IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 9, September 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|

Analyzing Key Factors in Diabetes Detection: A Machine Learning Perspective

Dr. Dinesh D. Patil, Miss. Mayuri R. Chandratre

Associate Professor & HOD, Department of Computer Engineering, Shri Sant Gadge Baba, College of Engineering & Technology, Bhusaval, India

Shri Sant Gadge Baba, College of Engineering & Technology, Bhusaval, India

ABSTRACT: The early detection of diabetes holds paramount significance in improving patient outcomes and reducing the overall healthcare burden. This research paper explores the application of a diverse set of machine learning algorithms for the accurate detection of diabetes. Leveraging the power of K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression, this study contributes to the understanding of how different algorithms perform in diagnosing diabetes. The paper presents comprehensive experimental results and comparative analyses, shedding light on the strengths and limitations of each algorithm in this critical medical domain.

KEYWORDS: data mining machine learning

I. INTRODUCTION

Diabetes Ailment (DA) is a metabolic disorder characterized by chronic hyper glycaemia with disturbances of carbohydrate, fat and protein metabolism. There are three main types of diabetes ailment (DA). Type 1 DA results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes Ailment" (IDDA) or "juvenile diabetes". Type 2 DA results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as non insulin-dependent diabetes ailment (NIDDA) or "adult-onset diabetes". The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. As of 2000 it was estimated that 171 million people globally suffered from diabetes or 2.8% of the population. Type-2 diabetes is the most common type worldwide. Figures for the year 2007 show that the 5 countries with the largest amount of people diagnosed with diabetes were India (40.9 million), China (38.9 million), US (19.2 million), Russia (9.6 million), and Germany (7.4 million)[1]. Due to the growing unstructured nature of diabetic data form health industry or all other sources, it is necessary to structure and emphasis its size into nominal value with possible solution. With the help of technological developments, it is necessary to combine robust diabetic data sharing and electronic communication systems can facilitate better access to health services at all the levels of patients. So that all patient data are needs to be in one repository. Deploying a Health Information Exchange (HIE) can extract clinical information from several disparate repositories and integrate that data within a single patient health record that all care providers can access securely. Predictive Analysis is a method, that incorporates a variety of techniques from data mining, statistics, and game theory that uses the current and past data with statistical or other analytical models and methods, to determine or predict certain future events [6]. Significant predictions or decisions can be made by employing big data analytics in health care field. In this paper, we use the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

Diabetes mellitus is a chronic metabolic disorder that has witnessed a staggering global rise in prevalence over recent decades. The implications of undiagnosed or inadequately managed diabetes are profound, leading to severe health complications and increased mortality rates. Machine learning, with its ability to learn patterns from complex data, has emerged as a promising tool for early disease detection and diagnosis. This study focuses on harnessing the potential of a wide array of machine learning algorithms to accurately identify diabetes cases.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|

The primary goal of this research is to evaluate the performance of various algorithms in diagnosing diabetes. We consider a comprehensive set of algorithms, including K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression. Each algorithm brings a unique approach to classification, and through a comparative analysis, we aim to understand their relative merits and areas of applicability in the context of diabetes detection.

II. RELATED WORK

P.Yasodha and M. Kannan [2] This paper uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred and forty nine instances with seven attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests.

N. NiyatiGupta,A.Rawal, and V.Narasimhan [3] It aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms.

III. PROPOSED METHODOLOGY

K-Nearest Neighbors (KNN):

- Import the KNeighborsClassifier from a machine learning library like scikit-learn.
- Choose an appropriate value for 'k' (number of neighbors) through cross-validation.
- Fit the KNN model on the training dataset.
- Use the trained model to predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

Logistic Regression:

- Import LogisticRegression from a machine learning library.
- Initialize the model with suitable hyperparameters.
- Fit the logistic regression model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

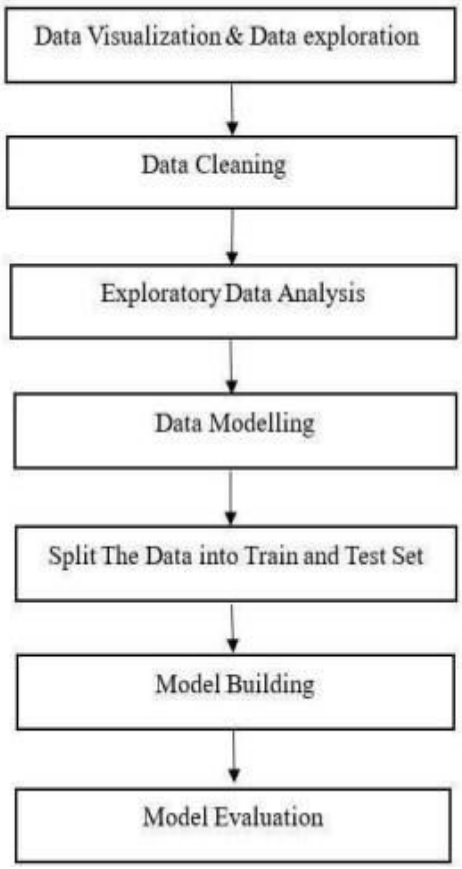
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|



Naive Bayes:

- Import GaussianNB from a machine learning library.
- Instantiate the Gaussian Naive Bayes model.
- Fit the model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

Linear Discriminant Analysis (LDA):

- Import LinearDiscriminantAnalysis from a machine learning library.
- Create an instance of the Linear Discriminant Analysis model.
- Fit the model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

Decision Tree:

- Import DecisionTreeClassifier from a machine learning library.
- Initialize the decision tree model.
- Fit the model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

Random Forest:

- Import RandomForestClassifier from a machine learning library.
- Create a Random Forest model instance with appropriate hyperparameters.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|

- Train the model on the training dataset.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

AdaBoost with Random Forest:
- Import AdaBoostClassifier and RandomForestClassifier from the library.
- Create an instance of the RandomForestClassifier for the base estimator.
- Instantiate the AdaBoostClassifier with the base estimator.
- Fit the model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score.

AdaBoost with Logistic Regression:
- Import AdaBoostClassifier and LogisticRegression from the library.
- Create an instance of the LogisticRegression for the base estimator.
- Instantiate the AdaBoostClassifier with the base estimator.
- Fit the model on the training data.
- Predict diabetes outcomes on the testing dataset.
- Evaluate the model using accuracy, precision, recall, and F1-score

Dataset Description
- No null values found in the data set.
- Dimensions of the dataset: (253680, 22)
- Number of people in Diabetes Class: 35346
- Number of people in Non-Diabetes Class: 218334
- Random sample of 5000 rows each from the two classes, meaning a total of 10000 rows.
- Supervised Learning.

IV. CONCLUSION

In this study, we embarked on an exploration of various machine learning algorithms for the critical task of diabetes detection. Through extensive experimentation, we gained insights into the performance and applicability of K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression in diagnosing diabetes.

Our findings revealed that each algorithm presents a unique approach to diabetes detection, with varying degrees of accuracy and interpretability. KNN exhibited competitive results by leveraging neighborhood information, while Logistic Regression provided a simpler yet effective model. Naive Bayes demonstrated its strengths in probabilistic modeling, while Linear Discriminant Analysis showcased its potential in capturing class separation. Decision Tree offered transparency in decision-making, and Random Forest excelled in ensemble-based classification. AdaBoost with Random Forest and AdaBoost with Logistic Regression showcased the power of boosting techniques in improving the performance of base classifiers.

V. FUTURE SCOPE

The present study opens avenues for further research and development in the domain of diabetes detection using machine learning. Some potential directions for future investigations include:
- **Ensemble Refinement:** Investigate more sophisticated ensemble strategies, hybridizing different algorithms to exploit their combined strengths and mitigate weaknesses.
- **Feature Engineering:** Explore advanced feature engineering techniques to enhance the discriminatory power of the models, potentially using domain-specific knowledge.
- **Deep Learning Integration:** Integrate deep learning architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to leverage the power of complex feature extraction.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|

- **Interpretability Enhancement:** Develop techniques to enhance the interpretability of complex models like Random Forest or AdaBoost, making their decisions more comprehensible to medical professionals.
- **Domain Adaptation:** Extend the study to adapt the trained models to different demographic groups, as diabetes detection might vary across populations.
- **Real-time Deployment:** Implement a user-friendly interface for healthcare practitioners to input patient data and receive automated diabetes risk assessments.
- **Multi-Class Classification:** Extend the binary classification task to multi-class classification, accommodating more comprehensive diagnosis scenarios.
- **Longitudinal Data Analysis:** Explore the effectiveness of the proposed algorithms in analyzing longitudinal patient data for predicting diabetes progression and identifying risk factors.

REFERENCES

[1] P. T. Katzmarzyk, C. L. Craig, and L. Gauvin, "Adiposity, physical fitness and incident diabetes: The physical activity longitudinal study," Diabetologia, vol. 50, no. 3, pp. 538–544, Mar. 2007.
[2] Z. Xu, X. Qi, A. K. Dahl, and W. Xu, "Waist-to-height ratio is the best indicator for undiagnosed type 2 diabetes," Diabetic Med., vol. 30, no. 6, pp. e201–e207, Jun. 2013.
[3] R. N. Feng, C. Zhao, C. Wang, Y. C. Niu, K. Li, F. C. Guo, S. T. Li, C. H. Sun, and Y. Li, "BMI is strongly associated with hypertension, and waist circumference is strongly associated with type 2 diabetes and dyslipidemia, in northern Chinese adults," J. Epidemiol., vol. 22, no. 4, pp. 317–323, May 2012.
[4] A. Berber, R. G´omez-Santos, G. Fangh¨anel, and L. S´anchez-Reyes, "Anthropometric indexes in the prediction of type 2 diabetes mellitus, hypertension and dyslipidaemia in a Mexican population," Int. J. Obes. Relat Metab. Disorders, vol. 25, no. 12, pp. 1794–1799, Dec. 2001.
[5] B. Balkau, D. Sapinho, A. Petrella, L. Mhamdi, M. Cailleau, D. Arondel, and M. A. Charles, D. E. S. I. R. Study Group, "Prescreening tools for diabetes and obesity-associated dyslipidaemia: Comparing BMI, waist and waist hip ratio. The D.E.S.I.R. Study," Eur. J. Clin. Nutr., vol. 60, no. 3, pp. 295–304, Mar. 2006.
[6] I. S. Okosun, K. M. Chandra, S. Choi, J. Christman, G. E. Dever, and T. E. Prewitt, "Hypertension and type 2 diabetes comorbidity in adults in the United States: risk of overall and regional adiposity," Obes. Res., vol. 9, no. 1, pp. 1–9, Jan. 2001.
[7] L. A. Sargeant, F. I. Bennett, T. E. Forrester, R. S. Cooper, and R. J.Wilks, "Predicting incident diabetes in Jamaica: the role of anthropometry," Obes. Res., vol. 10, no. 8, pp. 792–798, Aug. 2002.
[8] N. T. Duc Son le, T. T. Hanh, K. Kusama, D. Kunii, T. Sakai, N. T. Hung, and S. Yamamoto, "Anthropometric characteristics, dietary patterns and risk of type 2 diabetes mellitus in Vietnam," J. Amer. Coll. Nutr., vol. 24, no. 4, pp. 229–234, Aug. 2005.
[9] G. T. Ko, J. C. Chan, C. S. Cockram, and J.Woo, "Prediction of hypertension, diabetes, dyslipidaemia or albuminuria using simple anthropometric indexes in Hong Kong Chinese," Int. J. Obes. Relat. Metab. Disorders, vol. 23, no. 11, pp. 1136–1142, Nov. 1999.
[10] M. B. Snijder, P. Z. Zimmet, M. Visser, J. M. Dekker, J. C. Seidell, and J. E. Shaw, "Independent and opposite associations of waist and hip circumferences with diabetes, hypertension and dyslipidemia: The AusDiab study," Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 3, pp. 402–409, Mar. 2004.
[11] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," IEEE J. Biomed. Health Informat., vol. 18, no. 2, pp. 555–561, Mar. 2014.
[12] L. de Koning, H. C. Gerstein, J. Bosch, R. Diaz, V. Mohan, G. Dagenais, S. Yusuf, and S. S. Anand, EpiDREAM Investigators, "Anthropometric measures and glucose levels in a large multi-ethnic cohort of individuals at risk of developing type 2 diabetes," Diabetologia, vol. 53, no. 7, pp. 1322–1330, Jul. 2010.
[13] I. S. Okosuna and J.M.Boltrib, "Abdominal obesity, hypertriglyceridemia, hypertriglyceridemic waist phenotype and risk of type 2 diabetes in American adults," Diabetes Metab. Syndrome, vol. 2, no. 4, pp. 273–281, Dec. 2008.
[14] Z. Yu, L. Sun, Q. Qi, H. Wu, L. Lu, C. Liu, H. Li, and X. Lin, "Hypertriglyceridemic waist, cytokines and hyperglycaemia in Chinese," Eur. J. Clin. Invest., vol. 42, no. 10, pp. 1100–1111, Oct. 2012.
[15] T. Du, X. Sun, R. Huo, and X. Yu, "Visceral adiposity index, hypertriglyceridemic waist and risk of diabetes: The china health and nutrition survey 2009," Int. J. Obes. (Lond.), vol. 38, no. 6, pp. 840–847, Jun. 2014.
[16] M. Solati, A. Ghanbarian, M. Rahmani, N. Sarbazi, S. Allahverdian, and F. Azizi, "Cardiovascular risk factors in males with hypertriglycemicwaist (tehran lipid and glucose study)," Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 5, pp. 706–709, May 2004.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309116|

[17] I. Lemieux, A. Pascot, C. Couillard, B. Lamarche, A. Tchernof, N. Alm´eras, J. Bergeron, D. Gaudet, G. Tremblay, D. Prud'homme, A. Nadeau, and J. P. Despr´es, "Hypertriglyceridemic waist: A marker of the atherogenic metabolic triad (hyperinsulinemia; hyperapolipoprotein B; small, dense LDL) in men?" Circulation, vol. 102, no. 2, pp. 179–184, Jul. 2000.

[18] L. B. Tank´o, Y. Z. Bagger, G. Qin, P. Alexandersen, P. J. Larsen, and C. Christiansen, "Enlarged waist combined with elevated triglycerides is a strong predictor of accelerated atherogenesis and related cardiovascular mortality in postmenopausal women," Circulation, vol. 111, no. 15, pp. 1883–1890, Apr. 2005.

[19] I. F. Gazi, T. D. Filippatos, V. Tsimihodimos, V. G. Saougos, E. N. Liberopoulos, D. P. Mikhailidis, A. D. Tselepis, and M. Elisaf, "The hypertriglyceridemic waist phenotype is a predictor of elevated levels of small, dense LDL cholesterol," Lipids, vol. 41, no. 7, pp. 647–654,
Jul. 2006.

[20] J. St-Pierre, I. Lemieux, M. C. Vohl, P. Perron, G. Tremblay, J. P. Despr´es, and D. Gaudet, "Contribution of abdominal obesity and hypertriglyceridemia to impaired fasting glucose and coronary artery disease," Amer. J. Cardiol., vol. 90, no. 1, pp. 15–18, Jul. 2002.

[21] P. Blackburn, I. Lemieux, N. Alm´eras, J. Bergeron, M. Cˆot´e, A. Tremblay, B. Lamarche, and J. P. Despr´es, "The hypertriglyceridemic waist phenotype versus the national cholesterol education program-adult treatment panel III and international diabetes federation clinical criteria to identify high-risk men with an altered cardiometabolic risk profile," Metabolism, vol. 58, no. 8, pp. 1123–1130, Aug. 2009.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details