Milestone2 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

Milestone#2

Task4 : Compare those models

Model	Accuracy (%)	Training time (s)	Training Memory used (MB)	Testing time (s)	Testing Memory used (MB)	Model size (.pkl format) (MB)	Developer
SVM	69	0.05197358131	0.2578125	0.01800131798	~0	0.0274511719	Natnarin
Naive Bayes	71	0.005816	~0	0.001816	~0	0.83	Pipat
Logistic regression	69	0.3284509182	0.57421875	0.002897024155	~0	0.000689	Jakkrin
Linear Discriminant Analysis	72	0.01135492325	~0	0.001333475113	~0	0.00111	Yordyod

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

Metrics:

1.Prediction Accuracy:

- The Linear Discriminant Analysis (LDA) model has the highest accuracy at 72%.
- Naive Bayes comes next with 71% accuracy.
- Both SVM and Logistic regression tie at 69% accuracy.

2.Training Cost (considering both time and memory usage):

- Naive Bayes has the least training time of 0.005816 seconds and negligible memory usage.
- Linear Discriminant Analysis (LDA) has a training time of 0.0113549 seconds and negligible memory usage, making it the second most efficient in training.
- SVM and Logistic regression require relatively longer training times and more memory.

3.Inference Cost (considering both time and memory usage):

- Naive Bayes and Linear Discriminant Analysis (LDA) both have negligible memory usage for testing. Naive Bayes has a testing time of 0.001816 seconds, while LDA has a slightly higher testing time of 0.0013334 seconds.
- Logistic regression requires a testing time of 0.0028970 seconds with negligible memory usage, making it the third most efficient for inference.
- SVM has a testing time of 0.00180013 seconds but uses 0.2578125 MB of memory, which is relatively higher than the others.

4. Model Size:

- Logistic regression has the smallest model size of 0.000689 MB in .pkl format.
- Naive Bayes has the second smallest model size at 0.83 MB.
- LDA has a model size of 0.00111 MB, and SVM has a model size of 0.02745117 MB

In summary, 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.

Task 5: Measure the defined metrics (in Task#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.

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

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 regreesion	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 regreesion	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.

References

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