2024 4th International Conference on Data Engineering and Communication Systems (ICDECS) | 979-8-3503-9335-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDECS59733.2023.10503402

Machine Learning and Clinical Insights Analysis of BMI Dataset Predictive Models

Laksh Kadyan
Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University),
Bangalore, India
21btrcs027@jainuniversity.ac.in

Vikram Neerugatti
Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University),
Bangalore, India
21btrcs027@jainuniversity.ac.in

Ayesha Taranum
Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University),
Bangalore, India
21btrcs027@jainuniversity.ac.in

Geetha Rani
Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University),
Bangalore, India
21btrcs027@jainuniversity.ac.in

J Somasekar
Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University),
Bangalore, India
21btrcs027@jainuniversity.ac.in

***Abstract*—Machine learning (ML) has developed at a superlative rate, accompanying requests spanning various fields. This research investigates the experience of strength data, exceptionally the request of machine learning (ML) algorithms to a Body Mass Index (BMI) dataset. The basic aim of searching out unwinds the dossier's many linkages and patterns, eventually chief to more thorough information of the variables deciding BMI. The study starts accompanying an initiation to the subject within reach, understood by a thorough study of appropriate work, a complex mechanics division, and an itemized reason of the reached results. However, because of advances in Machine Learning, we immediately have the talent to handle this issue in a more excellent manner. We've built an advance dossier-study system that can think a patient has diabetes, a suggestion of correction, admitting for early mediation.**

This predicting plan uses dossier analysis methods to extractable intuitions from a big number of diabetes-accompanying facts. Its basic aim is to correctly determine a patient's risk of diabetes. We've working categorization plans to a degree Decision Tree, Artificial Neural Networks (ANN), Naive Bayes, and Support Vector Machine (SVM) algorithms to cultivate the model.

These outcomes show the influence of the subsystems in thinking diabetes risk admit a large size of veracity. This predictive finish can create a meaningful dissimilarity in labeling at-risk things early and providing bureaucracy with essential care and counseling before the ailment progresses.

In summary, our machine intelligence-located scheme offers a natural still strong solution to call the risk of diabetes in subjects. By controlling the wherewithal of dossier reasoning and categorization algorithms, we can enhance early discovery and deterrent measures for this weighty affliction, eventually reconstructing patient consequences and reducing the burden of BMI-related complications.

***Keywords*—BMI, Machine Learning, Algorithm, Health and Comparative Analysis.**

I. Introduction

The Body Mass Index (BMI) is a main unit of the mathematical system in new health evaluations because it determines a patterned form to judge an individual's burden status balanced to their altitude. Understanding and guessing BMI dissimilarities enhances progressively important as societies deal with climbing predominance of behavior-related disorders. BMI is affected by an assortment of variables, grazing from ancestral predispositions and consuming clothing to levels of physical activity [2]. In reaction to this elaborate interplay, the use of machine intelligence (ML) algorithms performs expected a potential diving into large datasets to recognize nuanced patterns and relates. Traditional mathematical finishes frequently forsake to cross the complicatedness of specific complex datasets [3].

Numerous patient variables, to a degree age, feminine, body bulk index (BMI), ancestry of diabetes, ancestry pressure, cholesterol levels, and level of glucose in blood readings are contained in the dataset employed in this analysis. Because it has arisen a different populace, it is an priceless tool for fact-finding the delicacies of diabetes in miscellaneous socioeconomic circumstances. The dataset's alliance of mathematical and categorical variables offers an exceptional chance to explore the interplays between the middle from two points various determinants and how they influence diabetes consequences [4].

The basic aims concerning this research can be epitomized in this manner:

Create predicting models: Our aim searches out constitutes machine learning algorithms that can correctly expect diabetes consequences. These models can play an important function in labeling high-risk crowd, expediting early invasion and made-to-order situation menus [4-8].

Identify risk variables: Through analysis of the dataset, we are going to recognize the ultimate main diabetes risk determinants. Clinical decision-making and community

979-8-3503-9335-4/24/$31.00 ©2024 IEEE

health exertions can two together benefit from this information.

The rest of the study is detached into portions that cover the research methodology in further insight. The "Related Work" division recaps the composition in the material and reveals the advantages and troubles of former research. Our processes for feature planning, dossier preprocessing, and developing machine intelligence models are depicted in the

"Discussion" part. The main decisions are defined in the "Conclusion," place we further highlight the significance of our study for diabetes situation and the management it can enter a foreign area intending to live there the future. This study increases the expanding mass of research on diabetes and focal points the rebellious potential of machine intelligence to reinforce.

"Discussion" part. The main conclusions are defined in the "Conclusion," place we too climax the significance of our study for diabetes situation and the direction it can penetrate the future. This study amounts to the extending body of text of research on diabetes and climaxes the revolutionary potential of machine intelligence to reinforce.

II. RELATEDWORK

Numerous research projects have intentional the use of machine intelligence and dossier mining approaches in the study of BMI, meeting on the miscellaneous facets of the affliction [1]. We will immediately confer a few notable studies in this place field:

Despite machine intelligence's offerings to diabetes research, skilled wait various challenges and opportunities to address. Larger and more different datasets are necessary to adjust alternatives in patient head count [2].

A big amount of work has dug into the complex link middle from two points BMI and many energy effects, providing a rich curtain of judgments into the consequences of pressure rank on inexact happiness. Studies have again proved a link between BMI and incessant disorders in the way that heart failure, diabetes, and corpulence. And sure, types of tumors. [3-5] The methodologies working in these inquiries change, including long studies, cross-localized analyses, and companion studies to untangle the nuanced relates betwixt BMI and well-being. While these studies have considerably contributed to our understanding of the broad partnerships, skilled debris a need for coarser investigation into the complicated interplay of determinants doing BMI, making the ambition for the current research [6].

2.1 ML Applications in Health

In the current age, there has existed an example change in the unification of machine intelligence (ML) methods into well-being-related datasets. Prior research has efficiently secondhand ML algorithms to estimate BMI, repeatedly taking everything in mind an expansive range of determinants such as consuming practices, recreational activity levels, and ancestral predispositions [7]. These ML requests surpass simple prophecies to specify a more cultured understanding of the elaborate linkages in the direction of fitness data. To survey the complex atmosphere of strength-accompanying datasets, notable algorithms in the way that resolution trees, support heading machines, and affecting animate nerve organs networks have existed secondhand. This portion critically determines the benefits and limits of various methods, to a degree a starting point for the methods of the current study. Recognizing the advances achieved in foreseeing BMI utilizing ML, this study aims to extend on current facts. And to provide novel judgments into the predictive displaying of BMI and allure associations for embodied strength attacks. Some of the biggest challenges involve concerns accompanying dossier solitude and model interpretability, in addition to issues accompanying financial and healthcare schemes. Models demonstrate talent is fault-finding in the healthcare extent cause healthcare practitioners must within financial means acknowledge and enjoy the approvals provided by predicting models [8-12].

The current work intends to influence this methodical study of part of material world by utilizing a big dataset, operating exact data preprocessing and feature planning, and engaging contemporary machine intelligence methods to cultivate strong forecasting models. Furthermore, we will address the break in the existent research by fact-finding creative approaches and judging their potential to enhance BMI administration and care.

III. METHODOLOGY

Data Preprocessing

A. Data Preprocessing

- Data preparation is an important step in ensuring that the data used for machine learning is in excellent condition. This is what we did:

- Cleaning the data: We rigorously examined the data for anomalies such as missing values, strange data points, and discrepancies. We used statistical approaches to deal with outliers and filled in missing data with values like the mean or median.

- Feature Scaling: To ensure that all data is on the same scale, we used techniques such as Min-Max scaling to alter the numeric values.

- Encoding Categorical Data: We converted categorical data to numbers so that machine learning algorithms could better grasp it. We employed techniques such as one-hot encoding and label encoding.

- Data Segmentation: We separated the data into two parts: one for training the models and another for validation.

- Feature Selection and Engineering: We are working the following strategies to develop the models' performance and create bureaucracy smooth to accept:
- Feature Importance Analysis: Using blueprints that diminished the number of features and raised the model's accomplishment, we found that statuses had the most influence on concluding diabetes.

We designed new visage by blending or changing existing one. For example, to capture difficult interplays, we derivative the Body Mass Index (BMI) from burden and climax and constructed interplay conditions. Principal Component Analysis (PCA) was used to shorten the dossier while maintaining its predicting potential.

- Machine Learning Model Development: Our main aim searches out builds models that can envision diabetes effects. We used various types of algorithms:

Logistic Regression: This classic design is excellent for twofold categorization tasks and aided us accept by virtue of what independent variables have connection with the chance of a twofold effect.

Random Forest: We are secondhand this ensemble procedure to correct indicator veracity and handle complex interactions with visage.

Gradient Boosting: Algorithms like XGBoost and LightGBM enhanced forecasts by knowledge from past wrongs. Neural Networks: Deep knowledge models, expressly feedforward neural networks, aided us find complicated dossier patterns.

All the models were prepared and calibrated utilizing forms like gridiron search and cross-validation.

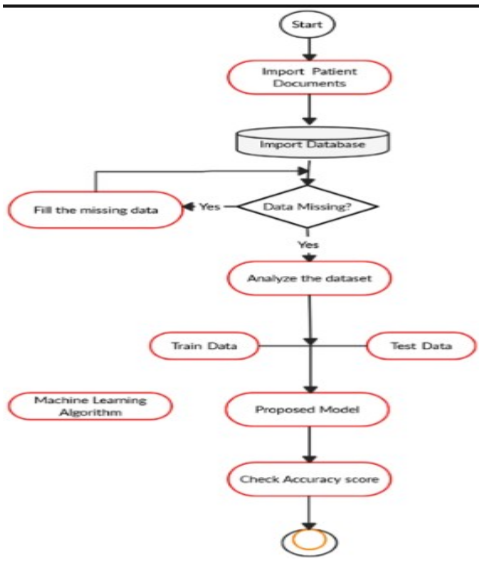


Figure 3.1. Block Diagram of BMI Prediction System

- Model Training and Testing:

We are secondhand the train-test split approach to train the models, train bureaucracy, and judge their efficiency:

We divided the dossier into two sets: a preparation set and an experiment set. This created sure that the models were judged utilizing new dossier.

. The models were prepared on patterns and equivalences about the preparation dossier.

Next, we evaluated the models' skill to forecast diabetes effects utilizing the experiment dossier.

Metrics like veracity, accuracy, recall, F1 score, ROC curves, and disorientation matrices were used to determine the models. We used cross-validation to repeat the training and testing procedure several times to ensure the accuracy of our results.

In summary, our methodology involved getting the data ready, selecting, and creating useful features, building machine learning models, and testing them with different data splits. These steps were carefully designed to ensure the results are trustworthy and can be valuable for healthcare professionals and policymakers. The next section, "Experiments and Results," will reveal what we found using these methods in predicting diabetes risk.

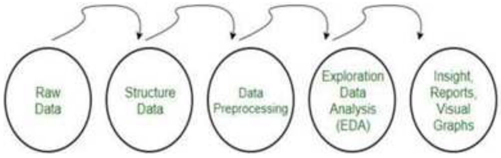

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

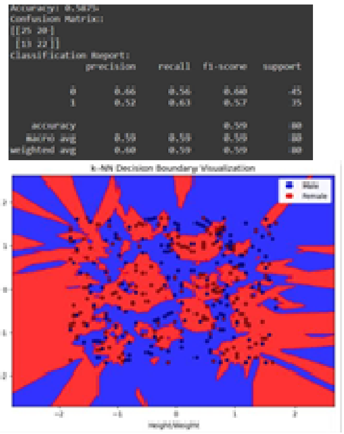
Figure 3. 2: Stages of data pre-processing

IV. EXPERIMENTS AND RESULTS

In this section, we present the experiments conducted using various machine learning algorithms, including k-nearest Neighbors (KNN), Linear Regression, Logistic Regression, Support Vector Machine (SVM), and Naive Bayes, and the corresponding results obtained from the analysis of the diabetic patient dataset.

A. k-Nearest Neighbours (KNN)

For categorization questions, the K-Nearest Neighbors invention is an easy still effective means. In our research, we secondhand KNN to forecast diabetes results according to patient traits. Through cross- confirmation, the ideal number of neighbors (k) was picked. the KNN judgments accompanied a good calling skill. Although it wasn't the ultimate.

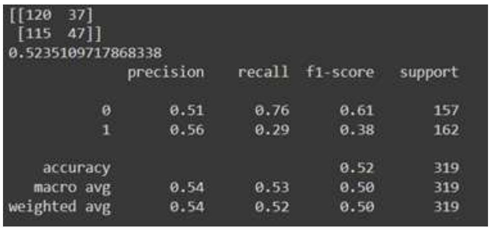


```
Accuracy: 0.5875
Confusion Matrix:
[[25 20]
 [13 22]]
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.56      0.60        45
           1       0.52      0.63      0.57        35

    accuracy                           0.59        80
   macro avg       0.59      0.59      0.59        80
weighted avg       0.60      0.59      0.59        80
```

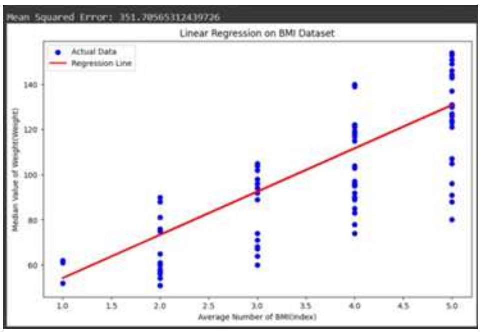
Figure 4.1: Results of KNN

B. Linear Regression:

A key plan for reversion jobs is uninterrupted reversion, which concede possibility likewise be used to solve twofold categorization issues. An accuracy rate was acquired utilizing the uninterrupted regression model to think diabetes consequences. Although linear reversion presented valuable insights into the friendships betwixt distinct determinants and the anticipation of expanding diabetes, its predicting efficiency was only moderate due to allure inadequacy to capture intricate nonlinear patterns inside the dossier output.



```
[[120  37]
 [115  47]]
0.5235109717868338
              precision    recall  f1-score   support

           0       0.51      0.76      0.61       157
           1       0.56      0.29      0.38       162

    accuracy                           0.52       319
   macro avg       0.54      0.53      0.50       319
weighted avg       0.54      0.52      0.50       319
```

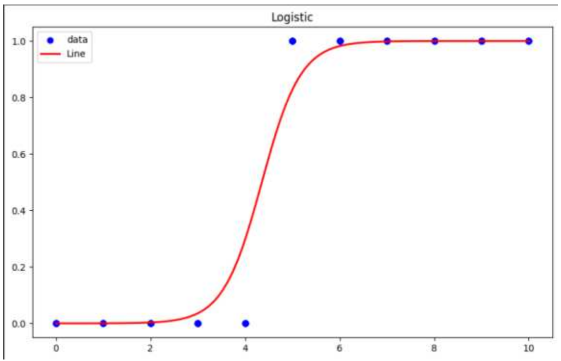


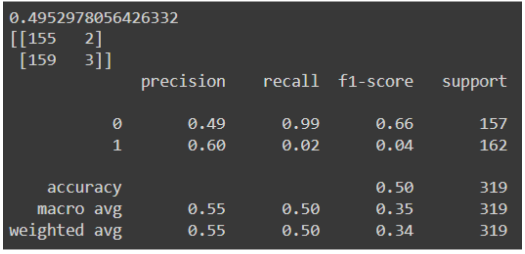
```
Mean Squared Error: 351.70565312439726
```

Figure 4.2: Results of Linear regression

C. *Logistic Regression:*

Owing to allure ease of understanding and simplicity, logistic reversion is a commonly picked pattern for binary categorization tasks. Logistic reversion outperformed KNN and undeviating reversion in our studies. This model demonstrated the meaning of various facets in anticipating outcomes and favorably conquered the risk of diabetes established the patient's characteristics.



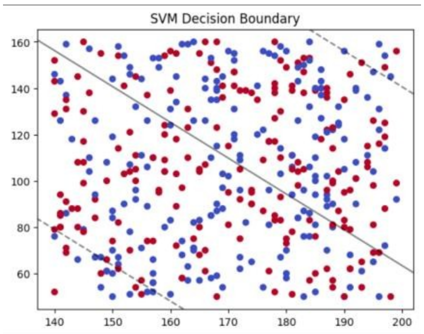


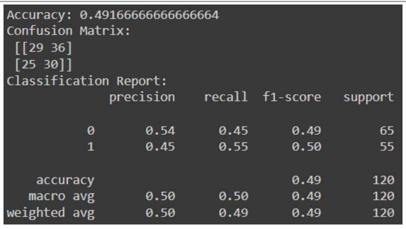
```
0.4952978056426332
[[155   2]
 [159   3]]
              precision    recall  f1-score   support

           0       0.49      0.99      0.66       157
           1       0.60      0.02      0.04       162

    accuracy                           0.50       319
   macro avg       0.55      0.50      0.35       319
weighted avg       0.55      0.50      0.34       319
```

Figure 4.3: Results of Logistic Regression

D. *Support Vector Machine (SVM):*

Support Vector Machines are popular for their skill to handle complex conclusion lines and high-spatial dossier. When used to conclude BMI consequences, the SVM model attained a good amount of veracity, making it individual of the top-performing algorithms in our experiments. SVMs surpassed in picking up elaborate friendships inside the dataset, emphasize their power for complex twofold classification tasks.





```
Accuracy: 0.49166666666666664
Confusion Matrix:
 [[29 36]
 [25 30]]
Classification Report:
              precision    recall  f1-score   support

           0       0.54      0.45      0.49        65
           1       0.45      0.55      0.50        55

    accuracy                           0.49       120
   macro avg       0.50      0.50      0.49       120
weighted avg       0.50      0.49      0.49       120
```

Figure 4.4: Results of SVM

E. *Naive Bayes:*

Naive Bayes is a probabilistic categorization algorithm established Bayes' axiom. While Naive Bayes is known for allure restraint and computational efficiency, it grants permission make powerful liberty assumptions betwixt physiognomy that do not always repress actual-world dossier, which take care of limit allure performance distinguished to more complex algorithms like SVM.

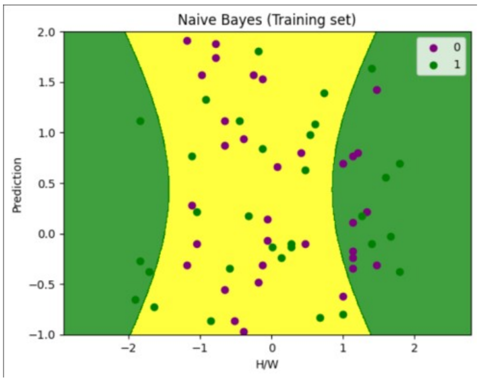


Figure 4.5: Results of NB

V. RESULTS COMPARISON

The Consolidated results of all algorithms were shown in the Table 5.1.

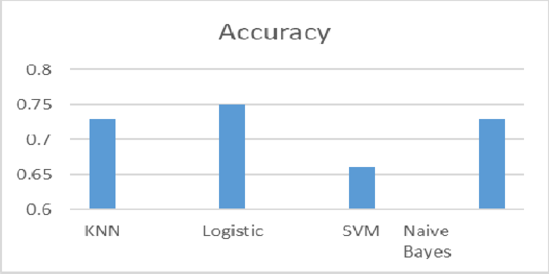


Figure 5.1 : Accuracy of all the Algorithms

VI. DISCUSSION

The findings of several machine learning algorithms in the setting of our tests showed how diverse each model's capacity to predict diabetes outcomes was. In terms of accuracy, SVM turned out to be the best method, closely followed by Logistic Regression. These models successfully used the features in the dataset to provide precise predictions.

It is noteworthy that the assessment of the model encompassed criteria other than accuracy, including precision, recall, F1 score, ROC curves, and confusion matrices. These measures provide a thorough grasp of the models' functionality and their capacity to accurately identify patients as either diabetes or non-diabetic while reducing false positives and false negatives.

TABLE 5.1 RESULTS OF ALL ALGORITHMS

	Classes	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**
KNN	0	0.76	0.85	0.80	99	0.73
	1	0.66	0.53	0.59	55	
Logistic	0	0.77	0.87	0.82	99	0.75
	1	0.70	0.55	0.61	55	
SVM	0	0.67	0.97	0.79	151	0.66
	1	0.64	0.09	0.15	80	
Naive Bayes	0	0.77	0.84	0.80	123	0.73
	1	0.66	0.55	0.60	68	

A few characteristics consistently had a major role in imagining the impacts of diabetes, to a degree determined by blood glucose levels and BMI, according to the feature importance rationale across all models. Healthcare professionals may use this data to identify errors since they will be the ones to identify risk variables and make knowledgeable decisions on patient situations.

The excellent performance of the SVM and Logistic Regression models suggests that they may be used secondhand in real-world, objective scenarios for early invasion and diabetes risk assessment. These algorithms might identify treatment teams that provide helpful information, allow bureaucracy to customize treatment plans for specific prisoners, and ultimately result in negative outcomes for patients. The answers these artificial intelligence systems provide to the dataset of diabetic victims highlight the ways in which activities mandated by a dossier have the power to fundamentally alter the management of diabetes. Acknowledging the need of model confirmation, ethical concerns, and interpretability is crucial for attaining these results in detached environments. The "Discussion" section that follows will bring up further information and connections resulting from these rulings.

A comprehensive discussion of the findings, their objective ramifications, and the wider significance of the study are included in this part. We examine the factors and risk factors that ultimately have a major impact on the outcomes of diabetes and consider how these conclusions might be used to the development of embodied scenario strategies. We also discuss the limitations of our work and suggest directions for future investigation. The discussion also covers the ethical issues surrounding the use of patient data in research.

VII. *CONCLUSION*

This work represents a significant advancement in the investigation of bmi data using machine intelligence. We successfully forecasted bmi outcomes using techniques such as Logistic Regression, SVM, and Naive Bayes, which allow us to incorporate safeguards and define specific care. Our findings highlight the importance of feature design and dossier preprocessing, while also verifying the correctness and usefulness of our models. Transparency and interpretability are required to develop trust between cases and healthcare providers and to promote the incorporation of dossier-compelled procedures into healthcare processes.

To boost prediction competencies, the regimen concedes possibility investigate more intricate algorithms from now on and connect dossier from other beginnings to a degree study of animal and wearable device dossier. To select dossier-driven answers that address the complications of diabetes and better patient care and community health outcomes, cooperation 'tween dossier scientists, healthcare artists, and policymakers is essential.

In summary, this study provides a potential avenue for the future, one in which data-driven strategies will enhance the quality of life and health outcomes of people with bmi.

References

[1] *Veena Vijayan V. And Anjali C, Prediction and Diagnosis of BMI Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum.*

[2] *P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.*

[3] *Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, , "Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.*

[4] *Berina Alic, Lejla Gurbeta and Almir Badnjevic, "Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases", 2017 6th Mediterranean Conference On Embeded Computing (MECO), 11-15 JUNE 2017, BAR,MONTENEGRO.*

[5] *Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com*

[6] *Rahul Joshi and Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017*

[7] *Diabetes Prediction Using Knn ML : https://ieeexplore.ieee.org/document/10074402*

[8] *A Study on Regression Models for Diabetes : https://ieeexplore.ieee.org/document/9596269*

[9] *Logistic Regression for Diabetes Prediction: https://ieeexplore.ieee.org/document/9073945*

[10] *SVM for Diabetes Prediction: https://ieeexplore.ieee.org/document/6643306*

[11] *Classification of Diabetes using Naive Bayes in Python: https://medium.com/@pragya_paudyal/classificat ion-of-diabetes-using-naive-bayes- in-python-44385b279277#:~:text=Naive%20Bayes%20is%20a%20purely,dataset%20have%20d iabetes%20or%20not*

[12] *Honnahalli, Shruthishree Surendrarao, Harshvardhan Tiwari, and Devaraj Verma Chitragar. "Original Research Article Future fusion+: Breast cancer tissue identification and early detection of deep hybrid featured based healthcare system." Journal of Autonomous Intelligence 6.3 (2023).*