www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1560-1565

e-ISSN : 2583-1062
Impact Factor : 7.001

COMPARATIVE STUDY OF RANDOM FOREST, LOGISTIC REGRESSION, AND K-NEAREST NEIGHBORS IN DETECTING DIABETES AND POLYCYSTIC OVARY SYNDROME (PCOS)

Sagar Parande[1], Owais Shamsi[2], Dr. Rakhi Gupta[3], Nashrah Gowalkar[4]

[1,2]Master of Science in Information Technology K.C. College, HSNC University, Mumbai 400 020, India.
sagarparande123@gmail.com
shamsiowais23@gmail.com

[3]Head of the Department I.T Department K.C. College, HSNC University, Mumbai 400 020, India.
rakhi.gupta@kccollege.edu.in

[4]Asst Professor I.T Department K.C. College, HSNC University, Mumbai 400 020, India.
nashrah.gowalker@kccollege.edu.in

ABSTRACT

Diabetes and Polycystic Ovary Syndrome (PCOS) are significant health issues that must be diagnosed as early as possible and classified correctly for proper treatment. In this study, three machine learning algorithms were used for the prediction of diabetes, and also PCOS. Data from clinical sources have been applied, including features such as glucose levels, insulin response, body mass index (BMI), and other relevant factors. The performances of the models were judged against the metrics such as accuracy, precision, recall, F1-score and AUC-ROC.

Keywords: Diabetes, Polycystic Ovary Syndrome, Logistic Regression, Random Forest, KNN

1. INTRODUCTION

Millions suffer from chronic diseases like diabetes and PCOS, and these are among the main causes of death and disability around the world. Diabetes is a metabolic condition due to high blood sugar as a result of either a lack of enough or functioning insulin, making complications with the heart, kidneys, nerves, and vision if left unchecked. In the same manner, PCOS in women of reproductive age would increase the risk for Type 2 diabetes, alongside resistive insulin. This project predicts diabetes and PCOS by using machine learning algorithms based on the following health indicators: glucose, insulin, BMI, and age. Early diagnosis with machine learning leads to better outcomes than conventional invasive methods in light of real-time analysis over large datasets.

We will make explicit use of Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) algorithms here as they are more efficient in classification problems. The proposed models will be assessed for their reliability in predicting the probability of diabetes and PCOS based on some selected features. Using a cross-validation technique along with performance metrics evaluation helps us derive the best-suited model for clinical use.

We will push for more transparent model predictions with the emergence of explainable AI techniques. This project will be based on a framework of web and mobile, which will cater to accessible tools for users in predicting and managing diabetes and PCOS. It will aid early intervention and better health management strategies.

2. LITERATURE REVIEW

The two datasets are designed to evaluate the performance of machine learning models on different medical conditions:

1. Diabetes Dataset

Source: A publicly available dataset from Kaggle. [6]

Attributes: This dataset is equipped with features, which include physiology-based attributes containing blood glucose level, insulin, BMI, age, and other physiological factors. These are considered very important for determining a patient's risk for diabetes.

2. PCOS Dataset

Source: A Kaggle dataset detecting Polycystic Ovary Syndrome, PCOS. [7]

Attributes: The dataset is equipped with menstrual cycle regularity, ovarian volume, and follicle size. All of these have a very important role in determining PCOS.

Several studies have applied machine learning techniques to predict diabetes and Polycystic Ovary Syndrome (PCOS). A model developed by Rahman et al. used glucose, BMI, insulin, and other factors, achieving improved classification accuracy with a new dataset [8]. In another study, XGBoost, combined with SMOTE and ADASYN techniques, achieved 81% accuracy in predicting diabetes. The authors also implemented explainable AI techniques to enhance model transparency [9]. Additionally, a comparative study on Random Forest, SVM, Logistic Regression, and Naive

Bayes found Logistic Regression to have the highest accuracy (82.46%) in predicting diabetes [10]. These works highlight the effectiveness of various machine learning algorithms in detecting diabetes and PCOS, emphasizing accuracy and precision as key performance metrics.

3. METHODOLOGY

3.1 Machine Learning Models

1. **Random Forest:** which has been highly sought in the domain of medical diagnostics due to its stability and with its capacity to handle high-dimensional data. In this model, it constructs a large number of decision trees and aggregates predictions, which helps avoid over fitting and increases the prediction's precision. Its studies demonstrate a potent capability to diagnose diabetes. For example, we trained Random Forest on a clinical dataset and achieved high classification accuracy in identifying diabetes patients. A model that easily manages the missing data and imbalanced classes is especially an advantage for medical datasets that may present such challenges.
2. **Logistic Regression** is much simpler than the Random Forest method, it is widely used in medicine because of interpretability: Logistic Regression can easily be used to understand multiple input factors like age, weight, family history, etc., which relate to the probability of developing diseases, such as Type 2 diabetes. We comparatively evaluated Logistic Regression with the K-Nearest Neighbors (KNN) algorithm towards the predictability of diabetes. In that study, although Logistic Regression was easy to interpret with coefficients that can be used to infer the importance of every feature in disease predictions, higher flexibility in data non-linearity was offered by the KNN algorithm.
3. **K-Nearest Neighbours (KNN)** is another well-known algorithm, the KNN is pretty simple yet effective in classifying instances based on how close other data points are to the to-be-classified instance. There are no assumptions needed for the assumption of the distribution of the data and makes it quite powerful for small datasets, but too much complicated for high-dimensional databases, because it depends highly on the distance metric used. For example, we noted that KNN did a better job than Logistic Regression in detecting complex patterns in the dataset to capture some non-linear relationships but then had issues when the dataset was getting large.

Diabetes Dataset:

Age
Gender
BMI
SBP (Systolic Blood Pressure)
DBP (Diastolic Blood Pressure)
FPG (Fasting Plasma Glucose)
FFPG (Final Fasting Plasma Glucose)
FFPG (Final Fasting Plasma Glucose)
Cholesterol
Triglyceride
HDL (High-Density Lipoprotein)
LDL (Low-Density Lipoprotein)
ALT (Alanine Aminotransferase)

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1560-1565

e-ISSN : 2583-1062
Impact Factor : 7.001

BUN (Blood urea nitrogen)
CCR (Creatinine Clearance)
Smoking Status: (1: Current Smoker, 2: Ever Smoker, 3: Never Smoker)
Drinking Status: (1: Current Drinker, 2: Ever Drinker, 3: Never Drinker)
Family History of Diabetes: (1: Yes, 0: No)

PCOS Dataset:

Age (yrs)	Weight (Kg)	Height(Cm)	BMI
Blood Group	Pulse rate(bpm)	RR (breaths/min)	Hb(g/dl)
Cycle(R/I)	,Cycle length(days)	Marriage Status (Yrs)	Pregnant(Y/N)
No. of abortions	I beta-HCG(mIU/mL)	II beta-HCG(mIU/mL)	FSH(mIU/mL)
LH(mIU/mL)	FSH/LH	Hip(inch)	Waist(inch)
Waist:Hip Ratio	TSH (mIU/L)	AMH(ng/mL)	PRL(ng/mL)
Vit D3 (ng/mL)	PRG(ng/mL)	RBS(mg/dl)	Weight gain(Y/N)
hair growth(Y/N)	Skin darkening (Y/N)	Hair loss(Y/N)	Pimples(Y/N)
Fast food (Y/N)	Reg.Exercise(Y/N)	BP _Systolic (mmHg)	BP _Diastolic (mmHg)
Follicle No. (L)	Follicle No. (R)	Avg. F size (L) (mm)	Avg. F size (R) (mm),Endometrium (mm)

3.2 Model Training and Evaluation

Evaluation metrics are used to measure the performance of a machine learning model. Here are the key ones used in classification problems, explained briefly:

1. **Accuracy**: The proportion of correct predictions out of all predictions. It gives a general idea of how well the model performs.

$$\textbf{Accuracy} = \frac{\textbf{Correct Predictions}}{\textbf{Total Predictions}}$$

2. **Precision**: This tells us how many of the predicted positive cases were actually positive. It's important when the cost of false positives is high (e.g., medical diagnosis).

$$\textbf{Precision} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Positives}}$$

3. **Recall (Sensitivity)**: The proportion of actual positive cases that were correctly identified. It is useful when the cost of false negatives is high.

$$\textbf{Recall} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}}$$

4. **F1-Score**: The harmonic mean of precision and recall. It balances both metrics, especially in cases of imbalanced data.

$$\textbf{F1} - \textbf{Score} = 2 * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1560-1565

e-ISSN : 2583-1062
Impact Factor : 7.001

5. **Area under the Receiver Operating Characteristic Curve (AUC-ROC):** AUC-ROC, or Area under the Receiver Operating Characteristic Curve, is the performance metric of how good a binary classification model is at discriminating between the two classes of interest.

$$\textbf{True Positive Rate (Recall)} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

$$\textbf{False Positive Rate (FPR)} = \frac{\textbf{FP}}{\textbf{FP} + \textbf{TN}}$$

These metrics help assess how well the model distinguishes between classes (e.g., diabetes types or PCOS detection) and are crucial in determining a model's effectiveness.

4. RESULTS AND DISCUSSION

The performance of three machine learning models is compared on two different medical datasets. They are Random Forest Classifier, Logistic Regression, and K-Nearest Neighbors (KNN). The metrics used for their performance assessment include Accuracy, Precision, Recall, and F1-Score..

1. Diabetes Dataset Results:

The Random Forest Classifier performed the best among all models, based on a result accuracy. Logistic Regression consistently performed well with all the evaluation metrics, with special focus on precision, as it shows the potential of strong case detection of diabetes cases without raising high false alarms. Results were somewhat mediocre in terms of accuracy for KNN but poor in terms of precision and recall, possibly because KNN seems sensitive to uncalled features.

2. PCOS Dataset Results:

The best performing model again was Random Forest Classifier, which did extremely well, especially concerning accuracy and recall, as the latter is crucial in the identification of PCOS cases. Logistic Regression proved competitive in accuracy, although it was still outperformed in recall by Random Forest. The KNN was the weakest one with low recall and precision scores. Its results are probably disturbed by the nature of the dataset and lack of feature scaling absolutely.

Diabetes:

	Accuracy	Precision	Recall	F1-Score
Random Forest	0.951201	0.950437	0.876344	0.911888
Logistic Regression	0.946553	0.954955	0.854839	0.902128
KNN	0.786212	0.666667	0.516129	0.581818

Metrics	Random Forest	KNN	Logistic Regression
Accuracy	95.20%	78.62%	94.66%
Precision	94.80%	66.67%	95.50%
Recall	88.17%	51.61%	85.41%
F1-Score	91.36%	58.18%	90.21%

PCOS:

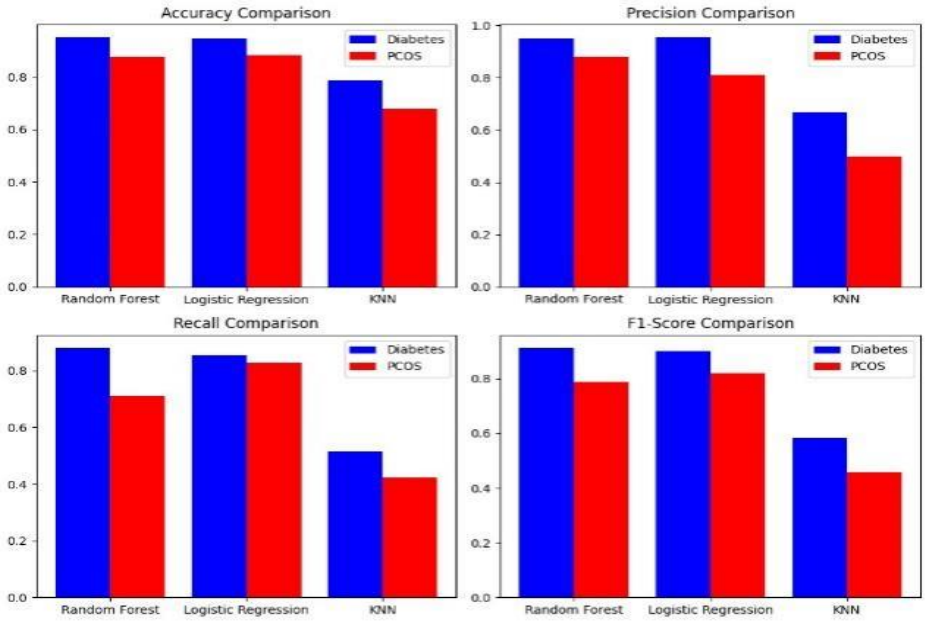
	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.882716	0.883721	0.730769	0.8	0.939948
Logistic Regression	0.882716	0.811321	0.826923	0.819048	0.902622
KNN	0.679012	0.5	0.423077	0.458333	0.661101

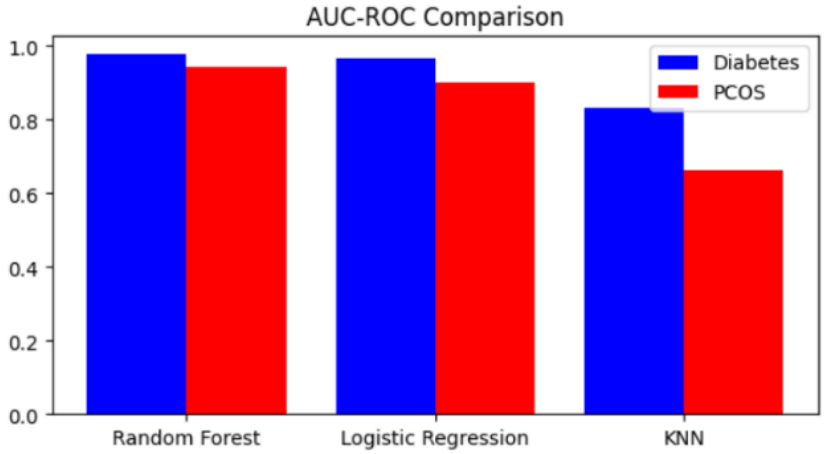
www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1560-1565

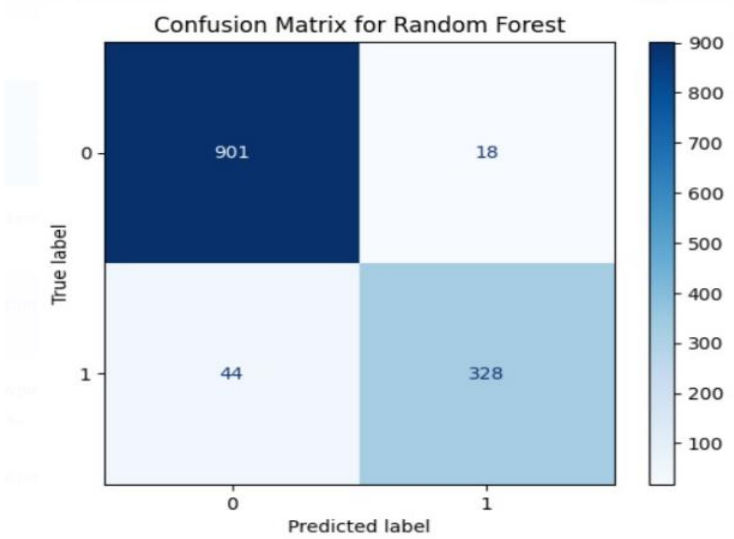
e-ISSN : 2583-1062
Impact Factor : 7.001

Metrics	Random Forest	KNN	Logistic Regression
Accuracy	87.65%	67.90%	88.27%
Precision	88.10%	50.00%	81.13%
Recall	71.15%	42.31%	82.69%
F1-Score	78.72%	81.90%	80.90%





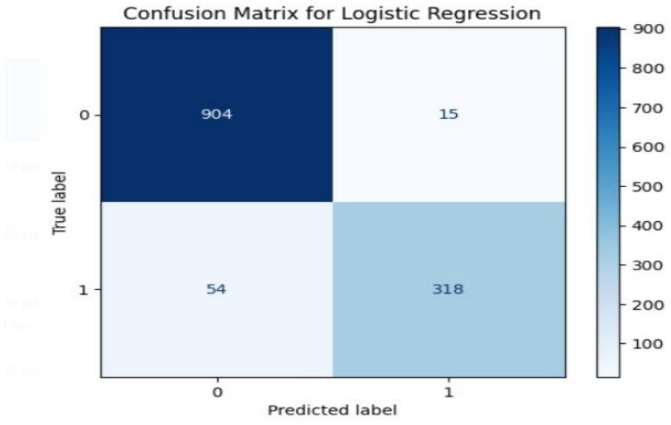
Confusion Matrix:

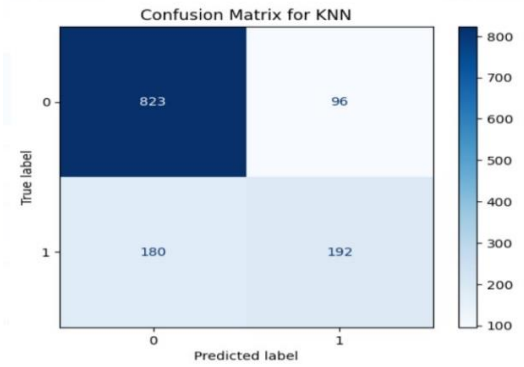


www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1560-1565

e-ISSN : 2583-1062
Impact Factor : 7.001





5. CONCLUSION

By the results of this study, it was found that Random Forest is a best-performing model in predicting diabetes types and detecting PCOS. The next best was K-Nearest Neighbors, followed by Logistic Regression. Random Forest is an ensemble model which makes it robust to over fitting. Moreover, it can learn complex feature interactions. Logistic regression, though interpretation, is limited in terms of linearity. Though KNN is relatively simple, correct feature scaling can bring competitive results.

6. LIMITATIONS

Datasets about PCOS and diabetes are focused upon, which would be a limited size, diversity, and representation with regard to various demographics. The findings and the models may not generalize well to other medical conditions or populations as is specific to that data.

7. FUTURE SCOPE

Further work would include application of deep learning models and more feature engineering to classify better and perhaps early detect the disease. More diverse sets would be able to improve the generalization of the developed models.

8. REFERENCES

[1] Escobar-Morreale H.F. Polycystic ovary syndrome: Definition, etiology, diagnosis and treatment. Nat. Rev. Endocrinol. 2018;14:270–284. doi: 10.1038/nrendo.2018.24. [PubMed]

[2] Norman R.J., Dewailly D., Legro R.S., Hickey T.E. Polycystic ovary syndrome. Lancet. 2007;370:685–697. doi: 10.1016/S0140-6736(07)61345-2. [CrossRef]

[3] McCartney C.R., Marshall J.C. Polycystic ovary syndrome. N. Engl. J. Med. 2016;375:54–64. doi: 10.1056/NEJMcp1514916. [Google Scholar]

[4] Barber T.M., Franks S. Obesity and polycystic ovary syndrome. Clin. Endocrinol. 2021;95:531–541. doi: 10.1111/cen.14421. [PubMed]

[5] Azziz R. Polycystic ovary syndrome. Obstet. Gynecol. 2018;132:321–336. doi: 10.1097/AOG.0000000000002698. [PubMed]

[6] Darabi, P. (2023). Diabetes Dataset with 18 Features. [Kaggle]

[7] Vedpathak, S. (2023). PCOS Dataset. [Kaggle]

[8] National Center for Biotechnology Information (NCBI), 2023. Polycystic Ovary Syndrome. PMC. [NCBI]

[9] ScienceDirect, 2022. PCOS and Diabetes Studies. ScienceDirect. [ScienceDirect]