Approaches to Type -2 Diabetes Mellitus Prediction with Machine Learning and Deep Learning

Satyam Chandrakant Chatrola, Prof. Hetal Gaudani

Gujarat Technological University, Chandkheda, Ahmedabad, Gujarat, India

Abstract:

Diabetes Mellitus is a severe health risk with very low awareness about its early symptoms. Approximately half of the people don't know that they suffer from Diabetes and they develop chronic diabetes when diagnosed. Therefore, its paramount to identify people with Diabetes, so they could be treated early and the severity of the disease can be mitigated. Machine Learning (ML) and Deep Learning (DL) techniques help us to classify people with diabetes and pre-diabetes. This is possible due to the health indicator dataset which contains answers of a health survey. The tree-based ML algorithms (Decision Tree, Random Forest, XGBoost) and Artificial Neural Network (ANN) perform well on the data. Algorithms that display high precision, high specificity and high recall are more suitable to medical problems. And the former mentioned algorithms yield scores greater than 93% in the evaluation tests.

1 Introduction

1.1 Diabetes Mellitus

Diabetes Mellitus, commonly known as Diabetes, is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. Beta cells of Pancreas secretes a hormone named Insulin which controls blood glucose level. There are 2 reasons for the elevated glucose levels found in the blood:

1. Decreased amount of Insulin secretion by the Pancreas.
2. Insensitivity of the receptors of Insulin or resistance to absorption by cells.

There are 2 mainly two types of diabetes:

1. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself.
2. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin.

According to WHO, about 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades.

Diabetes also brings health complications like Heart Attack, Stroke, Retinal Detachment, Chronic Kidney Failure, etc. Approximately 50% of the people don't know that they are suffering from diabetes. This makes early detection of diabetes paramount.

1.2 Machine Learning and its types:

Machine Learning is a branch of Artificial Intelligence and Computer Science with leverages the power of Mathematics to identify complex patterns in the data and imitates human-like intelligence. Machine Learning can help identify complex pattern from people's habits and health data and can infer whether a person has diabetes or not. There are 3 types of Machine Learning Techniques:

1. Supervised Learning:
   Supervised Learning is a subcategory of machine learning and artificial intelligence. It uses labelled datasets to train algorithms to classify data or predict outcomes accurately. The labelled dataset. Supervised learning help deal with two types of problems:

   a. Classification:
      It uses an algorithm to distinguish data into already specified number of categories. The model is already trained on the labelled dataset which consists of data containing all the categories that the model needs to classify. Then, the model is able to categorize similar data into known categories. For e.g., classifying emails into spam and not spam, categorizing people into groups like with diabetes, without diabetes, etc. Algorithms like Logistic Regression, Naïve bayes, Decision Trees, etc. are used for classification tasks.

   b. Regression:
      It uses an algorithm to determine relationship between the data and the label and predict the outcome of the data. In Regression, the output (label) is not bounded i.e., the label can take any value unlike the classification task

where the label is only one of the categories. For e.g., Predicting house prices, Loan Prediction, etc. Algorithms like Linear Regression, Lasso Regression, Support Vector Machine (SVM), etc. are used for regression tasks.

2. Unsupervised Learning:
   Unsupervised learning techniques works on a dataset with unlabelled data, analyses it and uncover hidden patterns from the data and creates clusters of similar data points. Its ability to subtly distinguish the data points on basis of similarity and differences without the need of human intervention makes it ideal for exploratory analysis, customer segmentation, image recognition, recommendation systems, etc. There are 3 main types of tasks that Unsupervised Learning helps to achieve.

   a. Clustering:
      It is a data mining technique which groups unlabelled data based on the hidden patterns that the data exhibit which is based on the similarities and differences of the data points. Algorithms like K-means Clustering, Hierarchical clustering and Probabilistic clustering are used for Unsupervised Learning.

   b. Association Rules:
      An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods group like data points with the insight of which data points affect other points and whether the relationship is vice-versa or not. Thus, this technique is crucial to Market Basket Analysis. For e.g., Market Basket Analysis could give an insight that whenever bread is purchased, eggs are also included in the order; but vice-versa is not true. This enables the retailers to sell breads and eggs side by side to increase the sales of eggs. Algorithms like Apriori, Eclat and FP-Growth are widely used in the industry.

   c. Dimensionality Reduction:
      As the name suggests, Dimensionality Reduction techniques reduces the dimension (size) of the data. Whenever the data is huge in amount, Dimensionality Reduction Techniques like Principal Component Analysis helps in extracting important features from the data and captures the significant amount of data in a relatively small dimension of the new data. This characteristic is particularly useful when the low computational power is available to train the data. Other algorithms like Singular Value Decomposition (SVD) are popular for compressing images while retaining most data and features.

3. Reinforcement Learning:
   Reinforcement Learning is a type of machine learning technique that enables an AI agent to learn in an interactive environment by trial and error by using feedback on its own actions. The model is rewarded for positive behaviour (correct actions) and punished for negative actions (incorrect actions). This approach is analogical to the way in which parents typically raise a child. The child is rewarded when it performs correct actions e.g., praising when a toddler walks correctly. This type of learning is useful in game playing and robotics.

1.3 Deep Learning:

Deep Learning is a machine learning technique that teaches computers to do what comes naturally to humans. It is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is essentially a neural network with more than 3 layers. This makes deep learning adapt to complex and non-linear data. It helps AI mimic the cognitive abilities or just a part of it accurately. In some cases, for e.g., driverless cars, the results are better than human accuracy.

2 Proposed System

2.1 Dataset:

The dataset is a health-related telephone survey (2015) that is collected annually by the Centre for Disease Control and Prevention (CDC) in the United States. It contains information on various health indicators. It is a subset of 22 columns of the original data of “Behavioral Risk Factor Surveillance System” of CDC.

Feature Name	Feature Values	Feature Value Meaning
Diabetes_012	0	No Diabetes
	1	Pre-Diabetes
	2	Diabetes
BMI	12 to 98	Body Mass Index
GenHlth	1	Excellent
	2	Very Good
	3	Good
	4	Fair
	5	Poor
Sex	0	Female
	1	Male

Feature Name	Feature Values	Feature Meaning
MentHlth	0 to 30	Number of Days Mental Health Not Good
PhysHlth	0 to 30	Number of Days Physical Health Not Good
Age	1 to 13	13-level age category (according to AGEG5YR)
Education	1 to 6	Education level (according to EDUCA)
Income	1 to 8	Income Scale (according to INCOME2)

Feature Name	Feature Value Meaning
HighBP	High Blood Pressure
HighChol	High Cholestrol
CholCheck	Cholestrol check in 5 years
Smoker	Have smoked at least 100 cigarettes (5 packets) in your entire life
Stroke	Had a stroke
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI)
PhysActivity	Had done some physical activity in past 30 days
Fruits	Consume fruits 1 or more times a day
Veggies	Consume Vegetables 1 or more times per day
HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
DiffWalk	Do you have serious difficulty walking or climbing stairs?

2.2 Algorithms used for the study:

2.2.1 Logistic Regression:

Logistic regression is a type of linear regression technique where its output is bounded between 0 and 1 by the logistic function. For the multiclass classification, we use 'One vs Rest' technique to apply Logistic Regression to the data. The logistic function is as follows:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$f(x)$ = output of the function

L = the curve's maximum value

k = logistic growth rate or steepness of the curve

x_0 = the x value of the sigmoid midpoint

x = real number

2.2.2 Naïve Bayes:

Naïve Bayes algorithm is a probabilistic classifier. It relies on Bayes Theorem to predict the outcomes based on chances. It makes an assumption that all the features have equal importance.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

2.2.3 K – Nearest Neighbour

K-Nearest Neighbour (KNN) is a supervised learning classifier which uses proximity to make classifications about the grouping of an individual data point. It assumes that similar points can be found near one another. It can use various metrics to calculate distances like Manhattan distance or Euclidean distance. In the end, it takes the vote of the majority of the outcomes of the most similar data points. In our experiment, we have used the outcomes of 5 most similar neighbours.

2.2.4 Decision Tree

A decision tree builds a model in the form of a tree structure. Its grouping exactness is focused with different strategies like entropy and Gini index. We have used Gini index for grouping. The tree splits the data based on the features that results in the cleanest groups.

2.2.5 Random Forest

Random Forest is a type of Ensemble method; it is made up of multiple decision trees and hence making it a 'forest'. Random Forests minimize the drawbacks of decision trees like bias and overfitting. Random Forest is based on one of the ensemble techniques; Bagging. In this method, a random sample of data in a training set is selected with replacement i.e. that the individual data points can be chosen more than once. After several data samples are generated, these models are then trained independently and majority of those predictions yield a more accurate estimate. This approach is commonly used to reduce variance within a noisy dataset.

2.2.6 XGBoost:

XGBoost stands for Extreme Gradient Boosting. It is a decision tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It uses an optimized Gradient Boosting Algorithm through parallel processing, tree-pruning, handling missing data and regularization to avoid overfitting / bias. It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting. XGBoost naturally admits sparse features for inputs by automatically 'learning' best missing value depending on training loss and handles different types of sparsity patterns in the data more efficiently.

2.2.7 Artificial Neural Network:

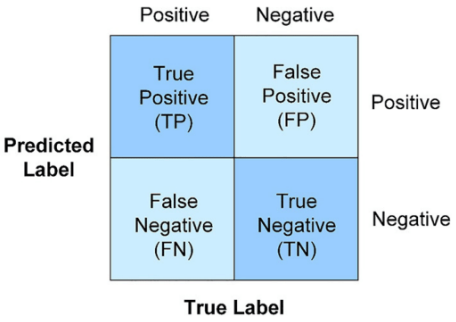
Artificial Neural Network (ANN) is analogical to a human brain. It has neurons interconnected to one another in various layers of the networks. These nodes have weights and biases. When the input, is provided to the node, it uses weights, biases and activation function to transform the input of previous layer to output which will eventually become the input for the next layer. There are also dropout layers which prevent overfitting of the model. ANNs are capable of discerning complex patterns. We have used the ANN of the following architecture to classify people into healthy, pre-diabetic and diabetics:

Layer (type)	Output Shape	Param #
dense_11 (Dense)	(None, 21)	462
dense_12 (Dense)	(None, 132)	2904
dense_13 (Dense)	(None, 254)	33782
dense_14 (Dense)	(None, 368)	93840
dropout_2 (Dropout)	(None, 368)	0
dense_15 (Dense)	(None, 511)	188559
dense_16 (Dense)	(None, 656)	335872
dense_17 (Dense)	(None, 503)	330471
dropout_3 (Dropout)	(None, 503)	0
dense_18 (Dense)	(None, 401)	202104
dense_19 (Dense)	(None, 283)	113766
dense_20 (Dense)	(None, 102)	28968
dense_21 (Dense)	(None, 3)	309

Total params: 1,331,037
Trainable params: 1,331,037
Non-trainable params: 0

2.3 Model Evaluation metrics:

The models are evaluated on metrics like Accuracy, Precision, FBeta score, F1 score, Specificity and Recall. For medical purpose, an algorithm that yields high Precision, high Specificity and high Recall is best suited. Confusion matrix can be used for further insights.



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

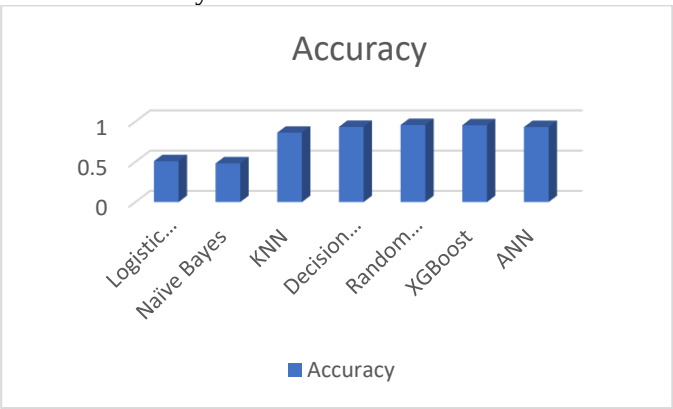
3 Results and Discussions

3.1 Experimental Setup:

The experiment was conducted with the help of libraries like NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn. The deep learning part was created with TensorFlow. The Artificial Neural Network (ANN) is a TensorFlow Sequential model. And all the code was run on cloud (Kaggle iPython notebooks and ANN was run with GPU on the same). The version of all libraries and framework are the default ones provided by Kaggle environment.

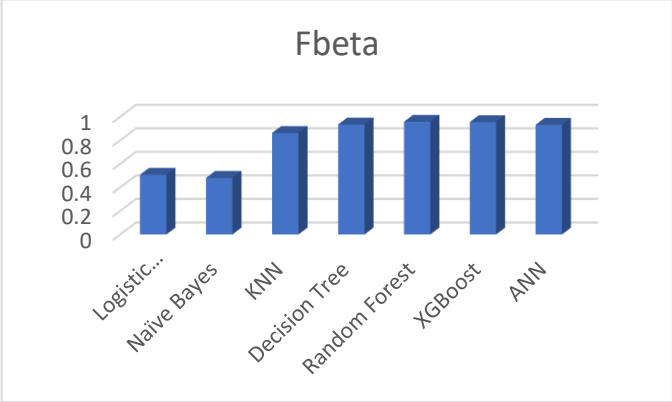
3.2 Result Analysis:

3.2.1 Accuracy



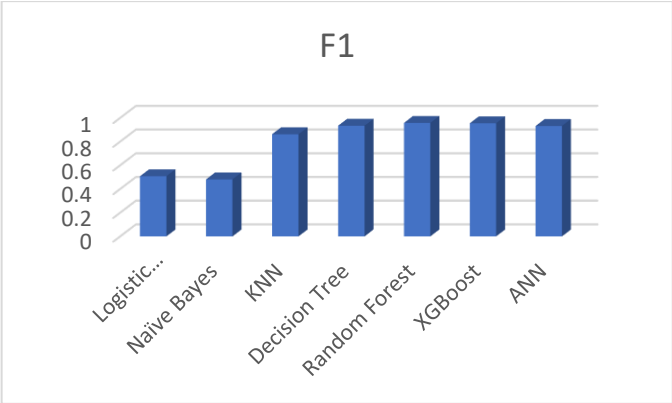
Random Forest and XGBoost have a similar accuracy score of 0.958 and 0.955 respectively.

3.2.2 FBeta:



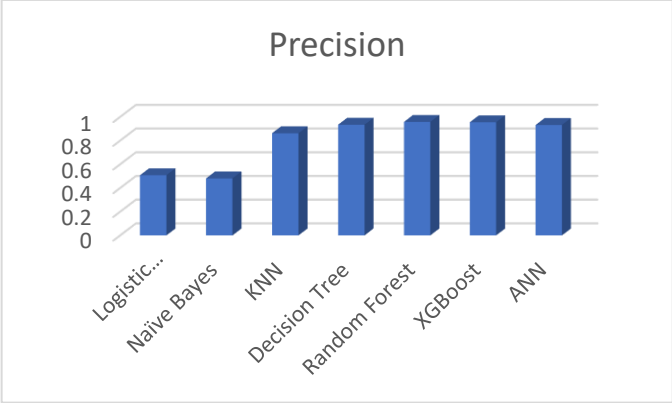
Random Forest and XGBoost have a similar FBeta score of 0.958 and 0.955 respectively.

3.2.3 F1:



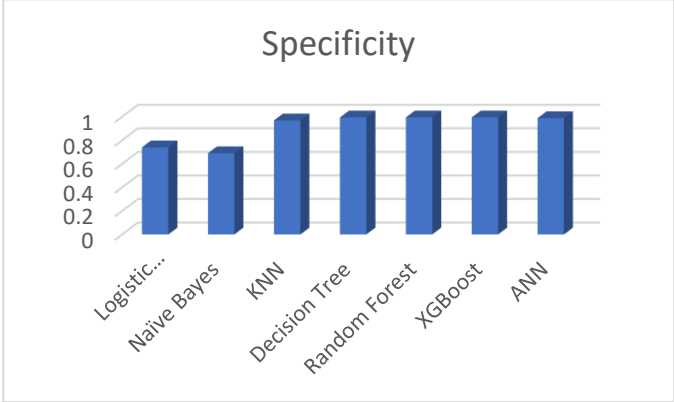
Random Forest and XGBoost have a similar F1 score of 0.958 and 0.955 respectively.

3.2.4 Precision:



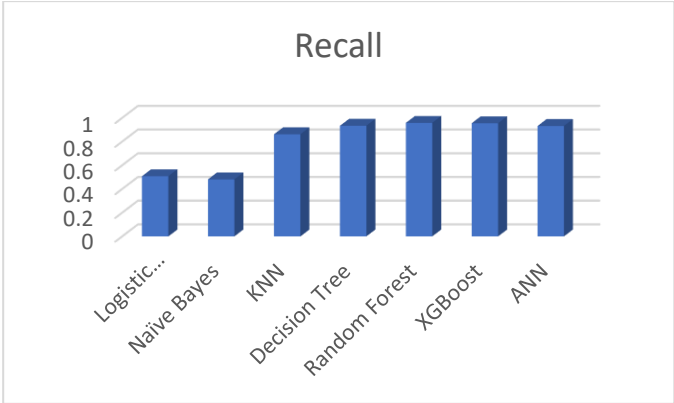
Random Forest and XGBoost have a similar precision score of 0.958 and 0.955 respectively.

3.2.5 Specificity:



Random Forest and XGBoost have same specificity score of 0.994.

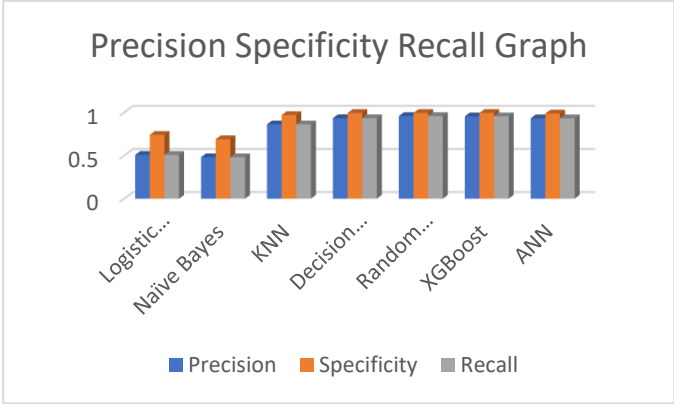
3.2.6 Recall:



Random Forest and XGBoost have a similar recall score of 0.958 and 0.955 respectively.

3.2.7 Precision-Specificity-Recall:

This graph indicates that whichever algorithm with all 3 scores high are more preferred for medical purpose.



Random Forest and XGBoost exhibits similar performance.

4 Conclusion:

While algorithms like Logistic Regression and Naïve Bayes struggle to perform on the dataset, tree-based algorithms and ANN shows promise in their performance. And KNN shows relatively better performance than Logistic Regression and Naïve Bayes.

XGBoost, Decision Tree and Random Forest beat ANN by a negligible margin while there is no identical difference in performance between XGBoost and Random Forest. Therefore, Random Forest, XGBoost, Decision Tree and ANN display tantamount performance levels.

5 References:

[1] Joshi TN, Chawan PP. Diabetes prediction using machine learning techniques. Ijera. 2018 Jan;8(1):9-13.

[2] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018 Jan 1; 132:1578-85.

[3] Behavioural Risk Factor Surveillance System (BRFSS) codebook 2015: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf