NORTH CAROLINA AGRICULTURAL AND TECHNICAL STATE UNIVERSITY

A Study of Disease Diagnosis using Machine Learning

Samin Poudel[*]

[*]*Computational Data Science and Engineering, North Carolina A&T State University, Greensboro, NC 27409*

Presentation follows as below:

- Introduction
- Data, Algorithms and Methods
- Result and Discussion
- Conclusion and Future Work

Introduction

- Artificial Intelligence (AI) and Machine Learning (ML) is successfully applied to practically in every domain like robotics, education, travel to health care
- Various applications of ML in healthcare as shown in the Figure 1
- In this past decade, the investment in AI in healthcare applications has increased significantly

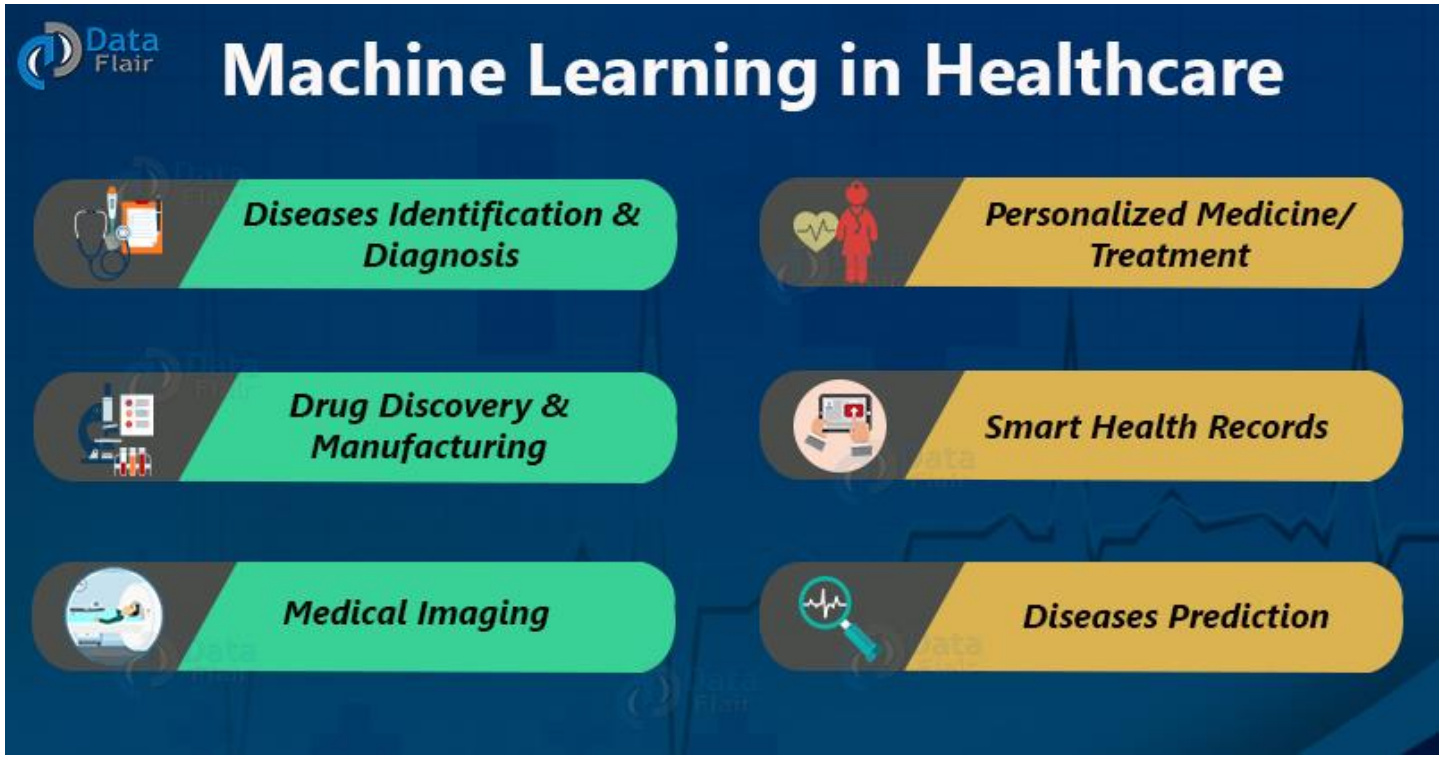

Figure 1: Applications of Machine Learning in Healthcare*

*https://data-flair.training/blogs/machine-learning-in-healthcare/

Introduction

- The analysis of the clinical data can lead to the timely diagnosis of the disease which will help to start cure for the patient in time as well
- Traditional approach of diagnosing disease is generally costly and time consuming
- ML techniques have not only been able to diagnose the common diseases but are also equally capable of diagnosing the rare diseases
- In general, a dataset table used to build a ML model for diagnosing a disease have columns for different attributes and a column variable for the class variable

Problem Statement:

- Accuracy of the ML in diagnosing the diseases is still a concern
- Improvement in the performance of ML to diagnose disease is a hot topic in healthcare domain
- Different ML approach perform differently for different healthcare dataset
- **Need to find the way to apply many state of art algorithms to same dataset in reasonable time with minimal lines of codes**, so that the search of best ML method can be pursued efficiently to diagnose a particular disease

Probable Solution:

- The use of libraries like AutoGluon can help to find the best performing ML approach out of many ML approaches in diagnosing the disease for a given dataset with optimal lines of codes.

Data:

- Dataset Used: Pima Indian Diabetes
- This data set has 8 attributes and one class variable named Outcome.
- Outcome variable has value of 0 or 1, 1 means tested positive for diabetes
- The dataset has 768 instances, 268 instances are tested positive for diabetes

Table 1. Statistical description of Data based on Attributes

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Count	768	768	768	768	768	768	768	768
Mean	3.85	120.89	69.10	20.57	79.79	31.99	0.47	33.24
std	3.37	31.97	19.35	15.95	115.244	7.88	0.33	11.76
min	0	0	0	0	0	0	0.078	21
25% (Q1)	1	99	62	0	0	27.3	0.24	24
50% (Q2)	3	117	72	23	30.5	32	0.37	29
75% (Q3)	6	140.25	80	32	127.25	36.6	0.63	41
max	17	199	122	99	846	67.1	2.42	81.0

Data, Algorithms and Methods

Data:

- Data Exploratory Visualization showed that ML models can be built without preprocessing of the data
- Every attribute may be important for the disease diagnosis with Machine Learning

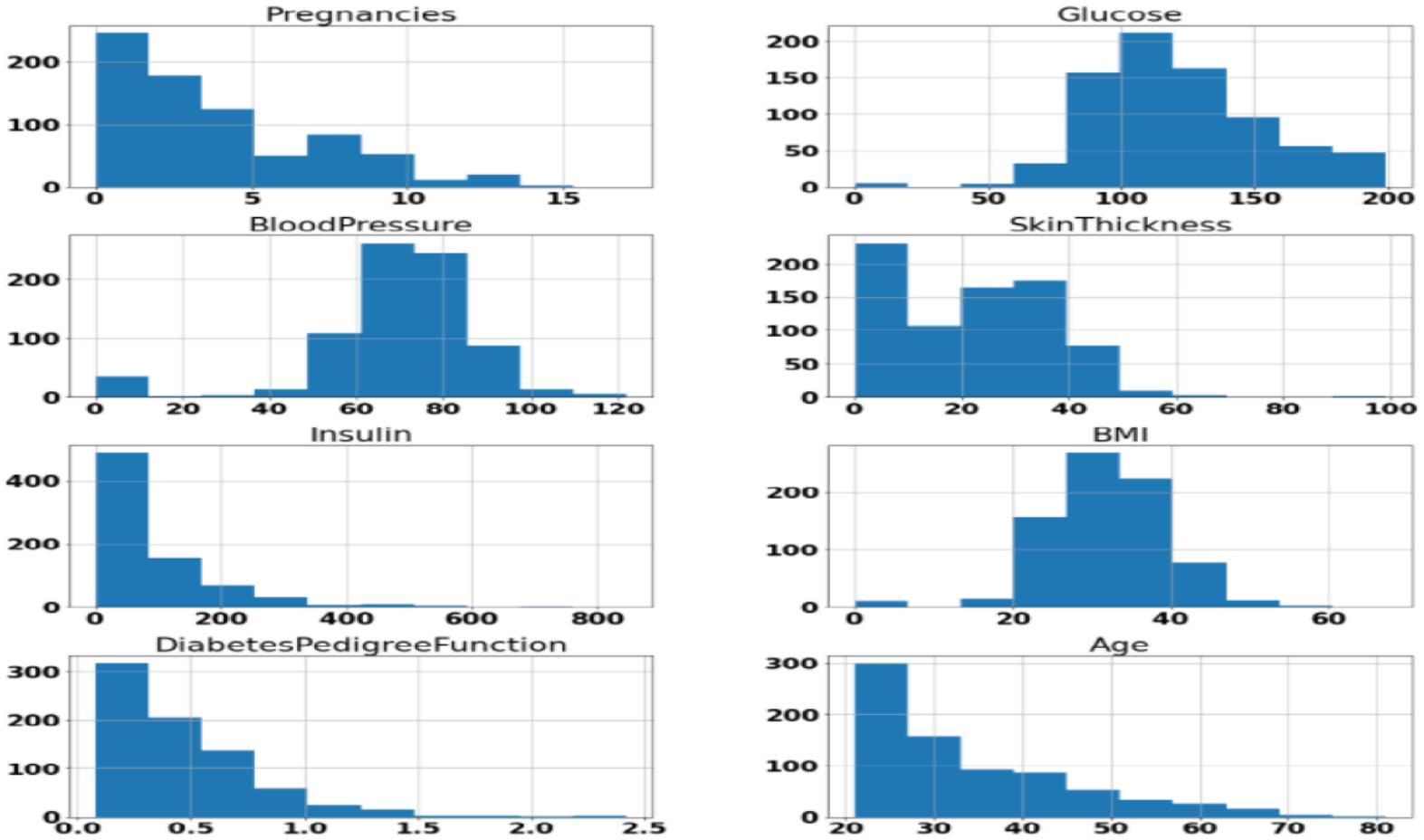

Figure 2: Histogram of Attributes

Machine Learning Algorithms Used:

- 20 Machine Learning Algorithms are used by importing from scikit-learn and AutoGluon Libraries in AWS SageMaker

Table 2. List of ML Algorithms Used

Library	ML Algorithm	Number of ML approaches
Scikit-Learn	Random Forest Classifier, Decision Tree Classifier, Naïve Bayes Classifier, Perceptron, Multilayer Perceptron, Voting Classifier	6
AutoGluon	WeightedEnsemble_L2, LightGBM_BAG_L1, LightGBM_LARGE_BAG_L1, NeuralNetFastAI_BAG_L1, CATBoost_BAG_L1, ExtraTreesGini_BAG_L1, LightGBMXT_BAG_L1, XGBoost_BAG_L1, RandomForestEntr_BAG_L1, RandomForestGini_BAG_L1, ExtraTreesEntr_BAG_L1, NeuralNetMXNet_BAG_L1, KNeighborsUnif_BAG_L1, KNeighborsDist_BAG_L1	14

Data, Algorithms and Methods

- **Overview of Methodology:**
- Data Loaded to Amazon SageMaker's Jupyter Instance
- Data Spitted to Training and Test set
- Machine Learning Algorithms trained and tested using scikit-learn and AutoGluon Library
- Training and Test set for each of the ML algorithm should be same for reasonable comparable among them. It was achieved by defining random seed while splitting data into training and test sets
- Evaluation of ML algorithms to diagnose diabetes are performed using classification metrics Accuracy, Precision, Recall and F1-score
- Detailed Implementation of the ML algorithms is in authors' GitHub page

Evaluation of ML Algorithms:

- Although being a classical ML algorithm, Naive Bayes performed better among the ML algorithms, based on combined analysis of all the evaluation metrics

Table 3. Evaluation of ML Algorithms

S. N	ML Algorithm	Accuracy	F1-score	Precision	Recall
1	Random Forest Classifier (Scikit-learn)	0.74	0.81	0.78	0.84
2	Decision Tree Classifier (Scikit-learn)	0.65	0.73	0.73	0.73
3	**Naïve Bayes Classifier (Scikit-learn)**	**0.77**	**0.83**	**0.80**	**0.86**
4	Perceptron (Scikit-learn)	0.49	0.47	0.71	0.35
5	Multilayer Perceptron (Scikit-learn)	0.68	0.76	0.75	0.77
6	Voting Classifier (Scikit-learn)	0.72	0.78	0.79	0.77
7	AutoGluon Best Performer	0.74	0.82	0.76	0.88

Result and Discussion

- Accuracy performance of different AutoGluon ML algorithms when trained with accuracy as validation metric is in Figure 3a. Similarly, performance in terms of F1-scores is shown when trained with F1-scores as validation metric in Figure 3b
- Weighted Ensemble_L2 ML technique performs better for both the cases and KNN based ML has the least performance for both the cases

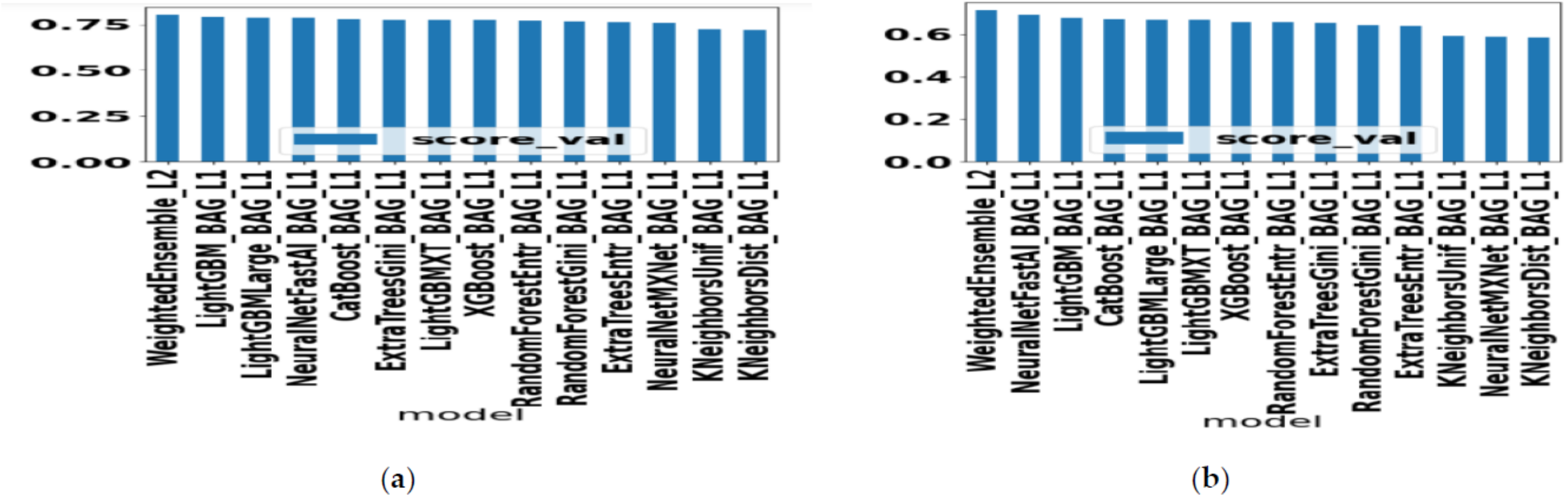

Figure 3. (a) Evaluation of AutoGluon ML algorithms when trained with accuracy as validation metric (**b**) Evaluation of AutoGluon ML algorithms when trained with F1-score as validation metric

Conclusion:

- Libraries like AutoGluon help comparing performances of many ML approaches in diagnosing a disease for a given dataset with optimal lines of code.
- This helps in finding the best performing ML algorithm for a particular dataset or a particular type of disease as well. And it decreases the probability of inaccurate diagnosis, which is a significantly important consideration while dealing with the health of the people.
- Performance of 20 ML approaches in diagnosing diabetes based on the Pima Indian Diabetes Dataset tested
- For the data set considered, Naïve Bayes algorithm performed better among the other algorithms. This shows that using the complex and computationally costly algorithms not necessarily improve the accuracy of diagnosing a disease.

Future Work:

- The possibility of the improvement in the performance of ML models in future can be started by finding the correlation among each attribute and dropping the highly correlated attributes. Because the highly correlated attributes can confuse a model in the learning phase.

THANK YOU