Final Report

Machine Bias in Diabetes Dataset

By

647020016-1 Mr. Yordyod Limrostham

655380002-2 Miss Natnarin Phonprapreut

655380003-0 Mr. Jakkrin Srihajak

665380009-9 Mr. Pipat Kumthinkaew

Present

Asst. Prof. Dr. Chitsutha Soomlek

SC348 810 Software Development and Project Management

for Data Science and Artificial Intelligence Semester 1/2566

College of Computing

Khon Kaen University

Dataset details (milestone 1)

1. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.
2. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Identify the quality issues in the selected dataset (milestone 1)

1. Dataset has some columns that include missing values.

2. Imbalance data Ex. Number of sampling in each classes class 0 (no diabetes) = 500, 1 (has diabetes) = 268

3. Sampling data were collected from specific 1 country, 1 gender and small size of sample in each class.

4. Some features in the dataset have many outlier values.

Goal and Suitable measure for the quality issues (milestone 1)

1. Arrange data to decrease bias in real dataset.
2. Develop model for predicting fairness predicted result.
3. Decrease data bias which is issues from collecting data processes.
    a. Missing data
    b. Data imbalance
    c. Data outliers
4. Analyse model bias.

Model exploration (milestone 1)

1.Naïve Bayes : used for classification, text analysis tasks and make predictions based on a set of features.

2. Logistic regression : is a statistical and machine learning model used for binary classification tasks, where the goal is to predict one of two possible outcomes (classes).

3. Linear Discriminant Analysis (LDA) : a statistical method used for dimensionality reduction and classification tasks in machine learning and pattern recognition. Widely use with multiple classes dataset.

4. Support Vector Machine (SVM) : used for both classification and regression tasks. Particularly effective in scenarios finding optimal hyperplane in a high-dimensional feature space and maximize the margin between the two classes while minimizing the classification error.

Model comparison (milestone 2)

Table 1 : Models comparison results

Model	Accuracy (%)	Training time (s)	Training Memory used (MB)	Testing time (s)	Testing Memory used (MB)	Model size (.pkl format) (MB)	Developer
SVM	69	0.051974	0.257813	0.018001	~0	0.027451	Natnarin
Naive Bayes	71	0.005816	~0	0.001816	~0	0.83	Pipat
Logistic regression	69	0.328451	0.5742188	0.002897	~0	0.000689	Jakkrin
Linear Discriminant Analysis	72	0.011355	~0	0.001333	~0	0.001111	Yordyod

Experiment Dataset: diabetes dataset, contains a total 768 instances and 9 features

From table 1, if you prioritize accuracy, the LDA model is the best. However, if you are more concerned with training and inference cost, Naive Bayes seems to be the most efficient. Lastly, if the size of the model is a critical factor, then Logistic regression is the smallest.

Metric measurement (milestone 2)

Our target of model prediction was to predict patients who tend to have diabetes, so we are going to focus in case patients have diabetes but predicted result give no have diabetes. (false negative values) than patient have diabetes but predicted result give no have diabetes. (false positive values) because patients won't get treatment in time and will become severe diabetes.

Our goals have 2 goals including:

1. 1st goal : Arrange data to decrease bias in real dataset to improve the prediction performance.
2. 2nd goal : Analyse model bias to decrease the prediction bias and have fairness predicted result.

1st goal : Arrange data to decrease bias in real dataset to improve the prediction performance.

Table 2 : Model performance comparison after applied sampling techniques

Model	Sampling Method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
SVM	Under-sampling	66	67	66	66
	Over-sampling	66	67	66	66
	Non-sampling	**69**	**69**	**67**	**67**
Naive Bayes	Under-sampling	70	70	70	69
	Over-sampling	**72**	**72**	**72**	**72**
	Non-sampling	71	71	71	71
Logistic Regression	Under-sampling	67	67	67	67
	Over-sampling	66	66	66	66
	Non-sampling	**69**	**69**	**69**	**69**
Linear Discriminant Analysis	Under-sampling	69	69	69	69
	Over-sampling	69	70	69	69
	Non-sampling	**72**	**72**	**71**	**71**

2nd goal : Analyse model bias to decrease the prediction bias and have fairness predicted result.

Table 3 : Model prediction bias comparison after applied sampling techniques

Model	Sampling Method	No. false negative	Precision(%)	Recall(%)
SVM	Under-sampling	56	67	66
	Over-sampling	75	67	66
	Non-sampling	**49**	**69**	**67**
Naive Bayes	Under-sampling	83	70	70
	Over-sampling	71	72	72
	Non-sampling	**64**	**71**	**71**
Logistic regression	Under-sampling	70	67	67
	Over-sampling	94	66	66
	Non-sampling	**65**	**69**	**69**
Linear Discriminant Analysis	Under-sampling	59	69	69
	Over-sampling	74	70	69
	Non-sampling	**54**	**72**	**71**

Summary

After we use sampling techniques to compare performance and prediction bias of each model showing in table2 and table3. For table2 training of non-sampling dataset can give higher performance accuracy from some machine learning models than doing over-sampling and under-sampling with original dataset except Naive Bayes model after training with over-sampling dataset can have better performance. As we focus on number of false negatives (FN) in table 3 the lowest values of FN came from model that training with non-sampling dataset.

Mitigating bias and Fairness measurement in dataset (Milestone 3)

For mitigating bias we use Reweighting technique showing in table 4 and measure fairness, then we have measured by using two fairness metrics including Mean Difference and Disparate Impact.

Table 4 : Mean Difference and Disparate Impact before and after apply Reweighting technique

	Reweighting	
	Mean Difference	Disparate Impact
Before	-0.3445893662	0.215488063
After	-2.22E-17	1

Table 4 represents comparison of before and after doing Reweighting technique and measure by Mean Difference and Disparate Impact showing that bias was decreasing and more equity in positive outcomes on the protected attribute for the privileged and unprivileged groups.

Table 5 : Classification performance comparison before and after applied bias mitigation techniques.

Model		Reweight			
		Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
SVM	Before	69	69	67	67
	After	**76**	**76**	**69**	**70**
Naive Bayes	Before	71	71	71	71
	After	**74**	**74**	**74**	**74**
Logistic regression	Before	69	69	69	69
	After	**74**	**71**	**68**	**69**
Linear Discriminant Analysis	Before	72	72	71	71
	After	**77**	**75**	**72**	**73**

Table 5 represents the classification performance of individual models, before and after applying bias mitigation techniques. As a result, All of the models have better performance after applying reweighting techniques.

Results analysis and possibility improvement

In terms of performance, we consider accuracy, precision, recall and F1-score by testing on different models. As a result, After the bias reduction process, the average models' accuracy is around 75.25 %.

Next term is about fairness, we consider two metrics Mean Difference and Disparate Impact after we are using bias mitigation methods (reweighting) . According to our experimental results, the bias is potentially reduced for testing various classifiers. The values in both farieness metrics (Mean Difference and Disparate Impact) of two methods are close to zero and close to 1 in order which indicate the improving of fairness.

In order to improve the quality of model, we suggest that the bias reduction process is very important. According to our experiment, without reducing bias, the quality of classifiers are not good enough after we are applying other methods such as data sampling and reweighting. In conclusion, our work we try only 4 machine learning models we suggest experimenting with another various types of machine learning techniques which can suitable to diabetes prediction problem.

Solution or best practices to mitigate the quality issues throughout the system's life cycle

For the solution or best practices, we are giving suggest perspectives description below:

- Recheck that the intelligent to produce result to solve mistakes which do not make more mistakes. (In our case the accepted mistake is that the model predicts a patient don't have diabetes but in fact they have diabetes . (false negative))
- Make sure to check the bias activities that the AI system made not to focus on specific features. (In our case, the system classifies as diabetes for a patient who have volumn of glucose higher than 100)
- Prepare and collect data following the necessary of evaluating processes.
- Setting up the right operating points (precision and recall points) is also an important practice to check which models performs well
- Monitor and check for bias and anomalies of data in the system.
- Avoid black-box models.
- Define system's goals and measurable metrics that we will use.

References

Diabetes dataset : https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?datasetId=2527538&sortBy=voteCount