Enhancing Early Diabetes Prediction Using Machine Learning: An Ethical and Practical Approach

Student ID: 23220052

***Abstract*—** In recent years, the healthcare industry has witnessed a rise in innovative solutions through the development of artificial intelligence (AI). This led to a revolutionary shift in medical practices and diagnosis of chronic diseases such as diabetes. Diabetes, a global condition, results in severe health complications if not managed properly, making early prediction and intervention crucial to enhance patient outcomes. Our project uses supervised learning on a Kaggle dataset containing health and lifestyle indicators. This approach consists in rigorous data analysis between diabetes and crucial variables such as blood glucose levels, HbA1c levels, age, BMI, gender and lifestyle habits to develop a Machine Learning (ML) model that predicts diabetes in patients with high accuracy on an unseen dataset. Finally, we discuss our results, emphasizing on the ethical considerations of inclusiveness and equity in healthcare, aiming to advance preventive healthcare strategies and promote better health outcomes globally.

Keywords: Machine Learning, Healthcare, Artificial intelligence, Diabetes, Predictive analysis

I. Introduction

The healthcare industry has witnessed a rise in AI-driven innovations. This led to a revolutionary shift in medical practices, diagnosis and patient- centric care [1]. Indeed, complex algorithms based on Machine Learning (ML) become increasingly integrated to clinical decision-making processes [2]. By providing accurate, efficient and scalable solutions, these AI technologies are critical to address healthcare challenges such as early disease detection.

Traditionally, diagnostic methods rely on symptom manifestation, which often lead to delayed and incorrect diagnosis. However, through ML models, predictive analysis leverages large patient datasets to identify early disease patterns and risk factors [3]. This allows timely intervention, improved patient outcomes and survival rates, which are crucial for conditions like diabetes which have better prognoses and treatment success rates when identified early [4].

In 2023, diabetes affects over half a billion people globally [5] with 1.5 million deaths each year [6]. Being a chronic disease, diabetes is characterized by elevated blood glucose levels due to the patient inability to produce or properly use insulin [7]. This metabolic disorder comes in 2 types: Type 1 which consists of an autoimmune disease; and Type 2, which is linked to lifestyle habits such as physical activity, diet, obesity, smoking and alcohol consumption [8]. However, diabetes symptoms are often unnoticed as they consists of frequent urination, extreme fatigue, and blurred vision [9]. This may lead to delayed diagnosis and complications such as Cardiovascular diseases (CVDs), retinopathy and Alzheimer's Disease [10]. These complications, arising from the disease lack of management, position diabetes as a significant public health challenge, with a projection to double to 1.3 billion people by 2050 [11]. However, given that its early detection is critical for effective management and complication prevention, it is vital to leverage AI to develop predictive models based on ML methods. Such models can identify high-risk individuals, predict disease risks and suggest preventive measures, enhancing patient outcomes and reducing healthcare costs by minimizing the need for more intensive treatments and hospitalizations [12].

In this report, we will explore potential predictive model for diabetes, develop our own, outline the methodology used, describe and discuss our results as well as shed light on the ethical implications of our work.

II. Literature review

Effectively managing diabetes is a significant public health challenge [13], that has led to various research focusing on its early detection and prediction by leveraging ML algorithms. These approaches analyze complex data patterns to predict high-risk individuals based on risk factors and lifestyle habits.

A. Logistic Regression

Logistic regression (LR) is a statistical and binary classification method used to estimate the probability of an outcome based on input features [14]. In terms of diabetes prediction, LR models have been widely used due to their simplicity and interpretability [15]. Indeed, Nguyen et al. (2019) used LR to predict diabetes risk using features such as age, BMI, and blood glucose levels [16]. The model achieved high accuracy and performance, and was particularly acknowledged for its simple interpretation. It was found that each feature was a significant predictor of diabetes, allowing better understanding of their impact on diabetes risk. Moreover, according to Rajendra and Latifi (2021), LR was also considered as an efficient algorithm in building prediction models [17]. However, its accuracy did not only depend on the chosen algorithm but also on data pre-processing. Indeed, removing duplications and null values, and using cross-validation was essential to improve its performance. Likewise, Anderson et al. (2015) conducted a study using LR and found higher accuracy with this model [18]. However, they raised concerns regarding the poor development of these predictive models due to the inappropriate selection of covariates, missing data and small sample sizes [19]. Furthermore, the reliability and quality of these predictive models highlighted significant variation based on geography, available data, and ethnicity

[16]. Indeed, LR performance presents limitations when dealing with complex, non-linear relationships in the data as it may not capture intricate patterns as effectively as more sophisticated algorithms [20]. However, Joshi and Dhakal (2021) study compared the LR method with the decision tree (DT) method and found that the risk factors for diabetes identified by LR were validated by DT, suggesting that this model can help classify high-risk individuals accurately and help in the prevention, diagnosis, and management of diabetes [21].

B. *Decision Tree*

On the other hand, DT are intuitive models that create a tree-like structure by separating data into branches based on feature values [22]. Being fast, easy to interpret and performing well even on large dataset, they learn simple decision rules provided by data features, handle well non-linear relationships and develop models that predicts the target value [23]. Indeed, Iparraguirre-Villanueva, et al. (2023) study demonstrated the application of DTs in diabetes prediction, showing the model's ability to manage complex interactions between BMI, insulin levels, and age [24]. However, according to Smolic (2023) they are also prone to overfitting, especially when they are deep and complex [25].

C. *Random Forest*

Furthermore, Random Forests (RF) are models that develop multiple DTs and combine their predictions [26]. In terms of predictive accuracy, this approach outperforms LR due to its ability to handle complex interactions between features, as well as DTs due to its ensemble approach that reduces overfitting [27]. Indeed, both studies by Wang (2024) and Mahboob Alam, et al. (2019) used RF on diabetes datasets, resulting in high accuracy[28][29], while Ahmed et al. (2021) study compared RF approach to others, and highlighted it having the highest level of accuracy, exceeding others [30]. This was explained by the effectiveness of RF method in identifying diabetes predictors, managing data imbalances and handling large datasets [28], as well as its ability to provide a balance between accuracy and interpretability [29]. However, despite its strengths, RF models can be computationally intensive and less interpretable than simpler models like LR [31].

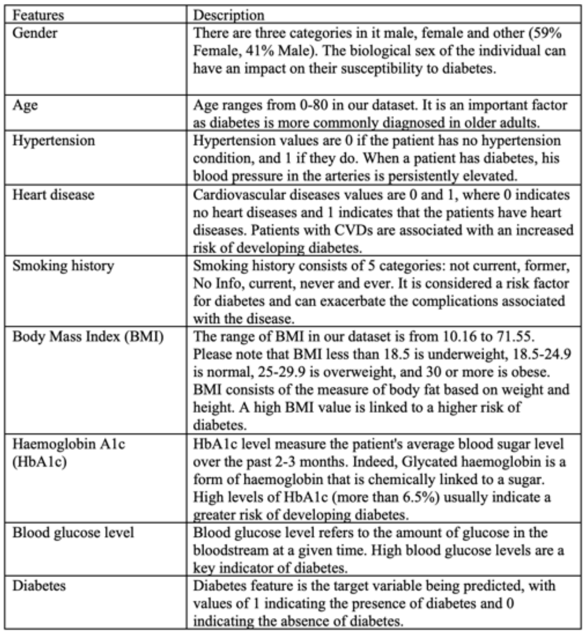
D. *SVMs*

Finally, other studies such as Sharma and Shah (2021), compared LR, DT, RF and SVMs methods for diabetes prediction and discovered that SVMs performed best [32], offering the highest accuracy and robustness [33]. However, according to GeeksforGeeks (2023), SVM method require complete and small dataset as they cannot handle missing values, consume a lot of memory and are very slow if many feature are within the dataset [34].

III. Methodology

A. *Our dataset*

The "diabetes_prediction_dataset.csv file" used for our ML model development was retrieved from Kaggle [35]. It comprises 100,000 patients medical and demographic data, including essential features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level and their diabetes status. These factors are vital to build a ML model that predicts diabetes in patients. Indeed, they allow Healthcare professionals (HCPs) to identify patients at risk of developing diabetes, and understand the correlation between these factors and diabetes risk. Therefore, our project aims to train our ML model on this labelled dataset with known outcome, to accurately predict patients at risk of diabetes.



Features	Description
Gender	There are three categories in it male, female and other (59% Female, 41% Male). The biological sex of the individual can have an impact on their susceptibility to diabetes.
Age	Age ranges from 0-80 in our dataset. It is an important factor as diabetes is more commonly diagnosed in older adults.
Hypertension	Hypertension values are 0 if the patient has no hypertension condition, and 1 if they do. When a patient has diabetes, his blood pressure in the arteries is persistently elevated.
Heart disease	Cardiovascular diseases values are 0 and 1, where 0 indicates no heart diseases and 1 indicates that the patients have heart diseases. Patients with CVDs are associated with an increased risk of developing diabetes.
Smoking history	Smoking history consists of 5 categories: not current, former, No Info, current, never and ever. It is considered a risk factor for diabetes and can exacerbate the complications associated with the disease.
Body Mass Index (BMI)	The range of BMI in our dataset is from 10.16 to 71.55. Please note that BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese. BMI consists of the measure of body fat based on weight and height. A high BMI value is linked to a higher risk of diabetes.
Haemoglobin A1c (HbA1c)	HbA1c level measure the patient's average blood sugar level over the past 2-3 months. Indeed, Glycated haemoglobin is a form of haemoglobin that is chemically linked to a sugar. High levels of HbA1c (more than 6.5%) usually indicate a greater risk of developing diabetes.
Blood glucose level	Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.
Diabetes	Diabetes feature is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

Table 1: *Dataset features description*

B. *Our code explanation*

Before starting our project, we first import all required libraries for our program solution.

a) Data exploration:

In order to predict diabetes, we first import and load our dataset from the 'diabetes_prediction_dataset.csv' file. Then, we thought it would be beneficial to generate a summary of our dataset to gain insights into the data structure, distribution, inconsistencies and missing values. Such insights allow us to understand the overall behavior and spread of our dataset, in order for us to select the appropriate analysis techniques and models.

b) Data pre-processing:

Our dataset undergoes pre-processing to ensure uniformity and compatibility for ML models. We noticed that it contains a mix of numerical and non-numerical data. To ensure consistent handling of our data, we convert all non-numerical data into numerical equivalents. Then, once our dataset is pre-processed and cleaned, we split it into training (70%), testing (25%) and deployment (5%) sets which we will use to train, evaluate and validate our model. However, prior to developing our model, we identified and visualized how many people have diabetes in our training and testing datasets, as well as checked for any data imbalances. Indeed, this is crucial as it ensures that ML models perform well across all classes, avoiding bias towards the majority class and providing more accurate and reliable predictions. Finally, we perform an analysis to understand the correlation between numerical features and the 'diabetes' target variable in our dataset.

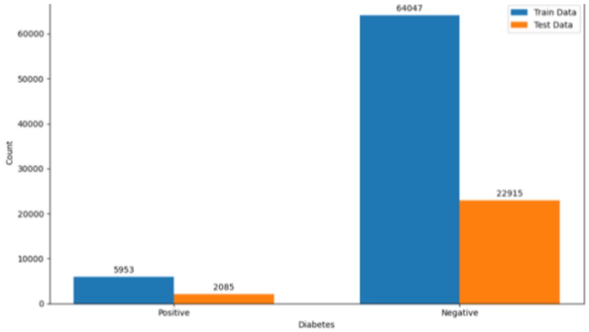


__Table 2:__ *Counts of Positive and Negative Entries in Train and Test Data*

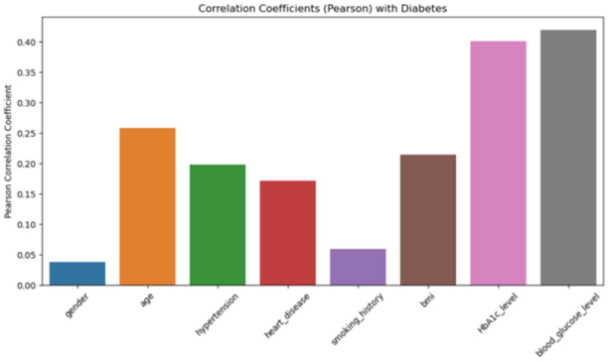


__Table 3:__ *Correlation Coefficients (Pearson) with Diabetes*

c) Model development:

Then, based on our literature review, we decided to evaluate LR, DT and RF models effectiveness in diabetes prediction. We compared their accuracy to determine which one is the most suitable for our model training. Moreover, for our current analysis, we have decided not to incorporate clustering. Our focus is on supervised learning tasks where we already possess labelled data. We have earmarked it for potential future exploratory analyses (AE2), where uncovering underlying patterns and enhancing predictive models could provide additional insights.

d) Training and evaluation:

Following the above, RF classifier showed the highest accuracy in predicting diabetes based on our dataset. Therefore, we trained our RF model on our training data and then evaluated its performance on our testing data. Finally, we generated a classification report which consists in providing precision, recall, F1-score metrics, allowing us to understand how well our model predicts each category.

e) Hyperparameter tuning:

Then, using the GridSearchCV function, we identified the best combination of parameters for our model and fine-tuned it to enhance its accuracy on our testing data. We then re-evaluated it using the optimized hyperparameters to further improve its performance.

f) Model validation:

Subsequently, we decided to do a cross-validation on our RF model across 5 folds of our training data. This helps mitigate the risk of overfitting and ensures the model generalization to new data.

g) Model deployment:

Once done, we save our best model so that it can be loaded and reused for predictions on new data, without retraining it from scratch. We then deploy our model on our unseen deployment data (5%, X_deploy, y_deploy). This is crucial because it provides insights into the model's ability to generalize to real-world scenarios and validate its robustness and reliability before it is used for making critical decisions. Finally, we calculate the accuracy score of our model on our deployment dataset to validate and maintain its efficacy in real-world applications. Please, note that we analyzed the dataset features to gain insight on the ones that have the most significant impact on our model's predictions.

IV. Results

Initially, our dataset had no missing values, allowing straightforward use. Upon splitting our dataset, our training data (70,000 cases), had 5,953 positive and 64,047 negative diabetes cases, with a difference of 58,094, while our testing dataset (25,000 cases), had 2,085 positive, and 22,915 negative, resulting in a difference of 20,830. Therefore, our dataset exhibited a slightly imbalanced distribution (8.50% positive in training, 8.34% in testing; 91.50% negative in training, 91.66% in testing). However, our model achieved high accuracies despite this class imbalance. Among evaluating all potential models (LR, DT and RF), we noticed that our RF model emerged as the most effective. Indeed, while LR achieved a mean accuracy of 96.08%, and DT 95.15% , RF achieved the highest one (97.02%). Therefore, we selected the RF model for further refinement and deployment due to its superior performance. Once our model trained, it yielded a significant training accuracy of 99.92% and maintained a testing accuracy of 96.99%. Indeed, the classification report showed a precision of 0.97 for class 0 (non-diabetic) and a recall of 0.68 for class 1 (diabetic), resulting in an overall balanced F1-score of 0.97. Then, we identified the best parameters for our RF model, which consisted in a max depth of 10, a min sample split of 2 and n_estimator of 200. When further evaluating the best-tuned RF mode, it achieved a slightly improved accuracy of 97.22%. Subsequently, while cross- validating our results, our model performance resulted in a mean accuracy of 97.02%. Finally, when using our model on our deployment data, we noticed that our RF still maintained a strong accuracy of 96.88%.

V. Discussion

The use of ML methods in our diabetes prediction study resulted into promising outcomes. Both our training and testing datasets exhibited a similar distribution of diabetes cases, indicating consistent representation of the target variable. Our model effectively identified both positive and negative cases of diabetes with high precision and recall on new, unseen data. This consistency further underscores our model's robustness and suitability for practical deployment in healthcare settings. However, ensuring our solution sustainability requires continuous monitoring and updating to maintain its relevance and accuracy.

Then, our Pearson correlation analysis highlighted a strong correlation between diabetes and key features such as HbA1c level, blood glucose, BMI, age, smoking history, hypertension, heart disease, and gender. These findings underscore the predictive potential of these variables, despite limitations related to encoding categorical variables as numerical. However, improvements in feature engineering and data pre-processing techniques could enhance model performance and reduce biases [36]. Notably, factors like genetics and alcohol consumption, not included in our dataset, significantly influence diabetes risk, which potentially introduces inaccuracies in our predictions [8].

Reflecting on the ethical implications of our model implementation reveals several critical considerations. Indeed, the inability to remove duplicates and convert categorical data to numerical may introduce biases and inaccuracies, affecting our prediction reliability [37]. This underscores the need for rigorous data pre-processing to ensure quality and integrity. Moreover, the absence of ethnicity data limits our understanding of diabetes prevalence across different demographic groups [38]. Including ethnicity data could provide nuanced insights and enhance the inclusiveness of our model, ensuring equitable healthcare outcomes. Our solution's impact on inclusiveness and diversity requires careful consideration, as predictive models can improve healthcare outcomes by enabling early intervention but may exacerbate disparities if access to technology is unequal [39]. Therefore, proactive efforts are needed to ensure equitable access and considerate deployment of predictive models in healthcare settings.

Moreover, safeguarding data privacy and confidentiality is crucial, especially for sensitive health information [40]. Therefore, ensuring compliance with data protection regulations and robust security measures are imperative to mitigate risks of data breaches and unauthorized access [41]. Finally, relying solely on automated prediction models without human oversight can lead to erroneous conclusions, particularly in complex scenarios, potentially causing HCPs to misinterpret the results and misjudge diagnoses [42]. Therefore, providing clear explanations of our model's operation, limitations, and potential biases is crucial for maintaining transparency.

VI. Conclusion

To conclude, our study aims to improve early diabetes prediction using ML models. Affecting over half a billion globally, diabetes presents a significant public health challenge. Therefore, our RF model offers precise and timely diabetes diagnosis, enhancing patient outcomes and reducing healthcare costs. Despite promising outcomes, continuous monitoring and updates are essential for sustainability. Indeed, improvements in data pre-processing and ethical considerations —addressing biases, including ethnicity data, safeguarding data privacy, and maintaining transparency—are crucial for trustworthy predictions. Therefore, to further refine our model predictive capabilities, it would be interesting to incorporate ensemble methods that integrate multiple model predictions in order to enhance robustness and generalizability.

References

[1] Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In Artificial Intelligence in Healthcare. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

[2] Iqbal, J., Cortés Jaimes, D. C., Makineni, P., Subramani, S., Hemaida, S., Thugu, T. R., Butt, A. N., Sikto, J. T., Kaur, P., Lak, M. A., Augustine, M., Shahzad, R., & Arain, M. (2023). Reimagining Healthcare: Unleashing the Power of Artificial Intelligence in Medicine. Cureus. https://doi.org/10.7759/cureus.44658

[3] Batko, K. and Ślęzak, A. (2022) The use of Big Data Analytics in Healthcare, Journal of big data. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8733917/ (Accessed: 04 July 2024).

[4] Herman, W.H. et al. (2015) Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: A simulation of the results of the Anglo-danish-dutch study of intensive treatment in people with screen-detected diabetes in primary care (addition-Europe), Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality:

A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512138/ (Accessed: 04 July 2024).

[5] Institute for Health Metrics and Evaluation (no date) Global diabetes cases to soar from 529 million to 1.3 billion by 2050. Available at: https://www.healthdata.org/news-events/newsroom/news-releases/global-diabetes-cases-soar-529-million-13-billion-2050#:~:text=June%2022%2C%202023%20%E2%80%93%20More%20than,published%20today%20in%20The%20Lancet%20. (Accessed: 04 July 2024).

[6] Diabetes (no date) World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=In%202019%2C%20diabetes%20was%20the,of%20cardiovascular%20deaths%20(1). (Accessed: 04 July 2024).

[7] What is diabetes? (no date) National Institute of Diabetes and Digestive and Kidney Diseases. Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes#:~:text=If%20you%20have%20diabetes%2C%20your,to%20some%20types%20of%20cancer. (Accessed: 04 July 2024).

[8] Sami, W. et al. (2017) Effect of diet on type 2 diabetes mellitus: A Review, International journal of health sciences. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426415/ (Accessed: 04 July 2024).

[9] Diabetes UK (no date) What are the signs and symptoms of diabetes?, Diabetes UK. Available at: https://www.diabetes.org.uk/diabetes-the-basics/diabetes-symptoms (Accessed: 04 July 2024).

[10] W, E. (no date) Complications of diabetes, Diabetes UK. Available at: https://www.diabetes.org.uk/guide-to-diabetes/complications (Accessed: 04 July 2024).

[11] Klein, H.E. (2023) Diabetes prevalence expected to double globally by 2050, AJMC. Available at: https://www.ajmc.com/view/diabetes-prevalence-expected-to-double-globally-by-2050 (Accessed: 04 July 2024).

[12] Toma, M. and Wei, O.C. (2023) Predictive modeling in medicine, MDPI. Available at: https://www.mdpi.com/2673-8392/3/2/42 (Accessed: 04 July 2024).

[13] Sugandh, F. et al. (2023) Advances in the management of diabetes mellitus: A focus on personalized medicine, Cureus. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10505357/ (Accessed: 05 July 2024).

[14] S, P. (2024) Logistic regression model: A guide to machine learning techniques and applications, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/#:~:text=Logistic%20regression%20is%20a%20Machine,likelihood%20of%20a%20specific%20outcome. (Accessed: 05 July 2024).

[15] Olusanya, M.O. et al. (2022) Accuracy of machine learning classification models for the prediction of type 2 diabetes mellitus: A systematic survey and meta-analysis approach, International journal of environmental research and public health. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9655196/ (Accessed: 05 July 2024).

[16] Nguyen B.P., Pham H.N., Tran H., Nghiem N., Nguyen Q.H., Do T.T., Tran C.T., Simpson C.R. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 2019;182:105055. doi: 10.1016/j.cmpb.2019.105055.

[17] Rajendra, P. and Latifi, S. (2021) Prediction of diabetes using logistic regression and ensemble techniques, Computer Methods and Programs in Biomedicine Update. Available at: https://www.sciencedirect.com/science/article/pii/S2666990021000318#sec0002 (Accessed: 05 July 2024).

[18] Anderson A.E., Kerr W.T., Thames A., Li T., Xiao J., Cohen M.S. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J. Biomed. Inform. 2016;60:162–168. doi: 10.1016/j.jbi.2015.12.006.

[19] Kalil A.C., Mattei J., Florescu D.F., Sun J., Kalil R.S. Recommendations for the assessment and reporting of multivariable logistic regression in transplantation literature. Am. J. Transplant. 2010;10:1686–1694. doi: 10.1111/j.1600-6143.2010.03141.x.

[20] Advantages and disadvantages of logistic regression (2024) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (Accessed: 05 July 2024).

[21] Joshi, R.D. and Dhakal, C.K. (2021) Predicting type 2 diabetes using logistic regression and machine learning approaches, International journal of environmental research and public health. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8306487/ (Accessed: 05 July 2024).

[22] Sharma, A. (2024) 4 simple ways to split a decision tree in machine learning (updated 2024), Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/ (Accessed: 05 July 2024).

[23] Cerchia, C. and Lavecchia, A. (2023) New Avenues in artificial-intelligence-assisted drug discovery, Science direct. Available at: https://www.sciencedirect.com/science/article/pii/S1359644623000326 (Accessed: 05 July 2024).

[24] Iparraguirre-Villanueva, O. et al. (2023) Application of machine learning models for early detection and accurate classification of type 2 diabetes, U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10378239/ (Accessed: 05 July 2024).

[25] Smolic, H. (2023) Demystifying ai decision trees: A practical guide to understanding and implementing, Graphite Note. Available at: https://graphite-note.com/demystifying-ai-decision-trees-a-practical-guide-to-understanding-and-implementing/#:~:text=However%2C%20like%20any%20AI%20model,poor%20generalization%20on%20unseen%20data (Accessed: 05 July 2024).

[26] Random Forest (no date) Random Forest - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/random-forest#:~:text=It%20is%20based%20on%20the,of%20the%20data%20and%20features. (Accessed: 05 July 2024).

[27] Sharma, A. (2024) Random Forest vs decision tree: Which is right for you?, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (Accessed: 05 July 2024).

[28] Wang, S. (2024) Diabetes prediction using random forest in Healthcare, Highlights in Science, Engineering and Technology. Available at: https://drpress.org/ojs/index.php/HSET/article/view/19875 (Accessed: 05 July 2024).

[29] Mahboob Alam , T. et al. (2019) A model for early prediction of diabetes, Informatics in Medicine Unlocked. Available at: https://www.sciencedirect.com/science/article/pii/S2352914819300176 (Accessed: 05 July 2024).

[30] Ahmed, N. et al. (2021) Machine learning based diabetes prediction and development of Smart Web Application, International Journal of Cognitive Computing in Engineering. Available at: https://www.sciencedirect.com/science/article/pii/S2666307421000279#:~:text=5.4.&text=For%20NB%2C%20DT%2C%20RF%2C,and%20exceed%20the%20other%20approaches (Accessed: 05 July 2024).

[31] Couronné, R., Probst, P. and Boulesteix, A.-L. (2018) Random Forest versus logistic regression: A large-scale benchmark

experiment - BMC Bioinformatics, BioMed Central. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5 (Accessed: 05 July 2024).

[32] Sharma, T. and Shah, M. (2021) A comprehensive review of machine learning techniques on diabetes detection, Visual computing for industry, biomedicine, and art. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8642577/ (Accessed: 05 July 2024).

[33] Aishwarya Mujumdar (2020) Diabetes prediction using machine learning algorithms, Procedia Computer Science. Available at: https://www.sciencedirect.com/science/article/pii/S1877050920300557?via%3Dihub (Accessed: 05 July 2024).

[34] GeeksforGeeks (2023) Support Vector Machine in machine learning. Available at: https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/ (Accessed: 05 July 2024).

[35] Mustafa, M. (2023) Diabetes prediction dataset, Kaggle. Available at: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data (Accessed: 05 July 2024).

[36] How do you preprocess data and engineer features in your machine learning model deployment? (2024) Data Preprocessing and Feature Engineering for ML Model Deployment. Available at: https://www.linkedin.com/advice/0/how-do-you-preprocess-data-engineer-features-jkhgc#:~:text=Data%20preprocessing%20and%20feature%20engineering%20can%20bring%20many%20benefits%20to,errors%2C%20and%20facilitating%20debugging%20and (Accessed: 05 July 2024).

[37] Ray, S. (2022) Simple methods to deal with categorical variables in predictive modeling, Analytics Vidhya. Available at: https://analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/ (Accessed: 05 July 2024).

[38] Ehrenstein, V., Kharrazi, H., Lehmann, H. and Taylor, C.O. (2019). Obtaining Data From Electronic Health Records. [online] www.ncbi.nlm.nih.gov. Agency for Healthcare Research and Quality (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK551878/.

[39] Grote, T. and Keeling, G. (2022) Enabling fairness in healthcare through Machine Learning - Ethics and Information Technology, SpringerLink. Available at: https://link.springer.com/article/10.1007/s10676-022-09658-7 (Accessed: 05 July 2024).

[40] Nass, S.J. (1970) The value and importance of Health Information Privacy, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Available at: https://www.ncbi.nlm.nih.gov/books/NBK9579/ (Accessed: 05 July 2024).

[41] What is Data Security? (2024) DataGuard. Available at: https://www.dataguard.co.uk/blog/what-is-data-security/#:~:text=By%20adhering%20to%20regulations%20like,to%20brand%20reputation%2C%20and%20lawsuits. (Accessed: 05 July 2024).

[42] McKendrick , J. and Thurai, A. (2022) Ai isn’t ready to make unsupervised decisions, Harvard Business Review. Available at: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions (Accessed: 05 July 2024).