Partizanska bb, 7000 Bitola, R. of N. Macedonia, E-mail: gjurovska.evgenija@uklo.edu.mk

[2] Snezana Savoska, University St. Kliment Ohridski Bitola, Faculty of ICT, Partizanska bb, 7000 Bitola, R. of N. Macedonia, E-mail: snezana.savoska@uklo.edu.mk

[3] Natasa Blazeska-Tabakovska, University St. Kliment Ohridski Bitola, Faculty of ICT, Partizanska bb, 7000 Bitola, R. of N. Macedonia, E-mail: natasa.tabakovska@uklo.edu.mk

[4] Kostandina Veljanovska, University St. Kliment Ohridski Bitola, Faculty of ICT, Partizanska bb, 7000 Bitola, R. of N. Macedonia, E-mail: kostandina.veljanovska@uklo.edu.mk

979-8-3315-2655-9/25/$31.00 ©2025 IEEE

healthcare institutions and are likely to be in different formats and types of databases. Collecting or integrating this data together is a huge challenge and is subject to many security mechanisms due to the different legal regulations and standards used in different countries around the world. A major challenge for doctors, managers of healthcare institutions that provide healthcare services and researchers who routinely receive and process data from a range of sources is obtaining data from heterogeneous and distributed data sources [7].

The paper is structured in the following way. After the Introduction section, the Objectives and the used dataset are considered. Section 3 explains the proposed model, RF Algorithm usage. Section 3 explains the data visualization of the distribution of each numerical features from the used dataset. Section 5 takes in consideration the application created for diabetes prediction using RFA model and a discussion about the results. The conclusion part considers the final deduction of the paper and future works.

II. Objectives and used dataset

The objectives of this research are to explore and implement various Key Performance Indicators (KPIs) in diabetes prediction using data visualization techniques alongside the RF algorithm. KPIs help measure the effectiveness of predictive models and provide actionable insights for healthcare professionals. By focusing on visual representation of KPIs, this study aims to enhance decision-making and improve the early diagnosis of diabetes.

In this paper, the authors used the data from the Indian Pima Indigenous Diabetes database (IPID), collected with the project of the National Institute of Diabetes, Digestive and Kidney Diseases in the United States of America. Data are accessible from the Kaggle website [8]. Data are an open-source dataset that has data for female patients. The dataset has 768 cases (Fig.1) and each case has a binary indicator - non-diabetic (0) and diabetic (1). This dataset has 500 cases classified as non-diabetic and 268 cases as diabetic. The dataset provides eight features: Pregnancies, Glucose level. Blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function and age.

1. Pregnancies are the number of pregnancies for each female.

2. Glucose Level is measured glucose concentration in plasma due to an oral glucose tolerance test.

3. Blood pressure is measured blood pressure in their body's cardiovascular system that is very important because both, high and low blood pressures can have health implications and can serve as a potential mortality indicator.

4. Skin Thickness measured in millimetres triceps offers a reliable estimation of obesity and body fat distribution. For this reason, it can serve as a good indicator in enabling insights in body constitution and fat distribution.

5. Insulin measured as insulin level in blood after a two-hour period -"mu U/ml" (micro-units per millilitre). The two-hour serum insulin level it is possible to detect metabolic disorder and a defect in islet function. These functions are connected with diabetes because a peptide hormone Insulin is primarily produced by the beta cells in pancreatic islets and for these reasons it serves as the main anabolic hormone in the body. The main role is facilitating the glucose absorption into the liver from the carbohydrates' metabolism of proteins and fats which means from the bloodstream into the liver as well as fat in the adipose tissue (fat). In addition, the glucose absorption in the skeletal muscle cells is enabled in this way.

6. BMI (Body mass index) is a measure of obesity and health used in statistical analysis. Obesity cannot be detected directly but it is related to peoples' height. BMI is calculated as the body weight divided by the body height squared.

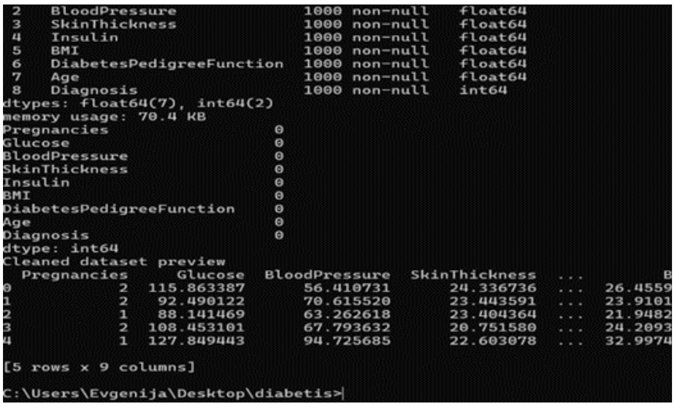
7. Diabetes Pedigree Function (DBF) indicates diabetes development probability (influence of familial background.

8. Age in years - from 21 to 81.

Therefore, we consider that in the classification process, we get outcome 0 – patient with Type II diabetes, and 1 - participant without Type II diabetes

III. Proposed model: RF Algorithm

Use dataset has a dependent variable "Outcome" as binary values (0 and 1), representing the presence or absence of diabetes. RF, a powerful ensemble learning method, is highly suited for predicting such outcomes. The RF algorithm works by creating a multitude of decision trees during training and outputs the most popular class (for classification problems) as the final prediction. It combines the predictions taken from several decision trees in order to enhance accuracy and reduce overfitting. RF is an extension of the decision tree algorithm and it benefits from the concept of Bootstrap Aggregating – "bugging", which use multiple subsets of the training data to create several models. Each decision tree is trained independently on a random data subset. The final prediction is created on the basis on all trees' majority vote. This process increases the model's robustness and generalizability compared to a single decision tree [1]. The key advantage of RF over other algorithms, like logistic regression, lies in its ability to handle complex relationships between variables and to automatically capture interactions between features without requiring prior feature engineering. Additionally, unlike logistic regression, which assumes a linear relationship between predictors and the log-odds of the outcome, RF does not make such assumptions and can model nonlinear relationships effectively.



```
 2   BloodPressure             1000 non-null   float64
 3   SkinThickness             1000 non-null   float64
 4   Insulin                   1000 non-null   float64
 5   BMI                       1000 non-null   float64
 6   DiabetesPedigreeFunction  1000 non-null   float64
 7   Age                       1000 non-null   float64
 8   Diagnosis                 1000 non-null   int64
dtypes: float64(7), int64(2)
memory usage: 70.4 KB
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Diagnosis                   0
dtype: int64
Cleaned dataset preview
  Pregnancies     Glucose  BloodPressure  SkinThickness  ...         B
0           2  115.863387      56.410731      24.336736  ...   26.4559
1           2   92.490122      70.615520      23.443591  ...   23.9101
2           1   88.141469      63.262618      23.404364  ...   21.9482
3           2  108.453101      67.793632      20.751580  ...   24.2093
4           1  127.849443      94.725685      22.603078  ...   32.9974

[5 rows x 9 columns]

C:\Users\Evgenija\Desktop\diabetis>
```

Fig. 1: The database and values

In our study, the RF algorithm will be used to predict the likelihood of diabetes (Outcome = 1) or the absence of diabetes (Outcome = 0) based on a range of predictor variables. This method's flexibility and power in handling large datasets, as well as its capacity to assess feature importance, make it an ideal choice for this prediction task.

IV. Data Visualization

The analysis below (Fig. 2) is based on the "Diabetes Prediction Data" dataset obtained from the Kaggle website. This dataset provides a rich source of information that serves as a basis for predicting diabetes risk. The data generated for diabetes prediction purposes contains a set of features related to diabetes risk factors. These features may include variables such as blood sugar levels, body mass index (BMI), age, family history, and other relevant health indicators. A diagnosis label indicating whether the individual has diabetes accompanies each set of feature values or not. This dataset is important for training machine-learning models to predict the likelihood of diabetes based on given risk factors. It can be used for research, analysis, and development of predictive models aimed at improving the diagnosis and management of diabetes.

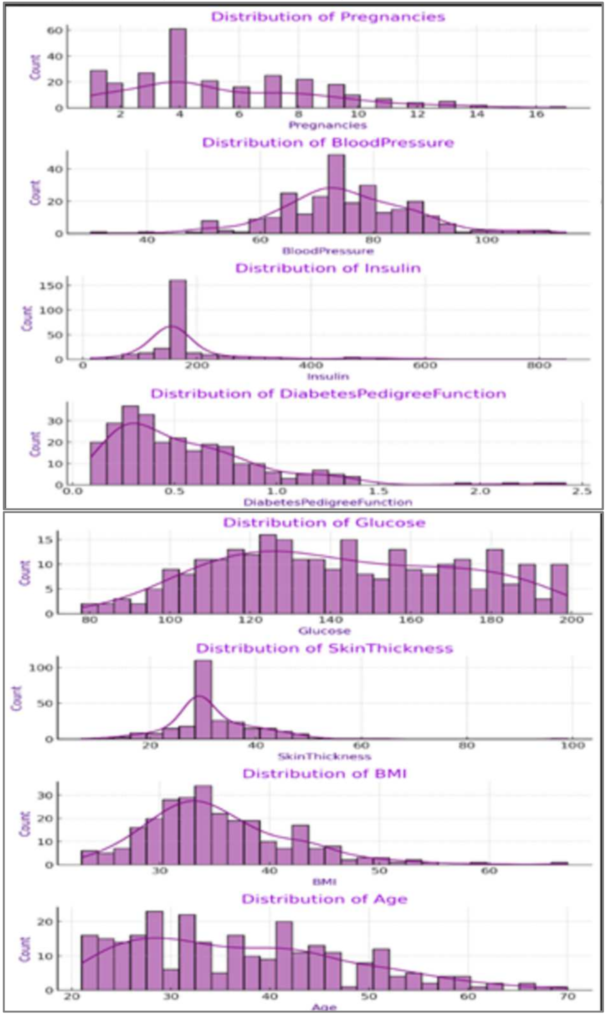


Fig. 2. The visualization explains the distribution of each numerical feature in the database

Most of the features in Fig. 2 show a right-skewed distribution (e.g., Insulin, Genetic predisposition). Glucose, BMI, and age appear normally distributed. Some features such as skinfold thickness and insulin have wide ranges, indicating potential outliers.

V. Model: Random Forest Algorithm

In the IPID dataset, the dependent variable "Outcome" has 0 and 1 binary values, indicating the presence or absence of diabetes. Given this binary classification task, the RFA emerges as a strong, powerful, versatile prediction approach. RF is an ensemble learning method combining the predictions of multiple decision trees in order to enhance model accuracy and reduce overfitting.

RF is particularly effective for scenarios where the relationship between predictor variables and the outcome is non-linear or complex and does not make strong data assumptions and training is made on a random dataset subset. Then, the final prediction considers the all-individual trees' majority vote. This process helps capture various interactions between features, making RF a robust model for binary classification tasks, such as predicting diabetes in this study.

One of the key advantages of RF over other algorithms, such as logistic regression, is its ability to automatically capture complex relationships between the predictor variables and the target outcome. Unlike logistic regression, which assumes a linear relationship between predictors and the outcome, RF is non-parametric and can handle intricate patterns without requiring prior feature transformation. Furthermore, RF provides insights into feature importance, helping to identify the most influential variables for predicting the outcome.

In our study, we will utilize RF to predict the likelihood of diabetes (Outcome = 1) or the absence of diabetes (Outcome = 0), based on several predictor variables. This method is particularly advantageous for handling large datasets with numerous features, as it efficiently handles variable interactions and helps prevent overfitting through the ensemble approach. The workflow of RF algorithm is shown in Fig. 3.

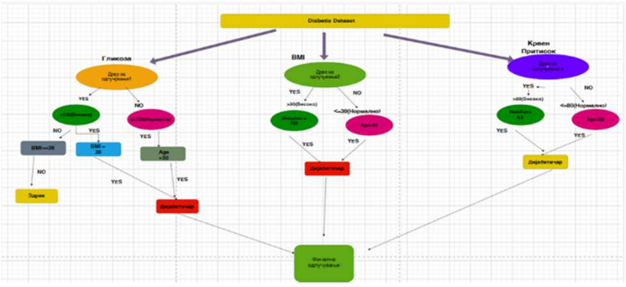


Fig. 3. Random Forest Algorithm

VI. Discussion

In the context of deductive methods, the paper applies a ML model, the RF algorithm, to predict diabetes based on patient data (Fig. 4). Deductive methods have also been used to extract Key Performance Indicators (KPIs) for the visualizations.

1. General principle (Main premise)

Medical research has shown that diabetes can be predicted based on measurable biological and demographic factors such as glucose levels, BMI, insulin levels, blood pressure, smoking history or heart disease. These factors significantly contribute to the risk of diabetes.

2. Specific observation (small premise)

By analyzing a dataset of 1000 patients, a ML model was applied to test whether these features reliably predict diabetes, the RF model was trained on these predictors.

Key Observations:

- Patients with high glucose (>126 mg/dL) and BMI (>30) are more likely to have diabetes.
- Younger individuals with normal glucose and insulin levels are less likely to develop diabetes.
- Blood pressure and a family history of diabetes also contribute to the risk of diabetes, but are weaker predictors than glucose and BMI.

3. Logical Deduction (Conclusion)

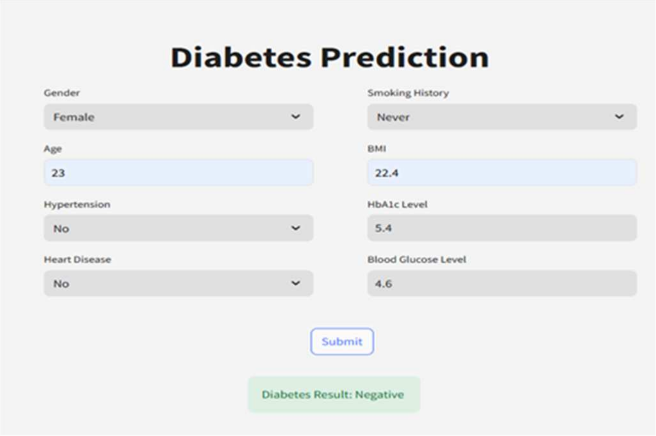
Since the database and model confirm established medical findings, the RF model can predict diabetes with reasonable accuracy. The prediction accuracy of 68.75% suggests that although the model is useful, improvements are needed to improve recall and sensitivity to diabetes cases.

4. Conclusion (Final Deduction)

VII. Conclusion

By analyzing a dataset of 1000 patients, a ML model was applied to test whether these features reliably predict diabetes, the RF model was trained on these predictors.

Key Observations:



Diabetes Prediction

Gender: Female

Smoking History: Never

Age: 23

BMI: 22.4

Hypertension: No

HbA1c Level: 5.4

Heart Disease: No

Blood Glucose Level: 4.6

Diabetes Result: Negative

Fig. 4. Negative result, the patient is negative for diabetes

- Patients with high glucose (>126 mg/dL) and BMI (>30) are more likely to have diabetes.
- Younger individuals with normal glucose and insulin levels are less likely to develop diabetes.
- Blood pressure and a family history of diabetes also contribute to the risk of diabetes, but are weaker predictors than glucose and BMI.

3. Logical Deduction (Conclusion)

Since the database and model confirm established medical findings, the RF model can predict diabetes with reasonable accuracy. The prediction accuracy of 68.75% suggests that although the model is useful, improvements are needed to improve recall and sensitivity to diabetes cases.

4. Conclusion (Final Deduction)

Given the performance analysis and the importance of the model features: Diabetes can be predicted using ML models based on medical and demographic factors. Glucose levels, BMI, and age are the most significant indicators of diabetes risk. The RF model, despite its reasonable accuracy, needs further optimization for better recall in predicting diabetes cases. This structured deductive method follows a logical flow of reasoning, leading from established general principles to specific conclusions based on data analysis.

Visualizations included:

KPI indicators:

- Statistical distributions of health-related parameters.
- Correlation heat maps to identify relationships between different health factors.
- Time series visualization to track patient trends over time.
- Geospatial maps to analyze health data.
- Bubble charts, pie charts, and histograms to effectively represent data distributions and category comparisons.

The combination of data visualization and predictive modelling offers a powerful solution to address healthcare challenges by improving the quality of care, improving decision-making, and enabling early diagnosis and intervention. However, to fully realize the potential of these technologies, the healthcare industry must address challenges related to data quality, model transparency, privacy, and ethical considerations. By prioritizing future research in these areas, we can move closer to creating a healthcare system that is more efficient, personalized, and equitable, ultimately improving patient outcomes and quality of life.

References

[1] Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.

[2] Sampath, P., Elangovan, G., Ravichandran, K. et al. Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. Sci Rep 14, 28984 (2024). https://doi.org/10.1038/s41598-024-78519-8

[3] Morgan P., Data Analysis from Scratch with Python, AI Scientist, ISBN-13: 978-1721942817 - 17/01/2025

[4] Ossama Embarak, Data Analysis and Visualization Using Python, Apress, 2018, https://doi.org/10.1007/978-1-4842-4109-7 17/01/2025

[5] American Diabetes Association; Standards of Medical Care in Diabetes—2009. Diabetes Care 1 January 2009; 32 (Supplement_1): S13–S61. https://doi.org/10.2337/dc09-S013

[6] Snezana, Savoska, et al. "Cloud based personal health records data exchange in the age of IoT: the Cross4all project." International Conference on ICT Innovations. Cham: Springer International Publishing, 2020.

[7] Ristevski, Blagoj, and Snezana Savoska. "Healthcare and medical Big Data analytics." Applications of Big Data in Healthcare: Theory and Practice (2021): 85.

[8] Kagggle website: https://www.kaggle.com/uciml/pima-indians-diabetes-database)