INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Diabetes Prediction and Recommendation Model Using Machine Learning Techniques and MapReduce

Ritika Bateja[1]*, **Sanjay Kumar Dubey**[2], **Ashutosh Kumar Bhatt**[3]

1 Ph.D. Scholar, Department of Computer Science and Engineering, ASET, Amity University Uttar Pradesh, Noida, India
2 Associate Professor, Department of Computer Science and Engineering, ASET, Amity University Uttar Pradesh, Noida, India
3 Associate Professor, School of Computer Sciences and Information Technology, Uttarakhand Open University, Haldwani, Uttarakhand, India

OPEN ACCESS

Received: 02-03-2024

Accepted: 07-06-2024

Published: 06-07-2024

Citation: Bateja R, Dubey SK, Bhatt AK (2024) Diabetes Prediction and Recommendation Model Using Machine Learning Techniques and MapReduce . Indian Journal of Science and Technology 17(26): 2747-2753. https://doi.org/10.17485/IJST/v17i26.530

* **Corresponding author**.

ritikalpr@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To deliver patient centric healthcare for diabetic patients using a fast and efficient diabetic prediction and recommendation model which will not only help in early diagnoses of disease but also recommend appropriate medicine for controlling it at stage 1. **Methods:** The Support Vector Machine Classifier is further enhanced with Particle Swarm Optimization (PSO) and used for the prediction of diabetes. Collaborative Filtering is used for drug recommendation, which produces a suitable list of medications that correspond to the diagnoses of diabetes patients. Improved Density-Based Spatial Clustering of Applications with Noise (I-DBSCAN) is proposed to cluster EHR data to get labels based on the symptoms of patients and map reduction is utilized to process the clustered data in parallel for quick recommendations. **Findings:** The accuracy of the SVM with the PSO model is 99.20%. The performance of I-DBSCAN is also compared with K-Means and regular DBSCAN using the Silhouette Score, Davies Bouldin Score, and the Calinski Harabasz Score. Also, I-DBSCAN was found to give a more accurate score. **Novelty:** The extensive volume of diabetes-related information stored in electronic health records (EHRs) through continuous monitoring devices poses a growing difficulty for healthcare professionals to effectively navigate and deliver patient-centered care. Machine Learning techniques like classification and recommendations can be utilized to facilitate early disease diagnosis and recommend appropriate medications.

Keywords: Electronic health records (EHRs); Collaborative Filtering (CF); Recommendations; Improved Density Based Spatial Clustering of Applications with Noise (IDBSCAN); SVM classifier

1 Introduction

Healthcare advancements use machine learning techniques like Recommender Systems (RSs) to diagnose and control chronic diseases like diabetes. These systems use extensive data from health institutions to suggest products based on patient profiles and preferences[1]. According to[2], the percentage of internet users who utilize social networks to search for health information is currently around 80%, and this number is rising continuously. Users can find people who share the same symptoms, learn about their ailments' potential origins, locate treatments for that disease, develop new healthy behaviors, and access general health information online[3].

RSs aid healthcare professionals in prescribing medications for chronic conditions like diabetes by analyzing patient data, symptoms, and medical history, and also aid in medical research by analyzing user ratings and reviews. An RS for cervical cancer was put up by Kuanr et al.[4] and showed high prediction model accuracy. To fight problems including malnutrition, obesity, and cardiovascular diseases, Poornima[5] created a daily nutrition RS for women that took into account physical data, preferences, and personal information. Other studies have mostly concentrated on recommending medications, doctors, and hospitals that are best matched to a particular patient profile, treatment suggestions for patients over time, health-related films, and even personalized meal plans. The application of RSs in diabetes has recently been the subject of various research[6], including some exploratory analysis of the condition, and forecasts of diet programs to combat diabetes. Furthermore, clustering is another popular machine-learning method for identifying patterns in patients with similar features. However, it's not widely adopted in the medical field. To provide personalized care and assist medical professionals in choosing medications, pharmacological RS based on clustering algorithms is recommended. The proposed research uses Collaborative Filtering (CF) based recommendation systems, which suggest items based on user's evaluations, rather than knowledge-based or content-based approaches.Table 1 presents different recommendation systems, techniques followed, and their advantages.

Table 1. Analysis of different recommender system, their techniques and advantages

References	Methods Used	Techniques Followed	Advantages
R. Yera et al.[7]	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020)	The present study creates a meal recommender system survey for individuals with diabetes.	The creation of a clear framework employing semantic technologies for research and development projects pertaining to dietary advice for diabetics.
Batalha et al.[8]	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), Cochrane's tool and grading of recommendations, assessment, development, & evaluation (GRADE).	The purpose of this study was to determine the features of treatments used to encourage behavior change in individuals with type 2 diabetes mellitus (T2DM).	Improvement in disease self-management and A1c level.
B. Janakiraman et al.[9]	Personalized Nutrition Recommendation for Diabetic Patients Using Optimization Techniques	The health records of patients with diabetes were examined, and recommendation algorithms were used to propose proper nutrition for enhancing their health.	Optimization improves the performance and accuracy during recommendations.
Bhat et al.[10]	Machine Learning Algorithms	Diabetes is diagnosed using Machine Learning Techniques and the right diet is recommended through a Diet Recommender System (DRS).	The pre-processing method provides a higher average accuracy for Naïve Bayes.
Gong et al.[11]	SMR Drug Recommendation System	Designed by bridging medical knowledge graphs and electronic medical records (EMRs) to create a high-quality heterogeneous graph	Used joint-learning-embedding models and heterogeneous graphs in to produce a drugs list
Yang et al.[12]	Neural Networks & Graph Representation	On the basis of drug interactions, drug molecule structures, procedure codes, and diagnosis codes	Compared to simply using diagnosis and procedure codes, this engine displayed some signs of improvement

Traditional categorization approaches were utilized in the majority of the drug recommendation studies which include Collaborative Assessment and Recommendation Engine (CARE), Markov model[13], Collaborative Filtering (CF)[14,15], and Bayesian methods[16]. The ever-increasing quantity of medical records, as well as the desire to use those records to better inform clinical decision-making, prompted the current investigation. The study evaluates collaborative filtering and classification methodologies for drug recommendation, hypothesizing improved DBSCAN clustering for better systems. Results can be used for clinical decision-making in drug prescription and verification.

2 Methodology

2.1 Proposed Methodology

The ML model for diabetes prediction and recommendation proposed in Figure 1 trains two models on two different datasets. One is for diabetes prediction and the other uses collaborative filtering to recommend medicines to patients. Diabetes Health Indicator Datasets (Teboul A. Diabetes Health Indicators Dataset. 2021) and Drug Recommendations Datasets (Cop C. Drug Recommendations. 2021) have been used here for prediction and recommendations respectively.

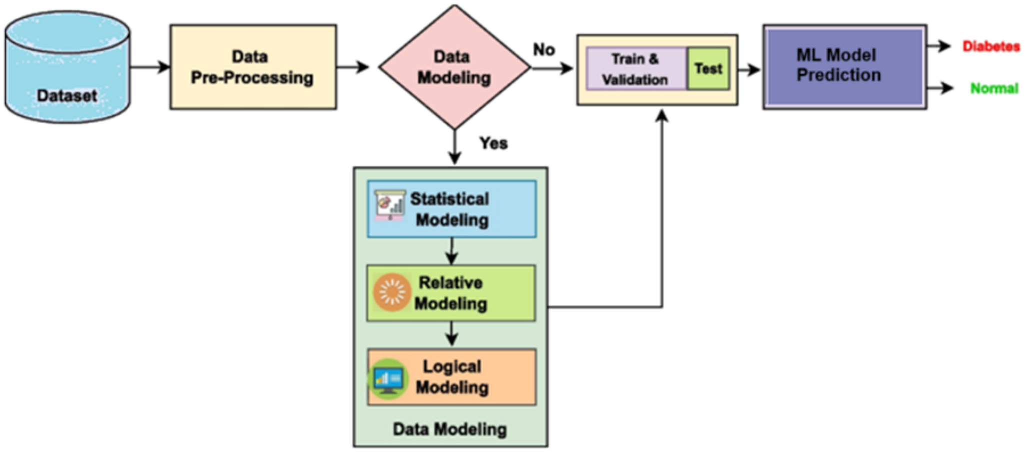

Fig 1. Diabetes Prediction and Recommendation Model

2.1.1 Dataset

The health indicator dataset contains 253,680 survey responses and includes three classes for diabetes: 0 (no diabetes), 1 (prediabetes), & 2 (diabetes), with an uneven distribution of classes. The dataset contains 23 columns with information like high BP, high cholesterol, BMI, smoking, Stroke, etc.

2.1.2 Data Pre-processing

Pre-processing involves data cleaning, null value removal, and unique count calculation, followed by Exploratory Data Analysis (EDA), dimensionality reduction, feature selection, and extraction, with irrelevant columns excluded for clustering and further processing.

2.1.3 Training and Testing of Data

The dataset is trained and tested using different classifiers like SVM, KNN, Decision Tree, Naive Bayes, and Adaboost, based on datatype, training data quantity, and performance scores. Table 2 represents the performance metrics of the state of the art classifiers on the raw data before preprocessing.

Table 2. Comparison of Performance Metrics on the Raw Data or Base Model

Models	Accuracy	Recall	Precision	F1-Score
SVM	**86.60%**	**85.87%**	**83.58%**	**83.74%**
KNN	83.44%	83.44%	80.53%	81.66%
Decision Tree	85.87%	85.87%	73.74%	79.34%
Naive Bayes	77.44%	77.44%	80.02%	78.61%
Adaboost	85.87%	86.60%	73.74%	83.58%

2.1.4 Proposed Improved DBSCAN Clustering Algorithm

A unique clustering approach for fuzzy samples is proposed. The algorithm is categorized under several indexes, and MapReduce is used for clustering. Classifiers are then utilized for training and testing. The following is one example of the pseudocode for the Improved DBSCAN **Algorithm 1**:

Algorithm 1: Improved DBSCAN Algorithm:

Inputs:

- **D**: The dataset with 9 integer columns.
- **eps:** The radius (maximum distance) for defining the neighbourhood of a data point.
- **MinPts:** The minimum number of data points required to form a dense region (including the point itself).
- **tau_min:** The minimum similarity threshold for merging clusters.

Outputs:

- **ClusterID**: A label indicating the cluster to which each data point belongs, or -1 for noise.

Initialization:

1. Initialize an empty list **visited** to keep track of visited data points and **clusters** to store the clusters.
2. Initialize **ClusterID** to 0.

Improved DBSCAN Algorithm:

```
1. For each unvisited data point P in the dataset D:
a. Mark P as visited
b. Find all data points in the $\varepsilon$-neighborhood of P. Form a cluster if there are at least MinPts data points within this neighborhood.
c. If the $\varepsilon$-neighborhood of P contains fewer than MinPts, data points, mark P as noise and continue to the next data point.
2. For each cluster formed in the previous step:
a. Assign a unique ClusterID to all data points in the cluster
b. Increment ClusterID for the next cluster
3. Tau_min, merge L followed by F into a single cluster
Result:
```

- The clusters and noise points are identified, and each data point is assigned to a cluster or labeled as a mapped index.

3 Results and Discussion

The SVM classifier has been observed to have a high level of accuracy, however, it is not as accurate as literature classifiers. The adoption of the PSO technique results in an improvement in the performance of the SVM. Because of the interaction between particles, PSO provides a reduced number of tuning elements. As a result, it is able to find the optimal solution at a slower speed due to the high dimensionality of the search space. Immediately following the implementation of the PSO, the SVM with the PSO model demonstrates an accuracy of 99.20%. In Table 3, this study presented a comparison of performance metrics for SVM with PSO, as well as for state-of-the-art classifiers, after they have been applied to the processing, clustering, and mapping of data.

Table 3. Comparison of Performance Metrics after the Application of Processing, Clustering and Mapped Data

Models	Accuracy	Recall	Precision	F1-Score
SVM+PSO	**99.20%**	**99.20%**	**99.22%**	**98.82%**
SVM	98.77%	98.77%	98.79%	98.39%
KNN	98.61%	98.61%	98.63%	98.08%
Decision Tree	98.77%	98.77%	98.79%	98.39%
Naive Bayes	98.77%	98.77%	98.79%	98.39%
Adaboost	98.77%	98.77%	98.79%	98.39%

3.1 Medicine Recommendations

Another dataset acquired from the Kaggle repository includes the name of the drug, as well as ratings and reviews from users as part of the process of recommending treatment. Following the implementation of collaborative filtering, the average ratings are calculated to be 20.79205095135711. The number of reviews that meet the quantile is 62.0. When a set of measurements is provided in order to estimate a value that belongs to a specific percentile, techniques known as percentile methods or quantile methods are applied. On the basis of the choices that were made previously, the suggestions that are recommended are presented in Table 4.

Table 4. Final Recommendations

Drug Name	Correlation
Acarbose	0.9382
ActoPlus Met	0.9105
Actos	0.8731
Aflibercept	0.8392
Afrezza	0.7934

3.2 Comparison of K Means and DBSCAN with IDBSCAN

As can be seen in Table 5, the performance metrics of K-Means and DBSCAN are compared with those of IDBSCAN. When compared to K-Means and DBSCAN, it is clear that IDBSCAN provides the highest performance score for the SVM classifier. As a result, it provides superior recommendations when utilizing the recommendation system.

Table 5. Comparison of Performance Metrics

Algorithm	Accuracy	Recall	Precision	F1-score
K Means	98.40	98.40	96.82	97.60
DBSCAN	98.66	98.66	98.68	98.19
IDBSCAN	**99.77**	**99.77**	**99.79**	**99.29**

Using the Silhouette score, the Davies Bouldin score, and the Calinski Harabasz score, additional comparisons are made between the clustering methods K-means, DBSCAN, and IDBSCAN. The results of the comparison of clustering algorithms employing the different scores are presented in Table 6.

Table 6. Comparison of Clustering Algorithms

Algorithms	Silhouette Score	Davies Bouldin Score	Calinski Harabasz Score
K-means	0.361	-0.446	0.486
DBSCAN	0.983	1.566	2.547
IDBSCAN	**1.264**	**0.306**	**18.360**

In order to properly manage massive amounts of dataset samples, Ramani et al. (2020) [17] have presented a modified artificial neural network (ANN) classification technique that utilizes MapReduce. This investigation makes use of the MapReduce technique, which is implemented in conjunction with an artificial neural network (ANN) classifier that has been updated. With the help of a trained persistent artificial neural network on the Pima Indian diabetes machine learning repository dataset, the objective is to successfully acquire the output that was predicted. Prognosticating the occurrence of diabetic chronic illness is the objective of this endeavor. An accuracy level of around 99.6% is achieved by the approach that has been proposed. It is possible to get improved output accuracy with the MapReduce method that has been suggested because of its dynamic architecture and linear scalability. According to the research conducted by Arun and Marimuthu (2024)[(18)], the MapReduce-based CapsNet system enables the most accurate categorization of diabetic conditions from massive amounts of data collection. SuiTab training of the MapReduce-based CapsNet helps enhance the classification of diabetes data and determines the risk profile of a patient. When it comes to classification accuracy, recall rate, and F-score, the simulation performed on the test datasets reveals that the MapReduce-based CapsNets framework performs better than traditional MapReduce and deep learning approaches such as RNN and DenseNet. Diabetes has emerged as a major global health concern, according to the authors Modak et al. (2023)[(19)]. Diabetes is associated with a number of serious complications, such as renal disease, loss of vision, and cardiovascular issues. Several other ensemble methods, including XGBoost, LightGBM, CatBoost, Adaboost, and Bagging, were studied in this work. CatBoost stands out as the most effective ensemble strategy among those that were assessed. It has an astounding accuracy rate of 95.4%, which is higher than the accuracy rate of 94.3% that XGBoost achieved. In addition, CatBoost's AUC-ROC score of 0.99, which is much higher than XGBoost's score of 0.98, provides additional evidence that CatBoost has the potential to be more authoritative than XGBoost.

4 Conclusion

This study concludes that machine learning techniques can provide timely predictions and recommendations for diabetic patients, enhancing patient-centric care and aiding medical professionals in making effective clinical decisions. Drug recommendation systems learn from the diagnostic and prescription data that is already stored in an EHR system in order to provide physicians with drug recommendations that correspond to the patient's diagnostic concerns. The dataset after IDBSCAN and map-reduce process has been trained and tested by using various machine learning classifiers. It is seen that SVM optimized with PSO gives the highest accuracy, precision, recall, and F1 Score. At the clustering stage, the IDBSCAN is compared with K-Means and DBSCAN using the Silhouette Score, Davies Bouldin Score and the Calinski Harabasz Score and it has been found that IDBSCAN leads in all the scores. After the dataset is trained and tested, the recommendation system using collaborative filtering is used and the required drugs are prescribed for the diabetes patients.

As part of future research, neural networks, and precision medicine can be considered for further training and testing the dataset. Deployability, prediction delay, clinical coding changes, and long-term maintainability are some of the practical and implementation concerns that need to be taken into consideration in the future.

References

1) Roy D, Dutta M. A systematic review and research perspective on recommender systems. *Journal of Big Data*. 2022;9:59–60.
2) Almuammar SA, Noorsaeed AS, Alafif RA, Kamal YF, Daghistani GM. The Use of Internet and Social Media for Health Information and Its Consequences Among the Population in Saudi Arabia. *Cureus*. 2021;13(9):18338. Available from: https://doi.org/10.7759/cureus.18338.
3) Jia X, Pang Y, Liu LS. Online Health Information Seeking Behavior: A Systematic Review. *Healthcare (Basel)*. 2021;9(12):1740–1740. Available from: https://doi.org/10.3390/healthcare9121740.
4) Kuanr M, Mohapatra P, Piri J. Health Recommender System for Cervical Cancer Prognosis in Women. In: 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India;vol. 2021. 2021;p. 673–679. Available from: https://doi.org/10.1109/ICICT50816.2021.9358540.
5) Princy J, Senith S, Kirubaraj AA, Vijaykumar PO. A personalized food recommender system for women considering nutritional information. *Int J Pharmaceut Res*. 2021;13(2). Available from: https://doi.org/10.1109/ICICT50816.2021.935678.
6) Almatrooshi F, Alhammadi S, Salloum SA, Akour I, Shaalan K. A recommendation system for diabetes detection and treatment. In: International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), Sharjah, United Arab Emirates. 2020;p. 1–6. Available from: https://doi.org/10.1109/CCCI49893.2020.9256676.
7) Yera R, Alzahrani AA, Martínez L, Rodríguez RM. A Systematic Review on Food Recommender Systems for Diabetic Patients. *Int J Environ Res Public Health*. 2023;20(5):4248–4248. Available from: https://doi.org/10.3390/ijerph20054248.
8) Batalha A, Ponciano IC, Chaves G, Felício DC, Britto RR, Silva LPD. Behavior change interventions in patients with type 2 diabetes: a systematic review of the effects on self-management and A1c. *J Diabetes Metab Disord*. 2021;20(2):1815–1836. Available from: https://doi.org/10.1007/s40200-021-00846-8.
9) Janakiraman B, Saradha A. Personalized Nutrition Recommendation for Diabetic Patients Using Optimization Techniques. *Intelligent Automation & Soft Computing*. 2020;26(2):269–280. Available from: https://doi.org/10.31209/2019.100000150.
10) Bhat SS, Ansari GA. Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques. In: and others, editor. 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India. 2021. Available from: https://doi.org/1109/INCET51464.2021.9456365.
11) Gong F, Wang M, Wang H, Wang S, Liu M. SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. *Big Data Res*. 2021;23:100174. Available from: https://doi.org/10.1016/j.bdr.2020.100174.
12) Yang C, Xiao C, Ma F, Glass L, Sun J. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In: In Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI 2021, Montreal, QC, USA. 2021;p. 19–27. Available from:

https://doi.org/10.24963/ijcai.2021/514.

13) Carvalho DFD, Kaymak U, Van Gorp P, Van Riel N. A Markov model for inferring event types on diabetes patients data. *Healthc Analytics*. 2022;2:100024–100024. Available from: https://doi.org/10.1016/j.health.2022.100024.
14) Tan WY, Gao Q, Oei RW, et al. Diabetes medication recommendation system using patient similarity analytics. *Sci Rep*. 2022;12:20910–20910. Available from: https://doi.org/10.1038/s41598-022-24494-x.
15) Morales LFG, Valdiviezo-Diaz P, Reátegui R, Barba-Guaman L. Drug Recommendation System for Diabetes Using a Collaborative Filtering and Clustering Approach: Development and Performance Evaluation. *J Med Internet Res*. 2022;24:37233–37233. Available from: https://doi.org/10.2196/37233.
16) Kidwell KM, Roychoudhury S, Wendelberger B. Application of Bayesian methods to accelerate rare disease drug development: scopes and hurdles. *Orphanet J Rare Dis*. 2022;17(186). Available from: https://doi.org/10.1186/s13023-022-02342-5.
17) Ramani R, Devi KV, Soundar KR. MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. *Soft Comput*. 2020;24(21):16335–16380. Available from: https://doi.org/10.1016/j.bdr.2020.100174.
18) Arun G, Marimuthu CN. Diabetes classification using MapReduce-based capsule network. *Automatika*. 2024;65(1):73–81. Available from: https://doi.org/10.1016/j.bdr.2020.100190.
19) Modak SK, Jha VK. Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications*. 2023;p. 1–27. Available from: https://doi.org/10.1016/j.bdr.2020.103456.