Effectiveness of Machine Learning Classifiers in Predicting Diabetes vs Non-Diabetes in a Population Sample

COMP3308 - Assignment 2 Report

Student 1: 510407902
Student 2: 510496102

Research Questions and Purposes

Today, people's bodies suffer from many chronic diseases, and diabetes is one of them. It does not only affect the health and living quality of patients but also poses certain burdens and challenges to their families and society. Our study in this report aims to explore from the perspective of the medical field and machine learning algorithms: whether a machine learning classifier can be based on certain human body health indexes as features, thereby effectively predicting whether the population sample has diabetes.

From a medical domain perspective, there is a strict definition of the "effectiveness" of prediction on diseases, which means that our research will adopt a series of evaluation methods to evaluate the effectiveness of the classification results of multiple classifiers.

From the algorithm point of view, by comparing multiple classifiers in the same task background for their classification output results, performance, efficiency, etc., to find the most suitable algorithm for the classification of diabetic patients.

The importance of this study is that if the classifier can effectively classify diabetics and non-diabetics, and satisfactory answers could be given within a fast and accurate range, it will greatly reduce the medical cost of diabetes detection. It helps to allocate limited medical resources more reasonably, thereby generating more benefits to society. But the premise of all this is that the results of the classifier must be highly accurate. If Type I and Type II errors are produced, this will have indelible and serious consequences throughout the research project. The explanation of the error is as follows:

Type I error: If the classifier misdiagnoses non-diabetics as diabetics, it may cause a healthy person to suffer from diabetes treatment, including drug side effects and irreversible psychological effects.

Type II error: If the classifier misdiagnoses a diabetic person as a healthy person, it will produce a very serious medical misdiagnosis event. The problem in the patient's body cannot be found, thus missing the best treatment time, making his physical condition worse and exacerbating the torment of the disease on the body.

1.1 Data

The Pima Indians Diabetes dataset is a widely-used dataset in machine learning and statistics. It contains information from 768 female Pima Indians, including their medical histories and the results of diagnostic tests. The dataset includes eight medical predictor variables: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age, as well as a binary target variable indicating whether the individual developed diabetes or not. This dataset has been used extensively in research on diabetes prediction and management, and has become a benchmark dataset for classification models. The details of the medical predictor variables in the dataset are as follows:

1. *Pregnancies: Number of times pregnant*
2. *Glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test*
3. *Blood Pressure: Diastolic blood pressure (mm Hg)*
4. *Skin Thickness: Triceps skin fold thickness (mm)*
5. *Insulin: 2-Hour serum insulin (mu U/ml)*
6. *BMI: Body mass index (weight in kg/(height in m)^2)*
7. *Diabetes Pedigree Function: Diabetes pedigree function, which represents the likelihood of diabetes based on family history*
8. *Age: Age (years)*

In total, the Pima dataset has a dimension of 768*9, which is 768 records of patients and 9 attributes (8 numeric variables + 1 binary variable outcome). Since the numerical variables in each column have their own units and measurement ranges, in order to manipulate the data better, we used the method of normalization to unify the numerical variable in the range [0,1] in data pre-processing step.

Next, we performed Correlation-based feature selection (CFS). This is a feature selection technique applied to machine learning for cleaning up redundant or irrelevant features in the dataset. The advantage of CFS is that it can handle noisy datasets and then be used with a variety of classification algorithms, which fits the needs of our study. A good result of CFS processing is to select a subset of features that are highly correlated with the target group, but not closely related to each other. Therefore, in the Pima dataset, our filtered subset of features contains these variables: Glucose, Insulin, BMI, Diabetes Pedigree Function, and Age. According to the above list are numbers 2, 5, 6, 7, and 8.

- *Glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test*
- *Insulin: 2-Hour serum insulin (mu U/ml)*
- *BMI: Body mass index (weight in kg/(height in m)^2)*
- *Diabetes Pedigree Function: Diabetes pedigree function, which represents the likelihood of diabetes based on family history*
- *Age: Age (years)*

1.2 Results And Discussion

All classifier results are obtained using 10-fold cross-validation. It does this by dividing the Pima dataset into ten subsets of roughly equal size and then training the model on nine of those folds and testing on the remaining one. By averaging the results from ten separate training and testing epochs, we can obtain a more reliable estimate of model accuracy than a single training and testing split.

We used 9 Weka's built-in classifier models for testing. At the same time, we also implemented two algorithms ourselves, the Naïve Bayes and the K-Nearest Neighbor in which K=1 and K=5 was used. All the accuracy in the table are displayed in the form of percentages. For the changes of each model before and after feature selection, we have taken the form of marking with different colors, where yellow means no change, green means increase, and red means decrease.

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM	RF
No feature selection	65.1%	70.8%	67.8%	74.5%	75.1%	71.7%	75.4%	76.3%	74.9%
CFS	65.1%	70.8%	69.0%	74.5%	76.3%	73.3%	75.8%	76.7%	75.9%

	My1NN	My5NN	MyNB
No feature selection	67.6%	75.5%	74.7%
CFS	67.7%	73.7%	76.3%

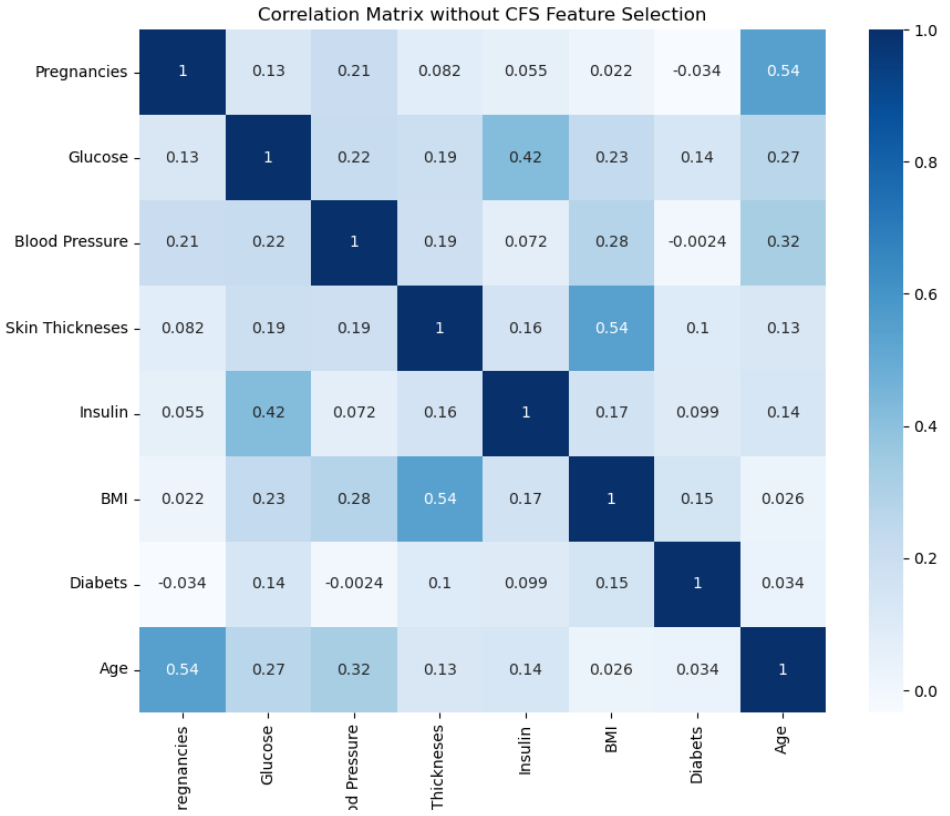
Overall, we end up with a total of 3 unchanged, 8 increased, and 1 decreased accuracy. Among them, Weka's built-in model maintains a constant or rising accuracy rate, but the algorithm implemented by us got a declining value. If no feature selection is used, the top three classifiers performing best in Weka are SVM, MLP, and NB; and in our group, it is KNN which ranks first when K=5. On the contrary, after feature selection, the first three optimal in Weka become SVM, NB, and RF; ours is NB. In both environments, Zero-R and KNN (K=1) are classifiers with low accuracy.

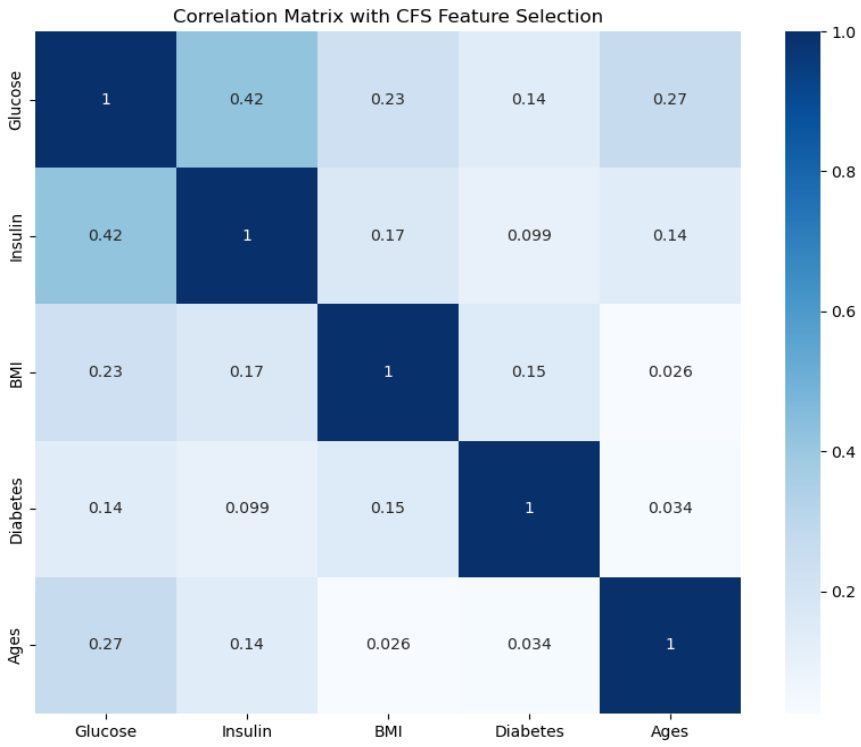
There is no doubt that the performance of the classifier will be better after the feature selection process because it shows a higher overall accuracy. And due to the reduction of the total number of features, it makes the training process faster. In this case, we think the two most suitable classifiers are RF and SVM. Firstly, the accuracy of the support vector machine (SVM) is the highest among all models. This is because of the complexity of its mechanism will makes it have a strong predictive ability, but its complexity makes it barely explainable. So, this is the reason why we also choose random forest (RF), although it is ranked third. One of its advantages is that it is relatively easier to understand and explain. Because the output of the hierarchical shape can clearly show the prediction process of RF, especially since there is a

decision threshold on each branch. As for the poorest classifier Zero-R, the accuracy rate is low because its mechanism is too simple. All outputs are directly predicted as majority class, so it is not sensible to be used in our medical background research.

The selected attributes are believed to be the most intuitive for expressing the presence or absence of diabetes. For example, one of the removed attributes by CFS - whether a woman is pregnant, does not sound like a decisive factor for diabetes screening. We used the same subset of features as Weka and it suppose to ouput the same result. The Naïve Bayes (NB) model did return the same, but the KNN model's output differed. It is possible that this was due to Weka using different distance metrics. In this study, we used Euclidean distance for our self-implemented KNN model to determine the nearest neighbors but Weka might use other.

Under the influence of CFS, the Naïve Bayes algorithm was the one that improved the most in accuracy among all algorithms. The accuracy rises from 74.7% to 76.3% which is a big leap and worth a deeper analysis and to discuss such a phenomenon. Naïve Bayes is a probabilistic classification algorithm that works on the assumption of feature independence, which means it assumes that the features are conditionally independent of each other given the class label. However, in the Pima dataset, this assumption is not quite satisfied because there is existing correlations between the features. CFS tries to avoid selecting correlated features that may be redundant or may not contribute much to the classification performance. This action is considered to be a direct benefit for the Naïve Bayes as it took the dataset closer to the “Naïve” assumption.





According to the correlation matrices shown above, Pregnancies and Age, BMI and Skin Thickness are two pairs of features with the strongest correlation among all features. The correlation coefficients between them are both 0.54. Since CFS removed Pregnancies and Skin Thickness from the dataset, such two pairs of features with strong correlations were eliminated. The correlation coefficient between Blood Pressure and Age is 0.32, which is also considered a relatively strong correlated pair, and Blood Pressure was removed by CFS. Overall, the correlation matrices showed that the CFS reduced the correlation between features, and better satisfied the "Naïve" assumption which significantly boosted the predictive accuracy of Naïve Bayes.

1.3 Conclusion And Future Work

Our research aimed to effectively predict the presence of diabetes in the sampled population, but the study concludes that machine learning classifiers cannot do so with sufficient accuracy. The average accuracy of all our classifiers (after feature selection) is only 72.9%, and the maximum accuracy is 76.7%, which are unacceptable for practical applications. We recognize that accuracy alone is not sufficient to validate the model and plan to incorporate additional evaluation metrics, such as precision, recall, F1-score, and AUC-ROC, to better assess the severity of type I and type II errors.

To improve the predictive ability and performance of the classifiers, we will fine-tune their hyperparameters and leverage more complex evaluation metrics in the future. Ensemble learning can also be used, such as bagging and boosting, to give full play to the advantages of multiple weak classifiers. Moreover, we aim to incorporate updated and modern diabetes-related datasets into our analysis to enhance the stability, robustness, and fairness of the models.

1.4 Reflection

As data science students, this study project has been a valuable learning experience for us. It has exposed us to a complete cycle of data analysis and machine learning projects, providing us with a broader perspective of the ecosystem at the core of modern big data. Throughout the project, we performed crucial data pre-processing steps, including normalization and utilized feature selection techniques CFS. We conducted a rigorous 10-fold cross-validation approach to split the data into training and testing sets for the different classifiers, enabling us to compare their prediction capabilities. Lastly, we applied domain knowledge to evaluate the performance and efficiency of each classifier, bringing the project full circle. Overall, this study has given us a more comprehensive understanding of data analysis and machine learning in real-world scenarios.