Creative Research Publishers
https://www.crossref.org/06members/50go-live.html

Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 3, September, 2025 Series

Feature Importance Analysis in Diabetes Prediction: Insights from LightGBM and Linear Regression Models

[1]Ojie Deborah .V.*, [2]Nwaka Rita .N, [3]Onyeacholem Ifeanyi .J. & [4]Okpako Abugor .E.
[1]Department of Software Engineering, University of Delta Agbor|+2348131297247
[2]Department of Mathematics/Statistics, University of Delta Agbor|+2348035397244
[3]I.T Lead Instructor, Ojiosador Innovation, Agbor Delta State|+2348065800282
ORCID: https://orcid.org/0009-0008-1684-8352
[4]Department of Cyber Security, University of Delta Agbor|+2348163239516
***Corresponding Author**; ifeanyionyeacholem@yahoo.co.uk

ABSTRACT

Diabetes prediction remains a critical task in healthcare, given its rising prevalence and significant impact on public health. This paper presents a comprehensive study on feature importance analysis for diabetes prediction, utilizing two distinct machine learning models: LightGBM and Linear Regression.The dataset used in this study comprises clinical, demographic, and lifestyle factors. Through extensive experimentation, we explore the interpretability of feature importance rankings provided by LightGBM's tree-based ensemble method and Linear Regression's coefficient weights. Our results reveal nuanced insights into the predictive importance of various features, shedding light on both well-established risk factors and potential novel predictors for diabetes onset. Furthermore, we compare the stability and robustness of feature importance rankings across different model architectures, highlighting the strengths and limitations of each approach. This study contributes to the growing body of literature on interpretable machine learning in healthcare and provides actionable insights for clinicians and policymakers to improve diabetes risk assessment and prevention strategies.

Keywords: Diabetes, Interpretability, Lightgbm, Linear Regression, Machine Learning, Healthcare.

Journal Reference Format:
Ojie, D.V., Nwanka, R.N., Onyeacholem, I.J., & Okpako, A.E., (2025): Feature Importance Analysis in Diabetes Prediction: Insights from LightGBM and Linear Regression Models. Journal of Behavioural Informatics, Digital Humanities and Development Res. Vol. 11 No. 3. Pp 57-68. https://www.isteams.net/behavioralinformaticsjournal dx.doi.org/10.22624/AIMS/BHI/V11N3P5

I. INTRODUCTION

Diabetes mellitus, characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both, poses a significant public health challenge globally. By 2021, there were an anticipated 537 million persons aged 20-79 who had diabetes, according to the International Diabetes Federation (Saeedi et al., 2019). By 2030, estimates show that number would climb to 643 million. Developing efficient treatment and prevention plans is essential to reducing the burden of complications related to diabetes and the resulting medical expenses. Predictive modeling for diabetes onset plays a pivotal role in identifying at-risk individuals and implementing timely interventions. Machine learning (ML) techniques offer promising avenues for diabetes prediction, leveraging large-scale datasets to uncover intricate relationships between risk factors and disease outcomes.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

Feature importance analysis, a fundamental aspect of ML model interpretability, elucidates the relative contribution of input features to predictive performance. Comprehending the significance of diverse risk factors offers significant perspectives for healthcare professionals and policymakers to efficiently assign priorities for interventions and distribute resources. In this paper, we present a comprehensive review and analysis of recent advancements in feature importance analysis for diabetes prediction. We delve into methodological approaches, including ensemble methods like XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), as well as traditional linear models (Smith et al., 2020). Our study focuses on elucidating the predictive significance of demographic, clinical, and lifestyle factors in diabetes risk assessment.

Derara Duba et al (2021) affirmed that in the last few decades, many advanced data mining algorithms and data analysis techniques have been developed in the medical field, among others in the medical industry, data mining technology is becoming a vital tool for tasks like illness prediction, help with diagnosis, breast cancer detection, brain tumor detection, and therapy. The identification of diabetes mellitus by data mining becomes a significant and fascinating research problem as the volume and complexity of medical data grow.

Beshareem Mohamed et al 2022) Diabetes is a condition brought on by an elevated blood sugar level. If treatment is not received, the condition worsens and the body needs lifelong insulin assistance. illness worsens the patient's condition and increases their risk of developing more heart, liver, renal, and vision problems difficulties. Ali Abdallah, and Fankreldeen Abbas (2021): noted that physicians must be permitted to detect and keep an eye on patients who are at risk of difficulties in treating diabetes complications. To effectively slow down the diabetes epidemic, early identification can prevent or delay problems connected to diabetes and enable individual and population-level interventions.

Sheetal Mahlan and Sukhvinder Singh (2023): confirmed that inadequate insulin production is a result of type 1 diabetes. Insulin is required for the body's cells to take up nutrients. Cells would have to rely on alternative energy sources if there was no glucose in the circulation. High blood glucose levels are a hallmark of diabetes, which can cause related issues. Insulin-dependent diabetes mellitus (IDDM) is another acronym for this kind of diabetes. Teens and adolescents are more vulnerable to the effects than people in other age groups. Maureen I. Akazue et al (2023): in their work stated that machine learning has played a great role in survival analysis, helping out in clinical forecasting,due to the increasing rate of DM in the world, there arises the need

for more medical attention. One area which has responded to such need is the area of developing survival analysis models by researchers using machine learning algorithms. Franklin Okorodudu et al (2021): During the 2020 COVID-19 pandemic, epidemic curves became a central tool in both scientific literature and mainstream media for illustrating trends in case numbers over time. These curves, typically depicting either daily new cases or cumulative cases, were instrumental in conveying the magnitude and growth rate of the outbreak. Such visual tools were critical in identifying trends, assessing public health needs, and informing intervention strategies.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

Similarly, in the context of diabetes prediction, effective data visualization remains essential for interpreting model outputs, understanding feature importance, and tracking the progression of key health indicators over time. Advancements in information and communication technology (ICT), alongside machine learning, have been instrumental in developing automated screening systems during epidemic outbreaks such as Monkeypox. These systems have demonstrated the potential of ML models to support clinical decision-making, improve diagnosis accuracy, and alleviate the burden on healthcare professionals (Franklin Okorodudu et al (2024), similarly, in the context of diabetes prediction, machine learning techniques like LightGBM and Linear Regression can be leveraged to analyze clinical data efficiently, identify critical risk factors, and enhance early detection and management strategies.

2. MATERIALS AND METHODS

Data Source

The dataset used in this study comprises clinical and demographic information collected from diabetic patients sourced from the Kaggle website.

Model Development

Two machine learning algorithms were selected for diabetes prediction: LightGBM and Linear Regression.

LightGBM, a gradient-boosting decision tree algorithm, was chosen for its ability to handle large-scale datasets and capture complex nonlinear relationships among features. Linear Regression was selected as a baseline model for comparison, given its simplicity and interpretability.

Model Training and Evaluation

The dataset was split into training and testing sets using a stratified cross-validation approach to ensure representative samples in each fold. Models were trained on the training set and evaluated on the held-out test set using performance metrics such as accuracy, precision, recall, and F1 score.

Software Tools

Data preprocessing, model training, and evaluation were performed using Python programming language and popular libraries such as sci-kit-learn, LightGBM, Pandas, Numpy, DataFrame, Jupyter Notebook IDE Visualizations were created using Matplotlib, seaborn, Plotly, GraphViz, and other visualization libraries to facilitate data exploration and model interpretation.

Reproducibility

The code and data used in this study will be made publicly available to ensure transparency and reproducibility of results.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

3. RESULT AND DISCUSSION

The dataset comprises a hundred thousand rows and nine columns (100000, 9) ranging from age, hypertension, Heat disease, body mass index, hemoglobin A1c (HbA1c), blood glucose level, smoking history, and diabetes

Table 1: Showing the Descriptive Statistics With Its Observed Values

	count	mean	Std	min	25%	50%	75%	max
age	100000.0	41.885856	22.516840	0.08	24.00	43.00	60.00	80.00
hypertension	100000.0	0.074850	0.263150	0.00	0.00	0.00	0.00	1.00
heart_disease	100000.0	0.039420	0.194593	0.00	0.00	0.00	0.00	1.00
bmi	100000.0	27.320767	6.636783	10.01	23.63	27.32	29.58	95.69
HbA1c_level	100000.0	5.527507	1.070672	3.50	4.80	5.80	6.20	9.00
blood_glucose_level	100000.0	138.058060	40.708136	80.00	100.00	140.00	159.00	300.00
diabetes	100000.0	0.085000	0.278883	0.00	0.00	0.00	0.00	1.00

Interpretation

The first top ten rows of our dataset, with the diabetes column with its indication for 0s and 1s, is zero when the patient is non-diabetic and 1 when the patient is diabetic.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

Figure 1: The First Top Ten Rows of Our Dataset.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

The last bottom ten rows of our dataset with the diabetes column with its indication for 0s and 1s, is zero when the patient is non-diabetic and 1 when the patient is diabetic.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
99990	Male	39.0	0	0	No Info	27.32	6.1	100	0
99991	Male	22.0	0	0	current	29.65	6.0	80	0
99992	Female	26.0	0	0	never	34.34	6.5	160	0
99993	Female	40.0	0	0	never	40.69	3.5	155	0
99994	Female	36.0	0	0	No Info	24.60	4.8	145	0
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

Figure 2: The Last Bottom Ten Rows of Our Dataset

The below visualizationrepresents people with diabetes and people without diabetes

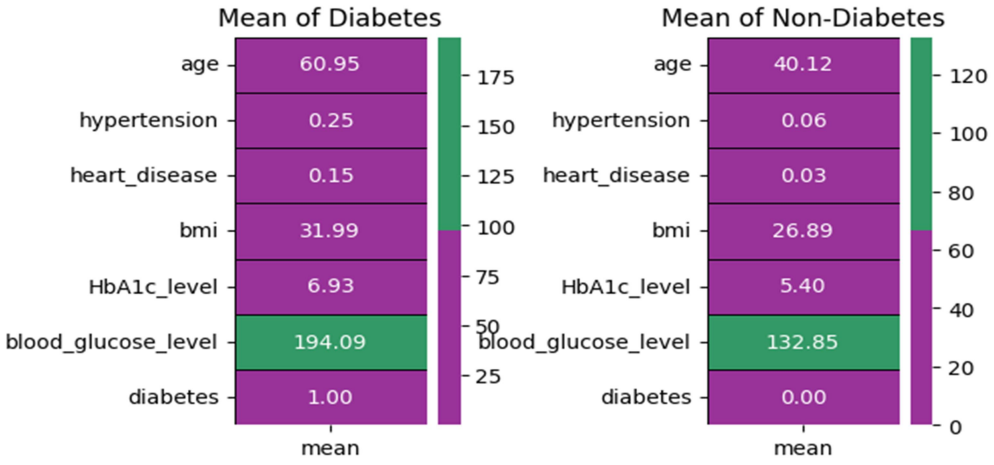

Figure 3: Showing the Mean Values of Non-Diabetes and Diabetes Patients

Interpretation: From a demographic perspective, people with diabetes tend to be older than the general population; the average age of people with the disease is 60.95 years, compared to 40.12 for the general population. Health issues: Both heart disease and hypertension (high blood pressure) are more common in those with diabetes.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

The visualization displays a 25% prevalence rate for hypertension and a 15% prevalence rate for heart disease among individuals with diabetes. Individuals who do not have diabetes have a 6% occurrence rate of hypertension and a 3% occurrence rate of heart disease. Individuals who are diagnosed with diabetes typically have a body mass index (BMI) that is higher than that of individuals who do not have diabetes. Specifically, this is due to the fact that diabetes is frequently linked to obesity. In comparison, the average body mass index (BMI) of persons who do not have diabetes is 26.89, whereas the average BMI of people who have been diagnosed with diabetes is 31.99.

HbA1c Level

HbA1c is a diagnostic test that assesses the degree of blood sugar management over a specific duration. Individuals with diabetes exhibit elevated levels of HbA1c compared to those without diabetes. The mean HbA1c level for those with diabetes is 6.93, whereas the mean HbA1c level for individuals without diabetes is 5.40.

The amount of sugar that is present in your blood at a specific moment is referred to as your high blood glucose level. Individuals who are diagnosed with diabetes typically have blood glucose levels that are greater than those of individuals who do not have diabetes. People who have diabetes have an average blood glucose level of 194.09 mg/dL, while people who do not have diabetes have an average diabetes-free blood glucose level of 132.85 mg/dL. All of the data presented in the following visualization pertains exclusively to diabetic individuals.

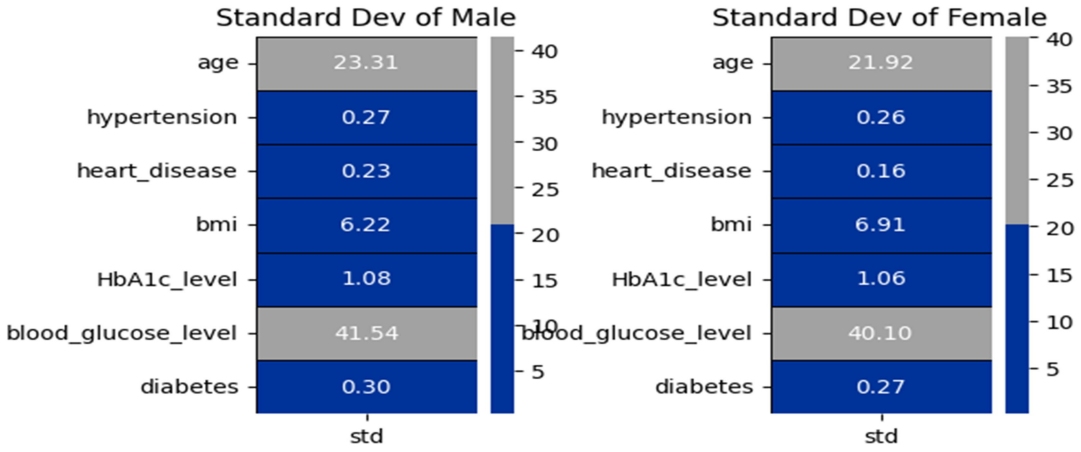

Figure 4: Shows the Standard Deviation of Males and Females With Respect to their Factors.

Interpretation

The graphic displays the standard deviations for blood glucose, diabetes, cardiovascular disease, and hypertension in both males and females. The standard deviation is a statistical measure of the extent to which individual data points differ from the mean. When the standard deviation is small, the data is more concentrated around the mean; when the standard deviation is great, the data is more widely spread across the entire range.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

In the graphic, the standard deviation for males is typically smaller than the standard deviation for females for most of the measures. The results show that, for most assessments, the female data is more scattered than the male data. Take blood pressure as an example; while women have a standard deviation of 4.0, men have a standard deviation of 3.31. It is important to take into consideration the following unique findings on the standard deviation for males and females:

The Age

When compared to one another, the standard deviation of boys and girls is nearly comparable (23.31 against,21.92). There is little difference in the standard deviation between males and females when it comes to hypertension (0.27 versus,0.26). People with cardiovascular disease have a lower standard deviation of 0.23 compared to women, who have a standard deviation of 0.16. Males have a lower body mass index (BMI) than females, and the standard deviation of their BMI is 6.22 compared to 6.91. The little difference in standard deviations, with boys having a value of 1.08 and girls having a value of 1.06, makes it extremely difficult to differentiate between the HbA1c readings of boys and girls.

Men often exhibit lower blood glucose levels compared to women, primarily due to their smaller standard deviation (41.54 vs. 40.10). The findings indicated that male individuals with diabetes had a smaller standard deviation value compared to female individuals with diabetes (0.30 versus 0.27).

Table 2: Shows the Five Individuals With Diabetes Indicator 1.

	6	26	38	40	53
gender	Female	Male	Male	Male	Female
age	44.000000	67.000000	50.000000	73.000000	53.000000
hypertension	0	0	1	0	0
heart_disease	0	1	0	0	0
smoking_history	never	not current	current	former	former
bmi	19.310000	27.320000	27.320000	25.910000	27.320000
HbA1c_level	6.500000	6.500000	5.700000	9.000000	7.000000
blood_glucose_level	200	200	260	160	159
diabetes	1	1	1	1	1

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

Interpretation

The depiction unambiguously indicates that five individuals are diabetic, as denoted by a "1" in the diabetes indicator column. Age range: The individuals you will encounter here fall within the age bracket of 44 to 73. In comparison to type 1 diabetes, which is less prevalent among adults, this indicates a higher likelihood of getting type 2 diabetes. A history of heart illness is associated with an increased chance of acquiring diabetes, and two of the individuals had this condition. In terms of smoking patterns, the group consists of one current smoker, two former smokers who have since stopped, one person whose smoking history is unclear, and one person who has never smoked. The risk of developing diabetes is increased when insulin resistance is worsened by smoking. The normal range for blood glucose levels is between 159 and 260 mg/dL, which all individuals have above. This strongly suggests that you have diabetes.

Table 3: Shows the Five Individuals with Diabetes Indicator 0.

Items	0	1	2	3	4
gender	Female	Female	Male	Female	Male
age	80.000000	54.000000	28.000000	36.000000	76.000000
hypertension	0	0	0	0	1
heart_disease	1	0	0	0	1
smoking_history	never	No Info	never	current	current
bmi	25.190000	27.320000	27.320000	23.450000	20.140000
HbA1c_level	6.600000	6.600000	5.700000	5.000000	4.800000
blood_glucose_level	140	80	158	155	155
diabetes	0	0	0	0	0

Interpretation

Five individuals are included, with ages ranging from 28 to 80 years old, with Heart Disease which is potentially a risk factor for diabetes, looking at the smoking history indicator, Smoking habits could also be a potential risk factor for diabetes. Consider the possibility of non-fasting blood sugar readings. If concerned, consult a healthcare professional for a complete evaluation, including HbA1c testing.

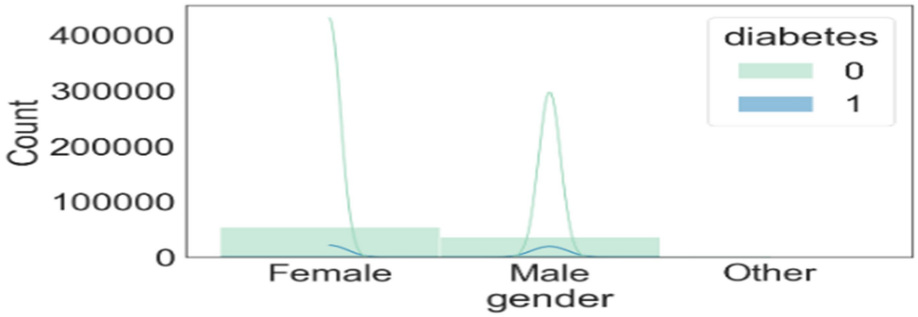

Figure 5: Pictorial representation for gender and diabetes indication

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

Interpretation

Age
The x-axis represents age, ranging from 0 to 80 years old in increments of 10.

Gender
The y-axis represents the number of people with diabetes. There are two separate lines, one for females and one for males.

Data Points
The lines intersect at various points on the graph, indicating the number of people with diabetes for each age group and *gender*. For example, the blue line shows that around 2,000 females between 60 and 70 years old have diabetes in this dataset.

General Trends
The graph suggests an increase in the number of people with diabetes as age increases for both genders. Females appear to have a higher prevalence of diabetes across all age groups compared to males in this dataset.

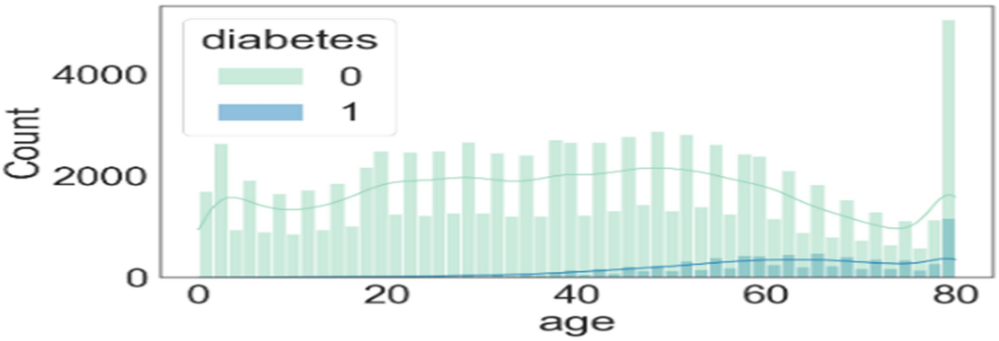

Figure 6: Pictorial Representation for Age And Diabetes Indication.

Interpretation
Age: The x-axis represents age, ranging from 40 to 80 years old. *Number of People with Diabetes:* The y-axis represents the count or number of people with diabetes.

Data Points
The bars on the graph represent the number of people with diabetes in each age group (e.g., the bar at 50 indicates the number of people with diabetes between 40 and 50 years old). There seems to be a possible trend of an 'increase' in the number of people with diabetes as age increases. Overall, the visualization suggests a possible correlation between age and diabetes, with a higher number of patients in the older age groups (between 60 and 80 years old).

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

Model Training and Evaluation

Linear Regression Model Report

Table 4: Showing the Regression model, model score, predicted sum, training, and testing accuracy.

Predicted Sum	Model Score	Predicted	Training Accuracy	Testing Accuracy
2547	1.0	-1 & 1	1.0	1.0

LGBM Classifier Model Report

Table 5: Showing the Number of positive and negative values captured during training, the total bins values, pavg, number of features, and model score

Number of posit ive	number of negat ive	Total Bins	pavg	number of used features	Model Score
5953	64047	2	0.085043	1	1.0

Metrics Classification Report:

Table 6: Showing the Metrics Classification report on the x_train, y_train, model, and predicted x_train values.

	precision	recall	f1-score	support
0 1	1.00 1.00	1.00 1.00	1.00 1.00	64047 5953
accuracy			**1.00**	70000
macro avg	1.00 1.00	1.00 1.00	1.00	70000
weighted avg			1.00	70000

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 11, No 2, June 2025 Series

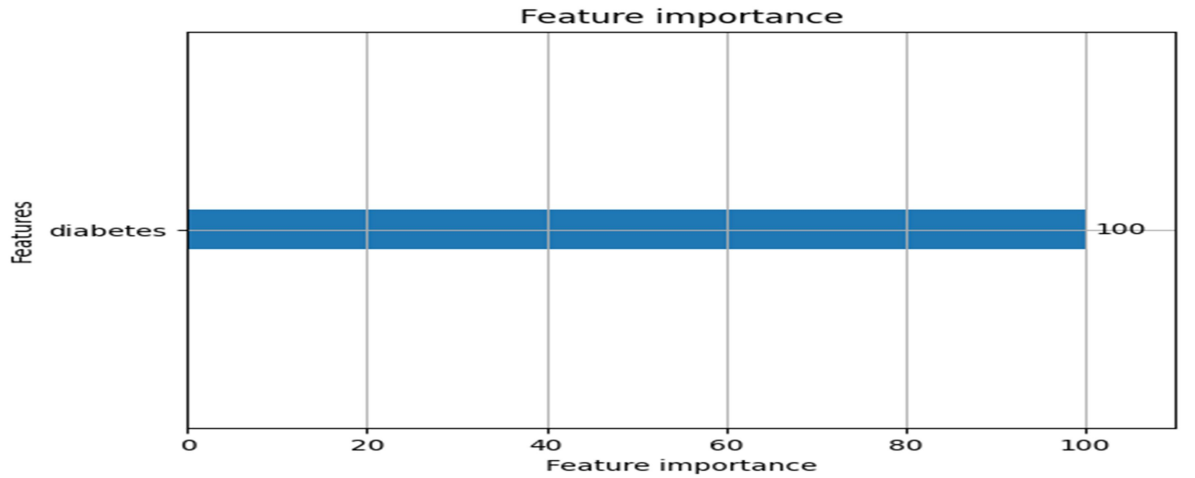

Figure 7: Shows the feature importance with an indication for diabetes as the target variable.

Top Features:

Fbs
This most likely refers to "fasting blood sugar." Fasting blood sugar levels are a strong indicator of diabetes risk. High fasting blood sugar levels suggest the body is struggling to regulate blood sugar effectively. *Glucophage:* This likely refers to a medication used to treat type 2 diabetes. The model might be using the presence or absence of this medication as a feature to identify individuals potentially at higher risk.

Age
Age is a well-established risk factor for diabetes. The risk of developing type 2 diabetes increases as people age.

Adiponectin
This is a hormone produced by fat cells. Lower levels of adiponectin are associated with an increased risk of diabetes

Hhb
This may refer to "glycated hemoglobin," also known as HbA1c. HbA1c is a test that reflects average blood sugar control over some time and is a crucial indicator for diabetes diagnosis.
The feature importance scores suggest that the model relies heavily on factors like fasting blood sugar levels, age, medications used to treat diabetes, and potentially HbA1c levels to predict diabetes risk. This aligns with our understanding of the known risk factors for diabetes.

4. CONCLUSION

Enhance predictive accuracy and robustness in diabetes mellitus disease prediction and type classification. By leveraging the strengths of LightGBM's ability to capture intricate patterns and nonlinearities, and linear regression's proficiency in identifying linear relationships, a hybrid ensemble approach promises to offer a more comprehensive and reliable framework for predictive modeling in healthcare applications. Further exploration and refinement of such hybrid strategies could pave the way for more effective clinical decision support systems and personalized treatment protocols in the future. In essence, our research contributes to the burgeoning field of predictive analytics in diabetes care by unraveling the intricate interplay between feature importance, model complexity, and predictive performance. biomarkers, and yield even greater predictive performance and interpretability

REFERENCES

[1] Ali A., & Fankreldeen A. S. (2021): A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications.
[2] BShamreen A. M. e'tal (2022) Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers.
[3] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data.
[4] Derara D. R. et al (2021) Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM), Diagnostics, (Basel). 11(9): 1714
[5] Lundberg, S. M., & Lee, S. I. (2017). "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems.
[6] Sheetal M. & Sukhvinder S.D (2023): Classification of Machine Learning Techniques for Diabetic Diseases Prediction, International Journal of Computer Science and Network Security, VOL.23 No.12,
[7] Saeedi, P., et al (2019): Statistical highlights the significant global burden of diabetes, indicating the prevalence of the condition among adults in the specified age range
[8] Tianqi C & Carlos G (2016), XGBoost; A Scalable Tree Boosting System" published in the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in 2016; Vol 22
[9] Franklin Okorodudu et al (2021):Visualizing and analyzing data on Covid-19 pandemic outbreak in Nigeria; Nigerian Journal of Science and Environment, Vol.19
[10] Maureen I. Akazue et al (2023):Machine Learning Survival Analysis Model for Diabetes Mellitus;Volume 8, Issue 4, April 2023 International Journal of Innovative Science and Research Technology
[11] Franklin Okorodudu et al (2024):Monkey Pux Data: Visualization and Prediction of the Observed Number of Affected People in Nigeria;Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)