JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING MEDICAL REPORT

[1]Raksha, [2]Krishna C, [3]Muhammed Mikdad U M, [4]Fatheen Yusuf,
[5]Dr. Jithendra PR Nayak, [6]Shreya Shetty, [7]Suresha D
[1,2,3,4]B.Tech Student , [5]Associate Professor, [6]Assistant Professor, [7]Professor
[1,2,3,4,5,6,7]Computer Science and Engineering,
[1,2,3,4,5,6,7]Srinivas Institute Of Technology Mangaluru India

Abstract:-The system tackles the problem of diabetes risk assessment in individuals by predicting "Diabetes Prediction Using Medical Reports" with the help of their detailed medical records from health institutions and applying logics and rules-based machine learning techniques on. The advanced system aims to analyze key key health metrics and potential risk factors including glucose concentration, body mass index (BMI), age, blood pressure, insulin concentration, family medicalhistory, and several other variables. By analyzing these factors, the system can develop a model and enabling proactive steps at an early stage. The flow of work in a project begins with feature selection highlighting the most influential factors. Then data normalization adjusting the scale of the features to a standard range and removing the missing entries that make the model unusable, known as data cleansing, is done. Logistic Regression works best for problems such as this with a clear delineation, and within a well-balanced dataset. Moreover, the dataset should best represent the entire population of patients in terms of age, sex, and other distinguishing features as this intends to improve prediction accuracy. The overall effectiveness is analyzed with additional advanced measures for model evaluation accuracy, precision, recall, F-score, especially in their collective form give a better image as the provide proper evaluation.

Keywords – Diabetes prediction, Medical Reports, Machine Learning, Feature Selection, Data Normalisation,Data Cleansing, Risk Assessment, Health Indicator

---

I. INTRODUCTION:-

Diabetes is a health condition that affects millions of people around the world as it can give rise to complications such as heart problems, kidney damage, and nerve damage if not diagnosed and treated early. Although existing while current methods for diagnosing diabetes are effective, they tend to be costly and require significant amounts of time, which underscores the urgency of finding faster and more convenient solutions. A powerful alternative lies within machine learning which has the capacity to estimate the risk of diabetes using day-to-day medical information. My goal in this project is to develop a Diabetes Prediction System that utilizes the Logistic Regression algorithm. It incorporates major health indicators including blood glucose, body mass index (BMI), and blood pressure to provide a reliable mechanism of early diagnosis. The system allows healthcare specialists to diagnose patients with increased risk so that intervention can be applied in good time, consequently improving the health outcomes of the patients, and in the long run, reducing diabetes-related complications and health issues.

II.LITERATURE SURVEY :-

1) Sujatha R. and Sharnitha J.

Sujatha and Sharnitha conducted an in-depth study on utilizing machine learning techniques to forecast the development of diabetes.Their study underscored the relevance of medical datasets like the PIMA Indian Diabetes Dataset, as well as data collected from hospitals. They considered multiple classification methods including Logistic Regression, Support Vector Machines (SVM), and Decision Trees.

2) M. A. Jabbar

M. A. Jabbar made notable contributions with his research and survey studies on diabetes prediction using machine learning and data mining techniques. His studies emphasized the processes of feature selection and the comparison of various algorithms in relation to the accuracy of their predictive outcomes.

3) R. Kavitha and K. Duraiswamy

R. Kavitha and K. Duraiswamy researched the use of classification methods and feature selection strategies to improve diabetes

prediction outcomes. Their research involved extensive performance testing with PIMA dataset and real-life medical data.

4) R. Subashini and Dr. V. Sundararajan

R. Subashini and Dr. V. Sundararajan performed a comparison of different machine learning techniques applied to diabetes prediction .Their study compared traditional techniques, like basic logistic regression, against more advanced models developed through Deep Learning.

III METHEDOLOGY:-

1)Glucose position

Tube glucose attention measured two hours after an verbal glucose resistance test.

2)BMI(Body Mass List)

An pointer of body fat calculated grounded on an existent's highness and weight.

3)Age : The case's age in a long time.

4)Blood Weight :Diastolic blood weight readings.

5) poke :The serum poke attention within the blood.

6)Skin Consistence:Estimation of triceps skinfold consistence.

7) Diabetes Family Work:A regard demonstrating the heritable inclination to diabetes grounded on family remedial history.

2)Software Tools and Libraries

The following programming languages, libraries, and tools were employed for developing and executing the machine knowledge model

Python Chosen as the main programming language because of its strictness and rich ecosystem of libraries that support data wisdom and machine knowledge tasks.

Pandas A Python library essential for managing and manipulating datasets, including loading, cleaning, and preparing data for analysis.

NumPy Used for effective numerical operations, particularly for handling arrays and performing fine computations during data preprocessing and model evaluation.

Scikit- learn A considerably- used machine knowledge library that eased model structure and evaluation. It offers various erected-in algorithms, analogous as Logistic Regression, and functions for tasks like dataset splitting, point scaling, and performance evaluation.

Matplotlib and Seaborn Visualization libraries employed to produce graphs and charts, helping to explore point connections, distributions, and performance criteria like ROC angles and confusion matrices.

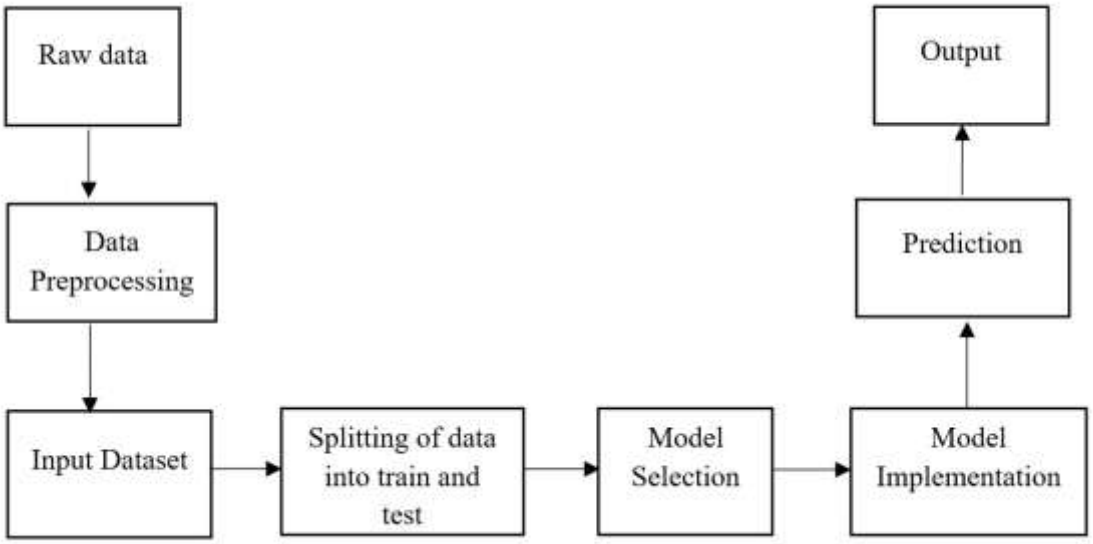
3)Machine Learning Algorithms:

Logistic Regression:

Logistic Regression served as the main algorithm for this project, ideal for tackling binary classification problems. It predicts the probability of diabetes occurrence based on input features by applying a logistic (sigmoid) function. The output values range between 0 and 1 and are subsequently categorized into twodistinct groups — diabetic or non-diabetic — based on a determined.

Data Splitting:

In order to assess the performance of a model accurately, the dataset was split into two parts, the train and the test subsets, with the common splitting being to give 80% of the data to training and 20% to testing. The split was thus performed using the train-test split function available in Scikit-learn.



IV. ARCHITECTURE OF THE DIABETES PREDICTION USING MEDICAL REPORTS

1)Data Collection Stage
Healthcare information is retrieved through PIMA Indian Diabetes Dataset and hospital databases.
The essential features include Blood Glucose, BMI, Age, Blood Pressure, Insulin Level, Skin Thickness and the Diabetes Pedigree Function and several others.

2) Data Preprocessing Stage
The identification and treatment of missing and inconsistent data points should include both value reduction when entries fail quality checks as well as value replacement strategies.A normalization process couples with scaling features to establish common value ranges through the utilization of MinMax Scaler and StandarScaler Tools from Scikit-learn. Selection of influential features should follow a process to predict diabetes properly.

3) Data Splitting Stage
The analyzed dataset gets split according to a 80/20 ratio This allocation is earmarked for training and testing purposes. The actual data set constituted 80% of Training Set to model the educative purpose, and 20% was kept as Testing Set for assessing the performance
Evidence now shows that twenty percent of the data has been withheld for the test phase towards the measurement of model evaluation.Having a testing and training subset even helped in diminishing overfitting; obviously there is no bias available for model evaluation.

4) Model Development Stage
For binary classification purposes the main algorithm will be Logistic Regression.
The model receives its training through the implementation of features extracted from the training dataset.
Additionally we use hyperparameter Optimization (optional) as a parameter enhancement technique for model performance enhancement.

5)Prediction Stage

The Logistic Regression model shows the ability to determine diabetic or non-diabetic status after processing new dataset information

V. FUTURE WORKS OF THE SYSTEM

The existing logistic regression approach yields satisfactory . The combination of biological indicators with clinical variables enables medical researchers to project treatment results from the data**.**

Critical data can be analyzed through existing methods but several improvements can be implemented.

1)Incorporation of Advanced Machine Learning Models:

Future versions should incorporate models like Random Forest, XGBoost, Support Vector Machines (SVM), and Neural Network-based approaches to enhance performance and handle intricate data relationships.

2)Integration of advanced deep learning methodologies:

The classification tasks use various deep learning models composed of Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs).serve as alternative procedures for to derive meaningful insights from intricate data patternswithin medical datasets (especially when working with image-based reports).

3)Real-Time Prediction System:

A web-based and mobile application targeting diabetes prediction provides instant feedback to users or doctors by allowing them to enter health data would enhance system accessibility and effectiveness.

4)Expansion of Dataset:

A consolidated analysis of broad and diversified health data from multiple medical facilities will enable The approach enables programmers to build predictive systems suitable for numerous population types spread across multiple geographical 13 regions.

5)Personalized Risk Assessment:Subsequent versions should use individual risk factors such as lifestyle behavior and genetic background and environmental exposures to give specific health.

VI. IMPACT ON SOCIETY

Early Diagnosis and Prevention:Early prediction of diabetes risk allows people to take preventive measures such as lifestyle modification along with dietary changes and regular check-ups which help to lower the frequency of severe health problems.

VII. RESULT OF THE SYSTEM

```
x=data.drop('Outcome',axis=1)
y=data['Outcome']

rm = RandomOverSampler(random_state=41)
x_res,y_res = rm.fit_resample(x,y)

print('old data set shape{}'.format(Counter(y)))
print('old data set shape{}'.format(Counter(y_res)))
```

```
old data set shapeCounter({0: 500, 1: 268})
old data set shapeCounter({1: 500, 0: 500})
```

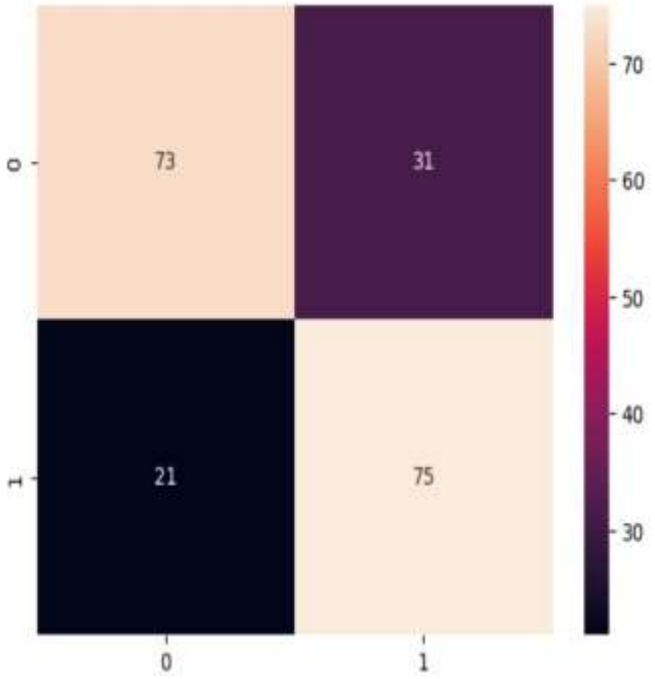


Figure.1: Confusion Matrix

```
sns.countplot(x='Outcome',data=data,palette=['green','red'])
```

```
<Axes: xlabel='Outcome', ylabel='count'>
```

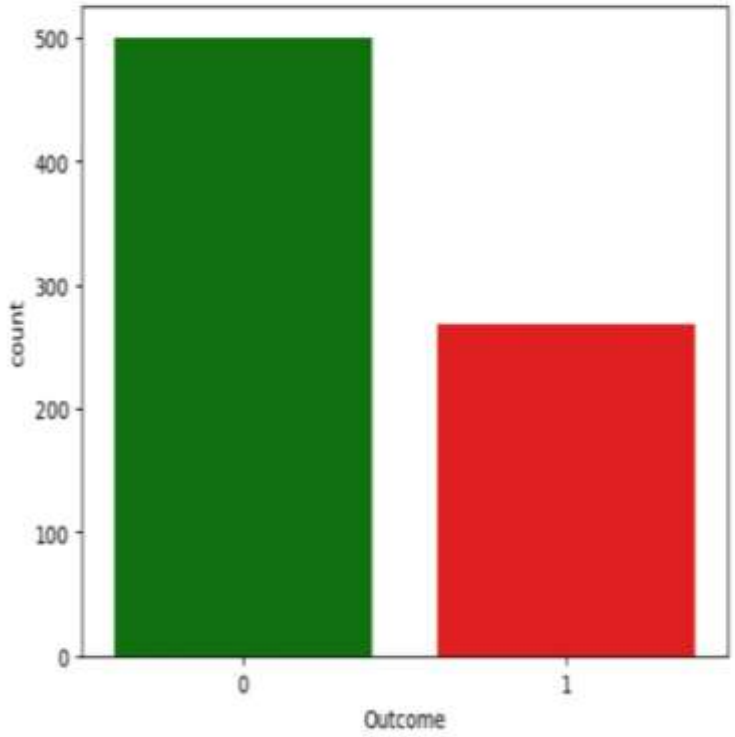


Figure 2: Balancing the imbalanced data using Random Over-Sampling

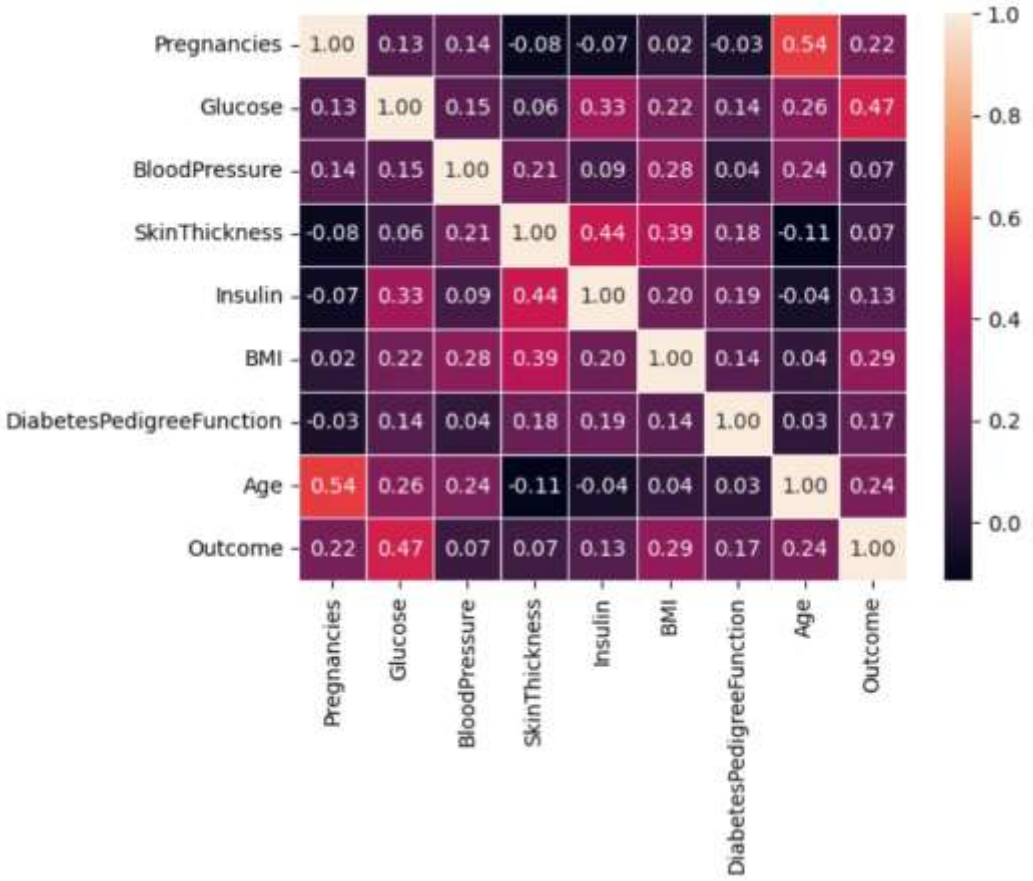


Figure 3: Heatmap of Correlation

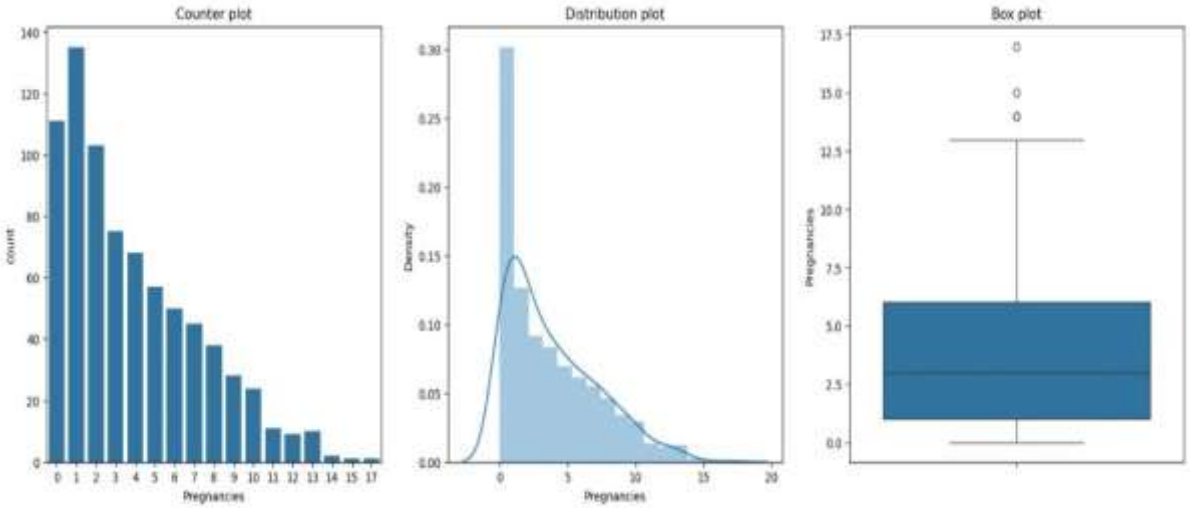


Figure 4. Counter Plot, Distribution Plot and Box Plot

```
LogisticRegression()

Accuracy is:  0.74
Recall is:  0.78125
f1 Score is:  0.7425742574257426
```

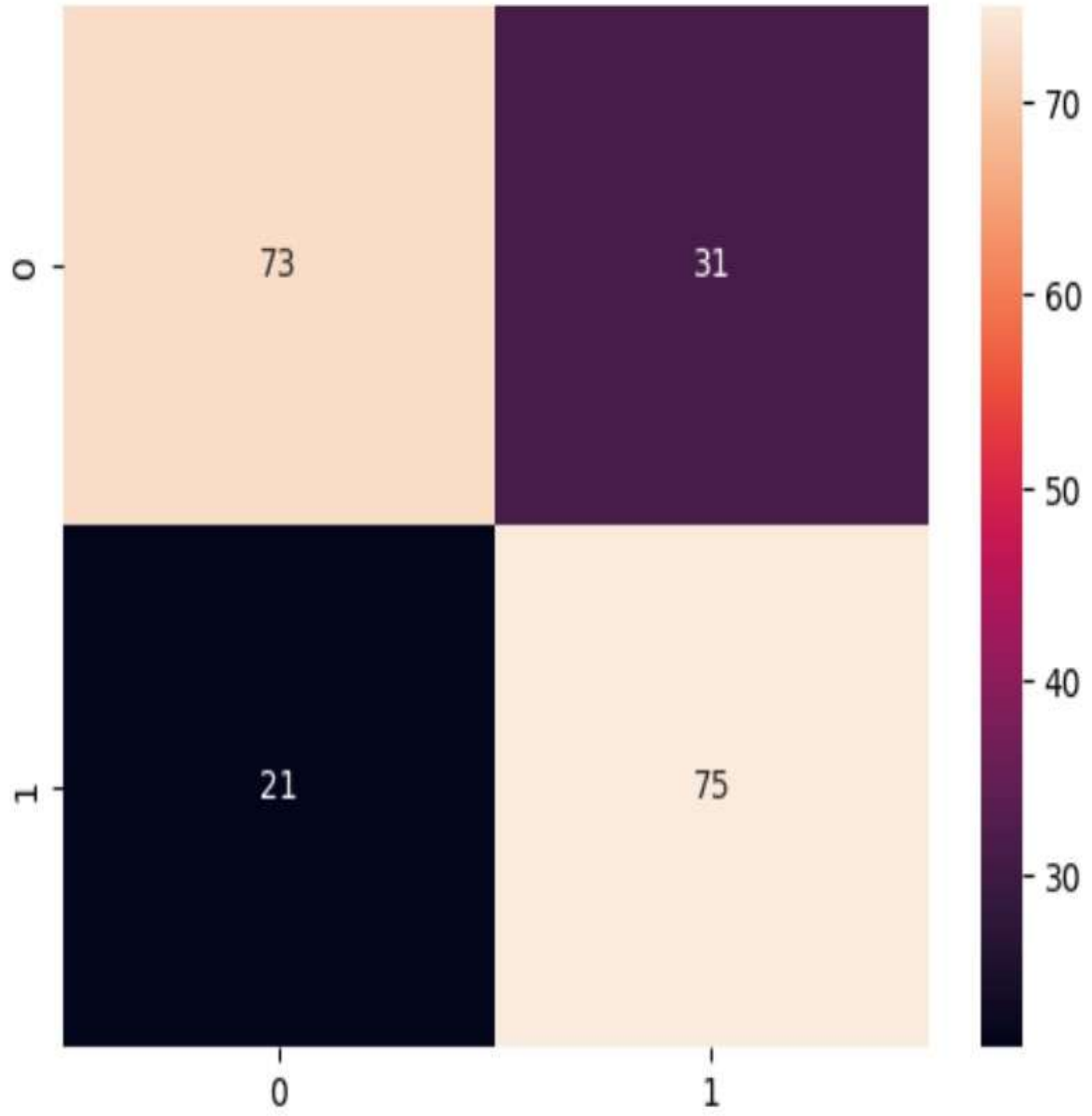


Figure 5 Confusion Matrix of the Mode

VIII. CONCLUSION

A successful machine learning implementation of the Logistic Regression algorithm defines the likelihood of diabetes occurrence in patients from medical features like glucose levels and BMI and age and blood pressure and insulin levels. A strategic preprocessing process was applied to address missing data values and the Random Over-Sampling method was implemented to handle class imbalance. Results from the Logistic Regression model proved promising by yielding high accuracy alongside high recall and F-score thus proving its effectiveness for diabetic diagnosis detection.The results from the Logistic Regression model indicated promising predictions of diabetes while showing excellent accuracy metrics and recall ability as well as F-score characteristics. The Random Over-Sampling data balancing method with relevant feature selection resulted in reliable performance of the model.

IX. ACKNOWLEDGMENT: -

My deepest appreciation goes to everyone who assisted me during the execution of "Diabetes Prediction Using Medical Report."The first point of appreciation goes to Your Institution/Organization Name because they supported my work on this project with both direction and research materials. The project guidance provided by [Guide's Name] stands as the most valuable asset I received because they consistently offered positive support through their wise guidance and motivational direction during every phase of this work.

The study's foundation derives from the medical experts together with data providers who provided both medical datasets and vital knowledge. Through their provided data scientists could successfully create and evaluate predictive models.

REFERENCES

[1] Alva, S., & Malini, P. (2019). A Survey on Diabetes Prediction through Machine Learning Algorithm Applications appears in this paper. International Journal of Advanced Research in Computer Science and Software Engineering, 9(4), 1-7.

[2] Jeberson, A., & Srinivas, K. (2020). "Diabetes Prediction using Logistic Regression". International Journal of Computer Applications, 175(4), 10-14.

[3] Chaurasia, V., & Pal, S. (2019). The research analyzes "Prediction of Diabetes Disease using Machine Learning Algorithms". Procedia computer science, 132, 975-982.

[4] Sathish, S. A., & Kannan, A. (2018). The research paper examines "Diabetes Prediction System using Machine Learning". Proceedings of the International Conference on Computational Intelligence and Data Engineering.

[5] Sahu, R. K., & Gupta, S. (2018). The research paper focuses on detecting diabetes through data mining methods. International Journal of Scientific & Technology Research, 7(9), 82-86

[6] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988).(pp. 261)

[7]]Kavakiotis, I., Tsave, O., Salifoglu, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017)15104–116.

[8] SisodiaD&Sisodia,D.S.(2018)132,1578–1585,132,1578–1585.https://doi.org/10.1016/j.procs.2018.05.122

[9] Smith, J., & Brown, L. (2018),, 2(3), 210–220.https://doi.org/10.1007/s41666-018-0020-x

[10] Kavakiotis, I., Tsave, O., Salifoglu, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017)15, 104–116.https://doi.org/10.1016/j.csbj.2016.12.005

[11] Choubey, D. S., & Paul, S. (2020); 9(3), 2277-8616.

[12] Nalluri,S.,&Babu,M.S P. (2021);37(Part 2), 2704–2709.https://doi.org/10.1016/j.matpr.2020.09.707

[13] Farhan, L. A., Reda, H. T. & T. Almustafa, K. M. "Utilizing Social Network for Tutoring." Int. J. Advances in Computing and Software Sys. 11(5), 672-678. 2020, https://doi.org/10.14569/IJACSA.2020.0110585.

[14] Patil, B. M., Joshi, R. C. & Toshniwal, D. "An Analysis of PCA and K-Means Clustering for Dimensionality Reduction and Clustering." Expert Systems with Applications 37, 12(2010), 8102-8108. 2010. https://doi.org/10.1016/j.eswa.2010.05.071.

[15] Kumari, V. A., & Chitra, R. "GPU-Based Implementation of C4.5 Algorithm." International Journal of Computer Applications 3.2 (2013): 1797-1801.

[16] Alaa, A. M., & van der Schaar, M. "Adversarially Learning Fair Representation." Proceedings of the 36th International Conference on Machine Learning 96 (2019): 2348-2355. https://proceedings.mlr.press/v96/alaa19a.html.

[17] Pérez, L., & Wang, J. "Learning from Uncertainty-Aware Agents." arXiv 1712.04621 (2017).

[18] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. "User-Centric Recommender Systems: A Survey." Health Information Science and Systems 19.1 (2019): 1-16. https://doi.org/10.1186/s12911-019-1004-8.

[19] Martin, A., Sturner A.E., Kaboose, T., Hagn, U., Queitsch-Maitland, M. & Mossköck, H. (2015).iomanip.https://doi.org/10.1007/978-3-642-49383-8_4

[20] Sahu, R. K., & Gupta, S. (2018). The research paper focuses on detecting diabetes through data mining methods. International Journal of Scientific & Technology Research, 7(9), 82-86