Midterm Report for the Reproduction of Diabetes Prediction Methods

Jeffrey Bird
jeffreyab@vt.edu

Abstract

For the midterm report, our scope included replicating the results of 4 different algorithms: Naive Bayes, logistic regression, decision tree, and random forest.

1 The Dataset

Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques - Islam, M.M.F., Ferdousi, R., Rahman, S., Bushra, H.Y. (2020)
https://link.springer.com/chapter/10.1007/978-981-13-8798-2_12

The dataset referenced in this paper contains 520 observations and 17 attributes that are collected using direct questionnaires and diagnosis results from the patients in the Sylhet Diabetes Hospital in Sylhet, Bangladesh. There are 2 demographic features, Age & Gender; the other 15 attributes are binary features that indicate whether the patient experienced a symptom of pre-diabetes such as: excessive urination, excessive thirst, episodes of sudden weight loss, etc.

1.1 EDA & Data Transformation

- The paper claims that 20 observations in the dataset have NULL values. After investigating, I did not locate any records with NULL values so I did not reduce the number of rows in the dataset.
- I converted all of the categorical fields containing (Yes,No) to binary features with 1=Yes and 0=No.
- In Figure 1, you can see the distribution of the age variable. It is on a much different scale from all of the other binary features so I applied min-max scaling to bring it closer to the scale of the binary variables. In the results, I compare the accuracy of the models using the scaled vs. the raw version of the age variable.

1.2 Summary Stats

Immediately from viewing the summary statistics in Figure 2, I noticed that all of the observations are male. Furthermore, the mean of the response variable "class" is 62% indicating that 62% of observations in the dataset are individuals who are positive for Diabetes. This is much higher than the overall prevalence of diabetes in the population of Bangladesh. "The overall age-standardized prevalence of diabetes was 12.8% (95%CI, 11.2-14.3%) with comparable estimates for men: 12.8%, 95%CI 10.8-14.7 and women: 12.7%, 95%CI 10.9-14.5". This indicates that the data was built on a specific subset of the population and likely cannot be extended to the general population. [1]

[1] https://www.medrxiv.org/content/10.1101/2021.01.26.21250519v3

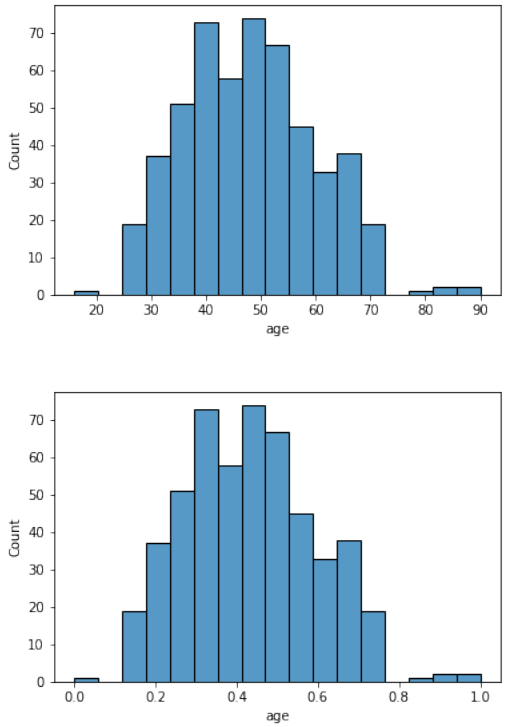

Figure 1: Distribution of Age Feature before and after using min-max scaling. Note that the distribution of the age feature does not change after scaling it between [0,1]

Attributes	Values	Count	Mean	Std
Sex	1.Male, 2.Female	520	1.00	-
Age	1.20–35, 2.36–45, 3.46–55,4.56–65, 6.above 65	520	48.03	12.15
Age_Scaled	Between [0,1]	520	0.43	0.16
Polyuria	1.Yes, 2.No.	520	0.50	0.50
Polydipsia	1.Yes, 2.No.	520	0.45	0.50
Sudden weight loss	1.Yes, 2.No.	520	0.42	0.49
Weakness	1.Yes, 2.No.	520	0.59	0.49
Polyphagia	1.Yes, 2.No.	520	0.46	0.50
Genital thrush	1.Yes, 2.No.	520	0.22	0.42
Visual blurring	1.Yes, 2.No.	520	0.45	0.50
Itching	1.Yes, 2.No.	520	0.49	0.50
Irritability	1.Yes, 2.No.	520	0.24	0.43
Delayed healing	1.Yes, 2.No.	520	0.46	0.50
Partial paresis	1.Yes, 2.No.	520	0.43	0.50
Muscle stiffness	1.Yes, 2.No.	520	0.38	0.48
Alopecia	1.Yes, 2.No.	520	0.34	0.48
Obesity	1.Yes, 2.No.	520	0.17	0.38
Class	1.Positive, 2.Negative.	520	0.62	0.49

Figure 2: Summary statistics for the features in the Sylhet dataset after applying scaling and one-hot encoding. The coloring indicates the attributes with the lowest and highest values of Mean and Std. The lowest values are colored red and the high values are colored green. Notice that after scaling the Age feature, the magnitude of the mean and standard deviation of the age feature is closely aligned with the rest of the dataset.

2 Models

I utilized the sklearn package to train the models. I started by training the following four models with the default parameters only with one exception.

```
from sklearn import ensemble, naive_bayes, tree, linear_model
models = [
    naive_bayes.BernoulliNB(),
    linear_model.LogisticRegression(),
    tree.DecisionTreeClassifier(),
    ensemble.RandomForestClassifier()
    ]
```

The 'naive_bayes' module has several options for training a naive bayes classifier. Because most of the features in the dataset are binary, I chose BernoilliNB. If you pass a non-binary feature to the BernoulliNB, it will automatically transform the feature into binary based on a threshold. [2]

```
    "thresholdfloat, default=0.0
    Feature values below or equal to this are replaced by 0, above it by 1.
    Threshold may not be less than 0 for operations on sparse matrices."
```

For the midterm, I allowed the model to use the default value of 0 so that the model will transform the age feature to 1 for every individual. For the final report, I will transform the age feature into a binary feature. I will bin the ages into buckets such as (0-18,21-25,etc.) and one-hot encode the buckets.

3 Validation

3.1 Train-Test Split

I replicated the 2 different validation methods used in the paper. The first method is an 80/20 train/test split of the data after which there are 416 observations in the training data and 104 observations in the testing data.

```
print(train1_x.shape, train1_y.shape)
print(test1_x.shape, test1_y.shape)
>(416, 16) (416,)
>(104, 16) (104,)
```

Figure 3 contains the accuracy metrics for each model trained on 80% and tested on the 20% test partition.

Model Name	Precision	Recall	F1	Train Accuracy	Test Accuracy
BernoulliNB	0.920635	0.90625	0.913386	0.855769	0.894231
LogisticRegression	0.9375	0.9375	0.9375	0.877404	0.923077
DecisionTreeClassifier	0.969697	1.0	0.984615	0.997596	0.980769
RandomForestClassifier	1.0	0.984375	0.992126	0.997596	0.990385

Figure 3: Model classification metrics

The results are directionally consistent with the results presented in the paper. The precision of the RandomForestClassifier and the Recall of the DecisionTreeClassifier are equal to 1.0; this is suspicious and requires further investigation.

In the final report, I will apply the improvements I previously discussed and I will also include the results of the 10-fold CV.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.naive$_bayes.BernoulliNB.htmlsklearn.naive_bayes.BernoulliNB::text = binarizefloat\%20or, of\%20binary\%20vectors.$