Comparative Analysis of Permanent Neonatal Diabetes Mellitus Prediction using various Machine Learning Techniques

Madhura Devendra Ranade
Assistant Professor
Department of Computer Science,
SVKM's Mithibai College of Arts, Chauhan Institute of Science & Amrutben
Jivanlal College of Commerce and Economics, Mumbai, India

ABSTRACT

This paper aims at analyzing the machine learning techniques for prediction of Permanent Neonatal Diabetes Mellitus (PNDM). PNDM is a serious condition responsible for neonatal mortality. This can be caused due to insufficient insulin levels.

The newborn baby in its month after birth is termed as "neonate". One of the Sustainable development goals provided by world health organization is to reduce neonatal mortality. Machine learning approach for disease prediction is a noninvasive method of dealing with the available data and its interpretation. The main objective of the paper is to assess the best suitable machine learning method for the prediction of PNDM. The dataset used in the analysis consists of six input features as age, information related to genetics, HbA1c level, medical attributes, laboratory details and maternal history. The output feature or target variable in this dataset is the presence or absence of PNDM. [1]. The dataset is divided in training and testing modules in the ratio of 8:2.

The dataset is validated and tested using various machine learning techniques such as Decision Tree classifier, Support Vector Machines, Logistic Regression, Naïve Bayes and Ensemble classifier. It was observed that, tree classifier was the best suitable choice for prediction of PNDM as it provided 99% accuracy and lowest training and validation cost as compared to other methods.

General Terms

Machine Learning Disease prediction.

Keywords

Permanent Neonatal Diabetes Mellitus (PNDM) , machine learning, neonatal healthcare, decision tree classifier, SVM classifier.

1. INTRODUCTION

The advances in artificial intelligence are transforming the world at very high pace. This technology has shown its wonders in the healthcare industry also. What if one applies it for prediction of neonatal health conditions? This paper tries to answer the previously asked question. Neonates are the newborn babies smaller than four weeks of age. These babies are so tiny that it is very difficult for healthcare professionals to perform invasive procedures on them. Some newborn have serious health conditions that they have to be treated to increase their lifespan. This paper shows an innovative method of using machine learning techniques to train the model for prediction of Permanent Neonatal Diabetes Mellitus. The dataset is an open source and freely available on internet. [1].

The paper is organized as various sections. First section provides introduction to the terminology used in paper. In the second section, the detailed information of PNDM condition is provided along with its symptoms and current diagnostics. The third section talks about various machine learning techniques in certain depth. The fourth section explains the proposed methodology used for prediction of PNDM. Section 5 covers the results obtained and its interpretation. In section 6, paper is concluded.

2. PERMANENT NEONATAL DIABETES MELLITUS (PNDM)

2.1 What is PNDM?

PNDM is a health condition which occurs in the newborn child less than half a year old, due to Hyperglycemia. It persists throughout their lifespan. It can be the effect of underlying insulin deficiency. Insulin is a hormone present in the blood which utilizes sugar in the body for energy. If the generation of Insulin is not as per requirement, then that condition is known as diabetes. 1 in around four lakhs babies is diagnosed with this condition. Some of the babies have transient symptoms meaning the condition is cured on its own within one or two years. The remaining children have to deal with PNDM throughout their life [2].

2.2 Causes of PNDM

This condition may be caused due to several gene mutations. It has been researched that the mutations in KCJN11, ABC88 gene is highly associated with PNDM.

2.3 Testing of PNDM

PNDM can be diagnosed with the help of measurement of plasma glucose concentration. If the plasma glucose level is greater than 150-200mg/dLin the children up to six-month age, then the underlying condition is diagnosed as PNDM.

The molecular testing is also done for checking genetic profiles. The pancreatic testing is also done using radiography.

2.4 Treatment for PNDM

The neonates diagnosed with this condition are advised to start Insulin therapy. These children have to be monitored on regular basis for blood glucose to avoid further complications due to diabetes.

These children require developmental and retinopathic evaluations periodically. Genetic counselling is also advised to the parents of such children.

3. MACHINE LEARNING TECHNIQUES

Machine learning is a branch of artificial intelligence which deals with training a machine using available data to learn to classify patterns.

There are various machine learning techniques broadly divided as supervised or unsupervised learning techniques.

The supervised techniques use labeled data for training whereas unsupervised methods do not use any labeling. Every method has its own pros and cons and still at a developing stage. It is very essential to apply these methods for finding out if they are really good at their classification job. In this paper, decision tree classifier and SVM, neural networks are used for training.

3.1 Decision Tree Classifiers

This technique comes under supervised learning. The structure of this classifier is equivalent to tree where every feature of the data is noted as a NODE and every branch is noted as decision rule.

The leaf node of this tree structure is noted as internal outcome. This is a graphical way of plotting all the possible combinations and then selecting a path based on certain decision rules.

3.2 Support Vector Machine Classifier

This technique is quite popular among the ML researchers due to its simplicity. This technique creates a decision boundary in the data space to divide data points into various classes. The best line for classification is called as hyper plane.

3.3 Naïve Bayes Algorithm

This supervised classification method is based on Bayes' Theorem. This algorithm performs classification based on probability.

The algorithm first calculates the frequency of each feature. It then determines the probability based on likelihood table. Then, the final posterior probability is found using Bayes' theorem.

3.4 Neural Networks

This is a type of unsupervised learning technique where classifier learns on its own without any guidance. Neural networks structure contains various layers, nodes for training the dataset.

It is one of the most popular process of ML. This network models the human brain for decision making and finding patterns.

4. METHODOLOGY

The proposed methodology describes the process of predicting PNDM based on machine learning. This process uses the open source dataset for training. The methodology is explained in figure below..

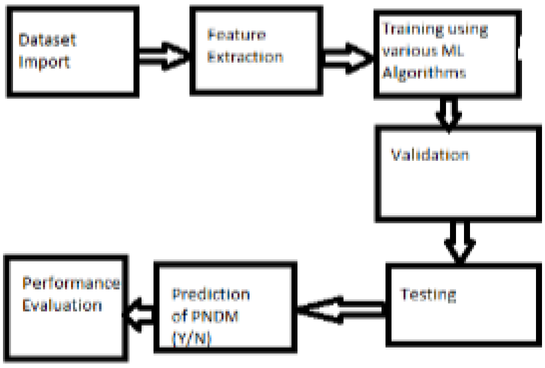

Fig 1: Proposed Methodology for prediction of PNDM

4.1 Dataset Import

The dataset used in this paper is available as an open source dataset [1]. It is a simulated database comprising of 6 features.

The dataset contains samples of around 1,00,000 children. The following Table 1 shows some of the rows of dataset.

Table 1. SAMPLE ROWS FROM DATASET

Age	HbA1c	Genetic Info	Family History	Birth Weight	Develop mental Delay	Insulin Level	PNDM
3	4.840	Mutation	Yes	3.128	No	5.585	0
3	5.694	Mutation	No	2.059	No	3.141	1
7	6.843	No mutation	No	2.718	Yes	4.6393	0
2	6.480	No mutation	No	3.087	No	6.217	0
4	7.0	Mutation	No	3.4	No	3.3688	0

The detailed information of features is explained in following points.

1) Age:

This feature gives the exact age of the child at the time of diagnosis. This dataset includes samples between age 1 to 11

2) HbA1c levels:

This feature represents the level of HbA1c serum. The level range varies between 2.83 to 11.3.

3) Genetic Info:

This feature provides the information about genetic mutations. (Mutations/No Mutations).

4) Family History:

This feature is stored as a categorical value "Yes" if the child has family history of PNDM and "No" if the child does not

have family history of PNDM.

5) Birth Weight:

This feature stores the weight of the child at the time of birth. The range in this dataset varies between 0.43 Kg to 4.91Kg.

6) Developmental Delay:

This feature is a categorical variable stored as "Yes" if the child has developmental delay and "No" otherwise.

7) Insulin Level:

This feature provides the measured insulin value. The insulin levels vary between -3.24 to 13.3 range in this dataset.

8) PNDM:

The last column of the dataset is a target variable which stores "0" if the child does not have PNDM and "1" if the child is diagnosed for PNDM.

4.2 Feature Extraction

This step is a preprocessing step where the correlation of features with target variable is studied and extraction is done. For this training, all features are included for training..

4.3 Training using various ML algorithms

Once the dataset is preprocessed, it is divided in training and testing sets in the ratio of (80:20). Then the training dataset undergoes training using various algorithms such as decision tree classifier, Naïve Bayes classifier, SVM, Neural Networks etc.

This is done by using MATLAB classification learner app. After training, the validation and testing is performed using same app in

MATLAB. The software then provides the various evaluation parameters such as accuracy, confusion matrix etc. These results are then tabulated for analysis.

5. RESULTS AND DISCUSSION

The trained dataset is analysed for finding the best suited algorithm for prediction of PNDM. The details of training and testing sessions are as follows:

Training Data: PNDM

• Observations: 80000

• Predictors: 7

• Predictor Names: Age, HbA1c, Genetic Info, Family History, Birth Weight, Developmental Delay, Insulin Level

• Response Name: PNDM

• Response Classes: 2

• Response Class Names: 0, 1

• Validation: 5-fold cross-validation Test Data: PNDM

• Observations: 20000

The obtained results are displayed in Table-2.

Table 2 Results of PNDM Prediction

ModelType	**Model Evaluation Parameters**			
	Accuracy % (Validation)	***Total Cost (Validation)***	***Accuracy % (Test)***	***Total Cost (Test)***
Fine Tree	**100**	**12**	**100**	**1**
BinaryGLM Logistic Regression	98.1987	1441	98.1	Not Applicable
Efficient Logistic Regression	98.1912	1447	98.225	355
Naive Bayes	98.5237	1169	98.565	298
Linear SVM	98.205	1419	98.1	380
EfficientLogistic Regression	98.2	1429	98.1	380
SVM kernel	99.3	597	99.2	158
Ensemble	95.1787	3857	95.175	965
Neural Network	**100**	**28**	**100**	**5**

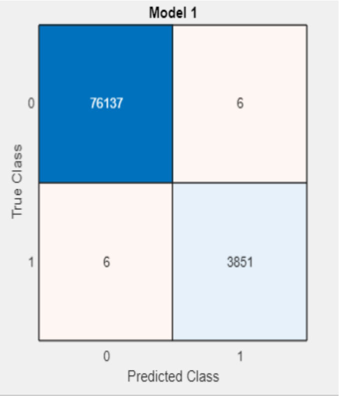

Fig 2: Validation confusion matrix for fine tree classifier

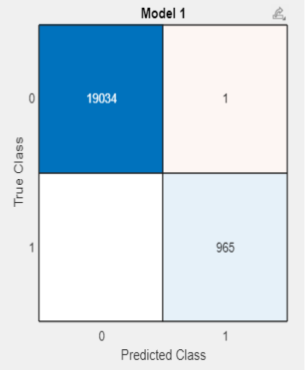

Fig 3: Test confusion matrix for fine tree classifier

It can be seen from the confusion matrix that very few samples are misclassified. Hence the accuracy is highest for fine tree

classifier. For above results PCA(Principal component analysis option was disabled. The next analysis was done on fine tree

classifier and Neural networks thereby enabling PCA option with 95% variance and 6 features.

Table 3 RESULTS OF PNDM PREDICTION WITH PCA

Model Type	**Model Evaluation Parameters**			
	Accuracy % (Validation)	***Total Cost (Validation)***	***Accuracy % (Test)***	***Total Cost (Test)***
Fine Tree	97.4	2084	97.3	539
Narrow Neural Network	97.5	2031	97.5	502

It can be seen from the table that both the models have reduced accuracy after enabling PCA. It may be the result of reducing number of features for training.

The another test was performed by setting optimizer option to Bayesian optimization. PCA is still kept enabled.

Table 4 Results of PNDM Prediction Wit PCA and Bayesian Optimizer

Model Type	**Model Evaluation Parameters**			
	Accuracy % (Validation)	***Total Cost (Validation)***	***Accuracy % (Test)***	***Total Cost (Test)***
Optimizable Tree	97.5	2018	97.5	500

Thus, the training and testing dataset is been performed on the dataset with various training classifiers including supervised and unsupervised training methods

6. CONCLUSION

This paper has proposed a methodology for training the dataset for prediction of Permanent Neonatal Diabetes Mellitus using supervised and unsupervised learning techniques. The dataset used in the study was taken from open source simulated data. The classification learner App from MATLAB software used for training, testing and validation. The results obtained show that PNDM can be successfully predicted using ML techniques with around 99% accuracy. Neural Networks and Fine tree classifier were best fit for prediction of PNDM.

The future scope of this paper is to apply more ML techniques to this dataset. The new dataset with added features can also be explored to get better results.

7. ACKNOWLEDGMENT

Author thanks the institute for availing resources to carry out this research.

8. REFERENCES

[1] https://www.kaggle.com/datasets/slmsshk/pndm-prediction-dataset

[2] Barbarini DS, Haslinger V, Schmidt K, Patch AM, Muller G, Simma B. Neonatal diabetes mellitus due to pancreas agenesis: a new case report and review of the literature. Pediatr Diabetes. 2009 Nov;10(7):487-91. doi: 10.1111/j.1399-5448.2009.00523.x. Epub 2009 Jun 3.

[3] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2022 Dec 14;10(1-2):1-10. doi: 10.1049/htl2.12039. PMID: 37077883; PMCID: PMC10107388..

[4] Zhou, H., Myrzashova, R. & Zheng, R. Diabetes prediction model based on an enhanced deep neural network. J Wireless Com Network 2020, 148 (2020). https://doi.org/10.1186/s13638-020-01765-7.Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.