Diabetes Prediction System using Support Vector Machine: A case study of the female Pima Indian Population.

Blaise Tanam FONGUH
Department of Computer Science
Kwame Nkrumah University of Science and Technology
btfonguh.st.knust.edu.gh

Yaw MARFO MISSAH
Department of Computer Science
Kwame Nkrumah University of Science and Technology
ymissah.cos@knust.edu.gh

***Abstract*—**This study aims to explore and analyze the Pima Indians Diabetes Database to gain insights into the factors influencing the occurrence of diabetes within the Pima Indian population. By examining key variables, such as age, BMI, glucose levels, and pregnancy history, this research seeks to identify potential risk factors associated with diabetes and develop predictive models to aid in early diagnosis and intervention. In this light, we extracted its dataset from Kaggle and then performed preprocessing on the data. The resulting dataset was evaluated on a supervised machine learning based approach – Support Vector Machine (SVM) model, based on Accuracy, Precision, Recall, F1-Score and Confusion matrix. The best performance was observed by Support Vector Machine – Accuracy Score with an Accuracy of 77.27%. We finished with an analysis and discussion of the results. Our data and code are available at: https://github.com/blaise-fonguh/Diabetes-Prediction-Using-MachineLearning.git

Keywords—Diabetes; Pima India; Prediction; Supervised machine learning; Support Vector Machine; Modelling

I. INTRODUCTION

Diabetes mellitus, a chronic metabolic disorder, continues to be a significant global health concern affecting millions of individuals worldwide. It is a condition in which the pancreas stops producing insulin, the hormone that helps regulate blood sugar levels [1]. Among the various communities involved, the Pima Indians, an ethnic group with a high prevalence of diabetes, have been subjected to intensive study due to their increased vulnerability to this condition [2]. In this context, the mini project aims to develop a Diabetes Prediction System utilizing the powerful Support Vector Machine (SVM) algorithm, focusing on the female Pima Indian population. [3] The Pima Indians Diabetes Database, a publicly accessible dataset, contains valuable information about the health and demographic attributes of female Pima Indians aged 21 years and above. This dataset is a valuable resource for investigating the factors contributing to diabetes development within this population. [4] By harnessing the potential of SVM, a supervised learning algorithm known for its robustness in handling complex classification tasks, we endeavor to construct a predictive model capable of accurately identifying individuals at risk of developing diabetes. [5] This mini project's primary goal is first to conduct an in-depth analysis of the dataset to understand the prevalence and distribution of diabetes among female Pima Indians. This includes exploring correlations between various demographic and health-related features, providing insights into potential risk factors. Secondly, by leveraging the predictive capabilities of SVM, we aim to develop a reliable and efficient Diabetes Prediction System capable of early detection and intervention, thus improving healthcare outcomes for the Pima Indian population [6]. Moreover, we'll also be creating a simple user interface that allows users to input their health data and receive predictions about their diabetes risk [7]. In the following sections, we will discuss the methodology employed for data preprocessing, feature selection, and model development and explain the rationale behind choosing SVM as our classification algorithm [8]. Subsequently, we will present the results obtained from the analysis, showcasing the performance of the developed Diabetes Prediction System. The implications of this research lie in its potential to contribute to the ongoing efforts in combating diabetes among the Pima Indians by facilitating timely diagnosis and targeted medical interventions. It is important to note that while this mini project focuses solely on the female Pima Indian population, the methodology and insights derived from this study could pave the way for future research encompassing broader demographic groups. [9] Furthermore, the lessons learned from this investigation can be extrapolated to other similar high-risk communities, thereby fostering the development of personalized healthcare approaches to combat the prevalence of diabetes on a larger scale. In conclusion, the Diabetes Prediction System presented in this mini project holds promise in addressing the challenges posed by diabetes among the female Pima Indian population. By harnessing the potential of SVM and the wealth of information from the Pima Indians Diabetes Database, we aspire to contribute to the ongoing global efforts to mitigate the burden of diabetes and improve the quality of life for individuals at risk. This mini project's primary goal is first to conduct an in-depth analysis of the dataset to understand the prevalence and distribution of diabetes.

II. RELATED WORK

Diabetes prediction using the Pima Indians Diabetes Database has been a topic of interest among researchers in recent decades. This section highlighted some of the methods used by the research to predict diabetes using the Pima Indians Diabetes Database and the accuracy achieved. According to [10], we learned

The following; Rado et al. [11] used random forest combined with recursive feature elimination, and the accuracy achieved was 73%. Rajni and Amandeep [12] achieved a classification accuracy of 72.9% by using the RB-Bayes algorithm. In this, the mean is used to handle the missing values. Ramana and Boddu [13] used the naïve Bayes classification algorithm, and the accuracy achieved was 76.34%. Kumari and Chitra[14] used SVM with RBF kernel to classify the data and achieved an accuracy of 75.5%. The primary obstacle faced by researchers in various techniques is the need to enhance the precision of early diabetes diagnosis systems. To address this challenge, this study proposes a two-phase fusion technique. The first phase involves data preprocessing utilizing feature scaling standardization, while the second phase entails employing Support Vector Machines for classifying diabetes cases with utmost accuracy. A comprehensive summary of different approaches employed by researchers and their corresponding achieved accuracies can be found in Table I.

TABLE I: COMPARATIVE STUDY OF EXISTING APPROACHES USED BY THE RESEARCHERS AND ACCURACY ACHIEVED.

Sr. no.	Method used	Reference	Accuracy(%)
1	Random forest combined with recursive feature elimination	[11]	73
2	RB-Bayes	[12]	72.9
3	Naïve Bayes	[13]	76.3
4	SVM(with RBF kernel)	[14]	75.5

III. METHODOLOGY

- *Data Collection*

For this project, we got a comprehensive dataset containing relevant features and corresponding diabetes outcomes from Kaggle; Pima Indians Diabetes dataset, made up of several medical KKpredictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age and so on for 768 female patients.

TABLE II: FIRST FIVE RECORDS IN THE PIMA INDIANS DIABETES DATASET

PN	Pregnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- *Data Preprocessing*

Next, we applied the following preprocessing techniques to clean the data: They are:

- Load the dataset Using Pandas and examine its structure.
- Handle missing values by imputing data using appropriate methods like mean, and standard deviation.
- Handle outliers by removing them
- Count the number of diabetic patients (Outcome=1) and non –diabetic (Outcome=0)
- Group the outcomes according to their mean
- Separate the whole data (X) from the Outcomes/Labels (Y) for standardization.
- Standardize the data by bringing them within a similar range using standard scaling.

IV. VALIDATION, EVALUATION, AND COMPARISON

After extracting the features, we split the data, 80% as train data, and 20% as test data by using shuffled sampling with the aim of training the Support Vector Machine(SVM). We then developed the Machine Learning Model-SVM from the library sci-kit-learn, to predict the existence of diabetes in the patients. Prior to feature extraction, we applied the preprocessing techniques highlighted above and performed hyperparameter tuning using GridSearch on our model, to find the optimal set of hyperparameters, which in turn maximizes our model's predictive accuracy. We then compared performance using four evaluation metrics: Accuracy, Precision, Recall, and F1-score after a cross validation experiment.

V. RESULTS

All experiments were performed on Google Colab and programs were written in Python.

Support Vector Machine Approach

TABLE III: ACCURACY SCORE RESULTS FOR TRAINING VS TEST DATA

Classifier	Data	
	Test	Training
SVM	0.7727	0.7866

TABLE IV: OVERALL PERFORMANCE RESULTS FOR "DIABETES" PREDICTION CLASSIFICATION

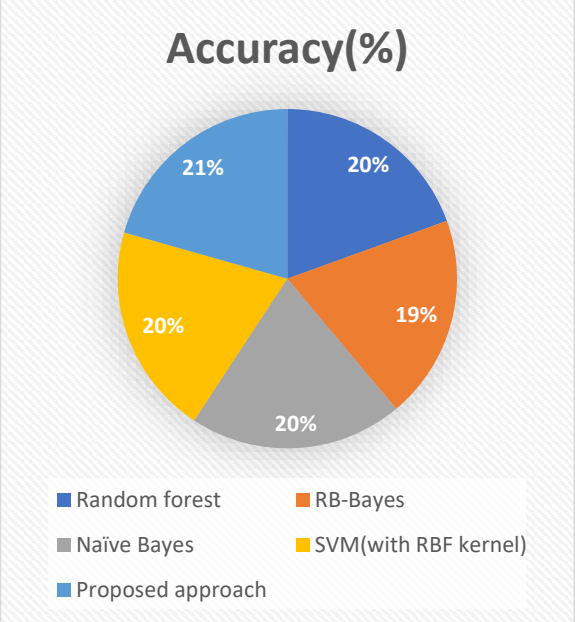
Classifier	Metrics			
	Accuracy	Precision	F1	Recall
SVM	0.7727	0.5185	0.6153	0.7567

Table III shows the performance measures of both our training data and test data in our SVM model. We notice a very little difference between both scores, for two data which performed very well with scores of 78.66% and 77.27% respectively.

Table IX shows how the SVM performed well and managed to have four different scores with its best being the Accuracy score with 77.27%, followed by Recall-75.67%, then f1 with 61.53% and finally its precision score with 51.85%. Comparing this Machine Learning Models to the other Classification models reviewed we have this table;

TABLE V: PERFROMANCE EVALUATION PF SEVERAL CLASSIFICATION MODELS

Methods used	Accuracy(%)
Random Forest	73
RB-Bayes	72.9
Naïve Bayes	76.3
SVM(with RBF kernel)	75.5
Proposed Method	**77.27**



VI. USER INTERFACE

A user-friendly interface was also developed with the help of flask in python, that allows users to input their health-related information .We integrated the trained SVM model into the interface to provide real-time predictions about the user's diabetes risk.

VII. CONCLUSION

This study suggested a brand-new way for diabetes patient categorization using exemplary data preprocessing techniques like; handling outliers, feature scaling, shuffled sampling and SVM. The test data of the database is designed using shuffled sampling which are then imputed into the SVM model. The predictions are then computer based produced from them. Utilizing the Pima Indians Diabetes Database evaluates the durability of our proposed methodology. This database encompasses data from 768 female patients and is widely accessible to the public. The training phase of our machine model relies on 80% of this dataset, while the remaining 20% is employed for testing purposes within the proposed framework.

Remarkably, we achieved a maximum accuracy rate of 77.27%. Future enhancements in classification rates can be anticipated by extracting more dependable attributes from this extensive database. In addition, integrating techniques such as decision fusion across multiple classifiers holds promise in further streamlining and advancing the classification process.

REFERENCES

[1] Type 1 Diabetes Diet: What to Eat and What to Avoid - Business Facts Hub. https://businessfactshub.com/type-1-diabetes-diet-what-to-eat-and-what-to-avoid/

[2] A genome-wide association study using a custom genotyping array identifies variant in GPR158 associated with reduced energy expenditure and increased Body Mass Index in American Indians | NIH Research Festival. https://researchfestival.nih.gov/2015/posters/genome-wide-association-study-using-custom-genotyping-array-identifies

[3] https: PIMA Indians Diabetes Database. (2016, October 6). Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[4] SVM-Awad, M., & Khanna, R. (2015). Support vector machines for classification. In Apress eBooks (pp. 39–66). https://doi.org/10.1007/978-1-4302-5990-9_3

[5] Shin, T. (2022, November 10). An extensive step by step guide to exploratory data analysis. Medium. https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

[6] Pisner, D., & Schnyer, D. M. (2020). Support vector machine. In Elsevier eBooks (pp. 101–121). https://doi.org/10.1016/b978-0-12-815739-8.00006-7

[7] 5 Best Blood Pressure Monitoring Apps To Download For Free - TheEMTSpot. https://www.theemtspot.com/5-best-blood-pressure-monitoring-apps-to-download-for-free/

[8] the methodology employed for data preprocessing, and explain the: García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1). https://doi.org/10.1186/s41044-016-0014-0

[9] last part-Petersen, B., Vesper, I., Pachwald, B., Dagenbach, N., Buck, S., Waldenmaier, D., & Heinemann, L. (2021). Diabetes management intervention studies: lessons learned from two studies. Trials, 22(1). https://doi.org/10.1186/s13063-020-05017-3

[10] Arora, N., Singh, A., Al-Dabagh, M. Z. N., & Maitra, S. K. (2022). A Novel architecture for diabetes patients' prediction using K-Means clustering and SVM. Mathematical Problems in Engineering, 2022, 1–9. https://doi.org/10.1155/2022/4815521

[11] O. Rado, N. Ali, H. M. Sani, A. Idris, and D. Neagu, "Performance analysis of feature selection methods for classification of healthcare datasets," in Proceedings of the Intelligent Computing-Proceedings of the Computing Conference, pp. 929–938, Springer, Cham, 2019, July. View at: Google Scholar

[12] R. Rajni and A. Amandeep, "RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset," International Journal of Electrical and Computer Engineering, vol. 9, no. 6, p. 4866, 2019. View at: Publisher Site | Google Scholar

[13] B. V. Ramana and R. S. K. Boddu, "Performance comparison of classification algorithms on medical datasets," in Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0140–0145, IEEE, 2019, January. View at: Google Scholar

[14] V. A. Kumari and R. Chitra, "Classification of diabetes disease using a support vector machine," International Journal of Engineering Research in Africa, vol. 3, no. 2, pp. 1797–1801, 2013. View at: Google Scholar