Comparison and Analysis of the Impact of PCA on Decision Tree Classification

Peizhi Han

Yixing Chen

CS890 Knowledge Discovery in Databases

Department of Computer Science

University of Regina

Abstract

Principal component analysis (PCA) is a data preprocessing method intended to reduce the dimension of data and improve the results. Decision tree is one of the most popular classification algorithms. The present research aims to discuss the impact of PCA on the decision tree algorithm by two parts. One is to compare the prediction accuracy between classifying with PCA method and without PCA; the second part is to compare the PCA-processed classification with selecting several attributes from the original data set. More specifically, the first part focuses on the general impact of PCA, while the second part is a more in-depth discussion on generating principal components or using original data attributes to predict the classes.

Keywords: principal component analysis, decision tree, classification algorithm, data preprocessing

Table of Contents

List of Figures

List of Tables

1. Introduction

PCA as a popular preprocessing method, it aims to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it (Shlen, 2014). PCA transforms the original data into a set of linearly independent representations of each dimension through a linear transformation. It can be used to extract the main feature components of the data. It is often used to reduce the dimensions of high-dimensional data.

Decision tree has been one of the main algorithms for classification. The basic idea behind a decision tree is that it recursively divides a training set until each division consists entirely or primarily of examples from one class (Sharda, Delen & Turban, 2019).

This research aims to study the impact of principal component analysis (PCA) on decision tree classification algorithm; get two principal components by PCA, then try every combination of two attributes from the data set, find which way could better predict the classification result.

Our research will use three data set from UCI Machine Learning data repository: heart disease ("Heart Disease Data Set", 1988), breast cancer ("Breast Cancer Wisconsin (Diagnostic) Data Set", 1995), and default of credit card clients ("Default of credit card clients Data Set", 2016) data set.

The tools of implementation are Google Colaboratory with Python and scikit-learn libraries.

The present research paper consists of the following sections: 1. Introduction on the project; 2. Comparison and discussion of several related works; 3. Statement of the problem; 4. Detailed explanation of the approach; 5. Discuss on the results; 6. Limitation and some possible extensions; 7. Conclusions of the research.

In Section 4 Approach, we will focus on the implementation steps. The method of the implementation follows the knowledge discovery from databases (KDD) process (Fayyad, Piatetsky-Shapiro & Smyth, 1996): data selection, data pre-processing, data transformation, data mining and evaluation. Since the purpose of the present research is to discuss the impact of PCA by comparison, we will divide the implementation phase into two sections: the approach with PCA and the one without PCA. To avoid the influence of different decision tree criterions, we will compute the accuracies with both Gini and information gain.

2. Related Work

The paper "A tutorial on Principal Components Analysis" (I Smith, 2002) is a technical report aiming to explain the basic ideas of PCA. This paper consists of four chapters. Chapter 1 is to introduce the tutorial, more specifically, the purpose and structure of the report. Chapter 2 reviews the mathematical background. This chapter uses an example to show how to calculate the PCA step by step instead of plainly presenting theories and formulas. Meanwhile, the readers can also learn concepts, including the standard deviation, variance, covariance, and covariance matrix, as well as matrix algebra, eigenvectors, and eigenvalues. Next is explaining the Principal Components Analysis with more details. The first step is to acquire data. In this report, the author creates a data set of two dimensions. The second step is to subtract the mean from each of the data dimensions. The third step is calculating the covariance matrix so that the covariance matrix will be 2 x 2. The fourth step, calculate the eigenvectors and eigenvalues of the covariance matrix. The fifth step is choosing components and forming a feature vector. The sixth step is to derive the new data set. Finally, the author shows how to get the original data back. Chapter 4 is an application with computer vision. It introduces the PCA to find patterns and PCA for image compression. In the appendix, the author also attaches the implementation code by Scilab.

“PCA Based Feature Reduction to Improve the Accuracy of Decision Tree C4.5 Classification” (Nasution, Sitompul & Ramli, 2018) is a research focusing on the PCA influence on decision tree C4.5. In this research, the authors use PCA to improve the accuracy of decision tree C4.5 classification algorithm with UCI Cervical cancer data. The problem and motivation of this research are: during the decision tree classification process, there is a problem of over-fitting resulting from noisy data and irrelevant features. Therefore, the authors propose a framework for

feature reduction using PCA to eliminate irrelevant features, selecting relevant and non-correlated features without affecting the information contained in the original data. This research is an approach of using PCA to improve the performance of decision tree C4.5 classifier. Between C4.5-Non-PCA and C4.5-PCA approaches, the accuracy has been improved, which proves the importance of feature reduction in the classification model.

Another research "The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data" (Howley, Madden, O'Connell & Ryder, 2006) applies PCA to five well-known performance of five well known machine learning techniques (Support Vector Machines, k-Nearest Neighbours, C4.5 Decision Tree, RIPPER and Naive Bayes) along with classification by Linear Regression are compared by testing on a high dimensional data set (Raman spectral data). During the phase of evaluation, the authors find that some classifiers suffer drastic increases in error within the range of PCs tested: PCR, RBF SVM, and k-NN (although not to the same extent as the previous examples). In contrast, the error for C4.5 never deviates too much from its lowest error at six PCs. This may be due to its ability to prune irrelevant attributes from the decision tree model.

The tutorial mentioned firstly is an educational-purposed report and focuses on the concept of PCA, which builds the theory cornerstone for the present research. It provides a step-to-step introduction and detailed explanation of PCA from a mathematical perspective. The second and third research paper are two approaches to discuss the influence of PCA on machine learning methods. The second paper discusses the theory and conducts an experiment on the PCA impact on the C4.5 algorithm. Then proves the importance of the PCA approach on the classification models. On the other hand, the third paper is a research aiming to find PCA impact

on several algorithms, including decision tree. Through this research, we could find that the influence of PCA on decision tree is less than other algorithms.

3. Statement of Problem

As mentioned in the previous section (Section 2. Related Work), there have been several related research papers discussing how PCA can improve the accuracy of classification algorithms, specifically on the decision tree algorithm. Despise the positive results achieved from this research; there are also research pointing out that PCA has little improvement in decision tree, among other classification algorithms.

This generates the motivation of the present research: conduct a research specifically on the impact of PCA on decision tree algorithms by comparing the prediction with the PCA approach and without PCA to pre-process the data. Moreover, as a dimensional reduction method, PCA transforms the attributes into some new attributes. In order to discuss further the impact, we also experiment by only selecting the same number of attributes from the original data set, try to predict the classification results by the selected attributes. Then compare the results to verify if the generation of principal components can perform better than just using the same number of original attributes to predict.

4. Approach

The implementation process will consist of data selection, data pre-processing, data transformation, data mining and evaluation; then divided by two approaches: with PCA and without PCA method for the comparison.

In the data selection phase, we select from UCI Machine Learning data repository: Heart Disease Data Set contains 14 attributes and 303 instances; Breast Cancer Wisconsin (Diagnostic) Data Set contains 31 attributes and 569 instances; Default of credit card clients Data Set contains 24 attributes and 30,000 instances. There are two relatively small data set, to avoid the influence of data size, we compare with one larger data set.

Furthermore, the decision tree classification model in scikit-learn library uses an optimized version of the CART algorithm. Hence, scikit-learn implementation does not support categorical variables ("1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART", n.d.). In this case, in order to utilize the function of decision tree classifier, the data need to be numerical variables.

4.1. Approach with PCA method

The approach with PCA method consists of data preprocessing with missing values, converting data type, and using PCA to generate two new attributes (principal components) to replace the original numerical attributes.

4.1.1. Data preprocessing and transformation

Heart Disease data set:

1. Check missing values: no missing values in this data set.
2. Five numerical variables and seven categorical variables. Hence, we use Dummy Variables to transform all categorical data into numerical data. For example, there are "male" and "female" values in the "sex" attribute which are categorical. After the

Dummy function, it is transformed into “sex_male”. The value of this attribute becomes numerical variables zero and one: zero means “not male”, one means “is male”. Therefore, after this step, the seven categorical attributes are transformed into 15 numerical attributes.

3. Apply PCA to reduce the dimension of numerical data into two dimensions. Using the two generated new principal components to replace the original five numerical attributes, then add up the 15 attributes from Step 2. A new data set is generated, as shown in Figure 1:

	PC 1	PC 2	target_Yes	ca	sex_male	cp_Atypical Angina	cp_Non-anginal Pain	cp_Typical Angina	fbs_True	restecg_Normal	restecg_ST-T wave abnormality	exang_Yes	slope_Flat	slope_Upsl
0	-1.267716	-0.082358	1	0	1	0	0	0	1	1	0	0	0	
1	0.932350	-0.070433	1	0	1	0	1	0	0	0	1	0	0	
2	1.419900	-0.389838	1	0	0	1	0	0	0	1	0	0	0	
3	0.920091	0.263924	1	0	1	1	0	0	0	0	1	0	0	
4	0.047331	1.591873	1	0	0	0	0	1	0	0	1	1	0	

Figure 1. New data set after PCA-heart disease data set

4. Plot correlation between the two principal components, as shown in Figure 2. The two colors of data points correspond to two target values: Yes (red) and No (blue). We can find from the plot that the two principal components have some overlaps and are not clearly separated:

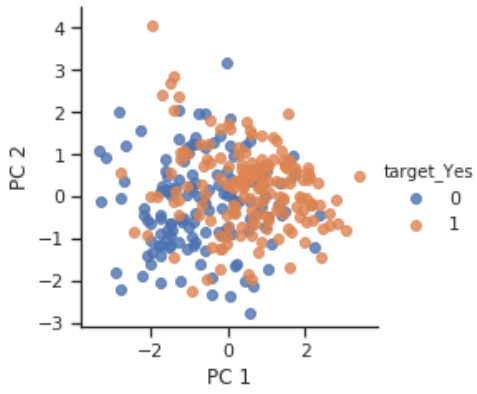


Figure 2. Correlation between PC1 and PC2 for heart disease data set

Breast Cancer data set:

1. Check missing values: no missing values in this data set.
2. Since all attributes are all numerical, we do not need to apply Dummy Variables to transform categorical data.
3. Apply PCA on all attributes then reduce the attributes into two dimensions. Therefore, we generate a new data set with only two numerical attributes (two principal components), as shown in Figure 3.

	PC 1	PC 2	diagnosis_M
0	9.192837	1.948583	1
1	2.387802	-3.768172	1
2	5.733896	-1.075174	1
3	7.122953	10.275589	1
4	3.935302	-1.948072	1
...	...	...	...

Figure 3. New data set after PCA-breast cancer data set

4. Plot correlation between the two principal components, as shown in Figure 4. The two principal components have less overlaps than the heart disease data set, and can be separated more clearly.

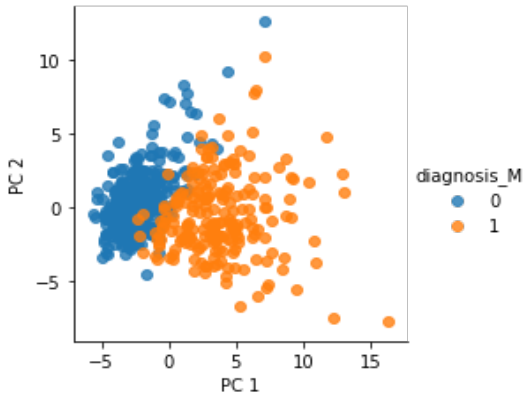


Figure 4. Correlation between PC1 and PC2 for breast cancer data set

Default of Credit Card Clients data set:

1. Checking missing values: there are missing values in this data set, we remove all the instances with missing values. Then get a new data set with 15336 instances.
2. Since the credit card data set consists of both categorical variables and numerical variables, the preprocessing step is the same as heart disease data set.
3. Same as Step 2, the transformation is the same as heart disease data set as well.
4. The plot of two principal components is shown in Figure 5. The overlaps of the two components is larger than the second data set.

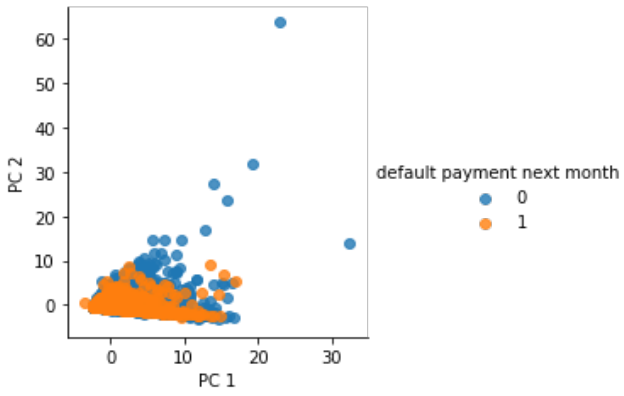


Figure 5. Correlation between PC1 and PC2 for credit card data set

4.1.2. Modeling

To build the model, we first split the data set into 80% and 20%. Then set the random seed of splitting to be 4, so that the splitting will be consistent as in Figure 6:

```
# split the data
x_train, x_test, y_train, y_test = train_test_split(final_df.drop('target_Yes', 1),   # delet
the label
                                 final_df['target_Yes'],        # label set
                                 test_size = .2,         # test size 20%, train size 80%
                                 random_state = 4)       # random seed, same seed same output
```

Figure 6. Split data for training and testing

Next, we import the decision tree classifier from scikit-learn. Since the supported criteria is “gini” for the Gini impurity and “entropy” for the information gain

("sklearn.tree.DecisionTreeClassifier", n.d.), we utilize these two criteria for further computation of accuracy, as shown in Figure 7:

```
clf = tree.DecisionTreeClassifier(criterion='gini')  clf = tree.DecisionTreeClassifier(criterion='entropy')
```

Figure 7. Decision tree criterion: Gini and information gain

After setting the classification criterion, we train the model and plot tree as in Figure 8:

```
clf = clf.fit(x_train, y_train)
tree.plot_tree(clf.fit(x_train, y_train))
```

Figure 8. Training the decision tree model

4.1.3. Testing and evaluation

The accuracy of classification is computed by comparing the predicted results with the testing data. Figure 9 shows the accuracy of heart disease prediction under Gini criteria:

```
score = clf.score(x_test, y_test)
print('The accuracy is:', score)
```

```
The accuracy is: 0.7704918032786885
```

Figure 9. Model evaluation on heart disease data set using Gini

4.2. Approach without PCA method

Section 4.1 explains the approach with PCA to reduce dimension, hence, the next part proceeds with the preprocessing without PCA. As indicated in the first approach, new attributes and values are generated. However, these new attributes and values have no logical meaning in the data set. Therefore, in order to compare the influence of PCA, we utilize the same number of attributes from the data set, which means we select two attributes from the original data set. So the idea of how to make selection and comparison is critical.

For non-PCA approach, firstly, we extract all attributes into a list. Secondly, the list can do combinations with its elements. Thirdly, programming runs the loop to find all combinations of two attributes among all of the attributes, then proceed with the training and testing in

decision tree. Fourthly, calculate the prediction accuracy of each combination. Finally, compare all accuracies, find a combination with the highest accuracy rate, and visualize these two attributes and data.

4.2.1. Data preprocessing and transformation

The data preprocessing and transformation for the approach without PCA is the same as Step 1 and 2 in Section 4.1.

4.2.2. Modeling

The modeling step of the non-PCA approach is the same as the PCA approach.

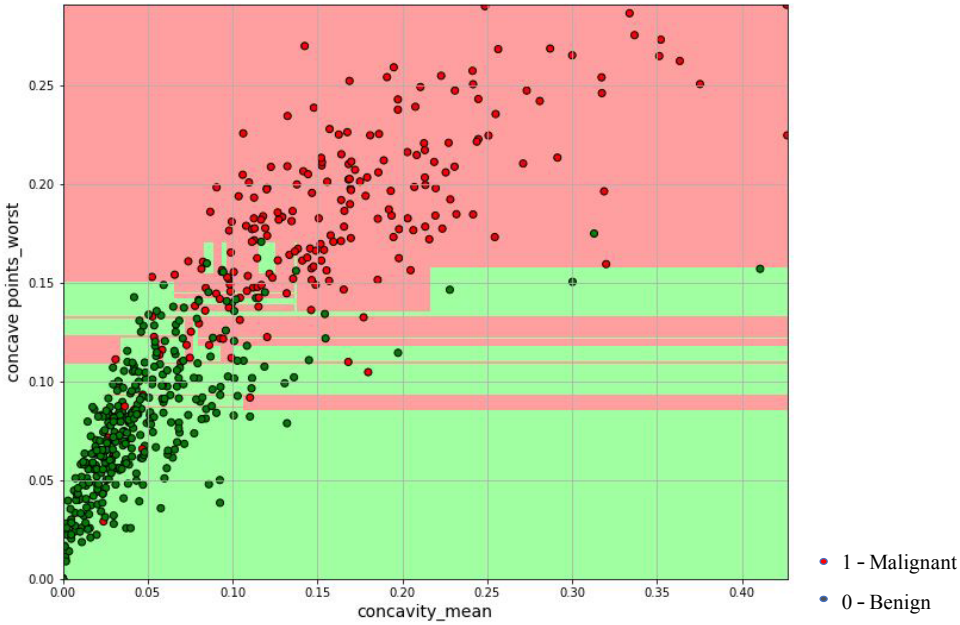
4.2.3. Testing and evaluation

This part has an additional visualization approach, which displays the predicted attribute combination and test results in two-dimensional graphs.

It is worth noting that the non-PCA approach uses original data, and the PCA approach gets two new attributes and values based on all the attributes and values. Although PCA can reduce the interference of noisy data, non-PCA only uses two original attributes and values. This also means in non-PCA method, the influence of much noisy data is reduced.

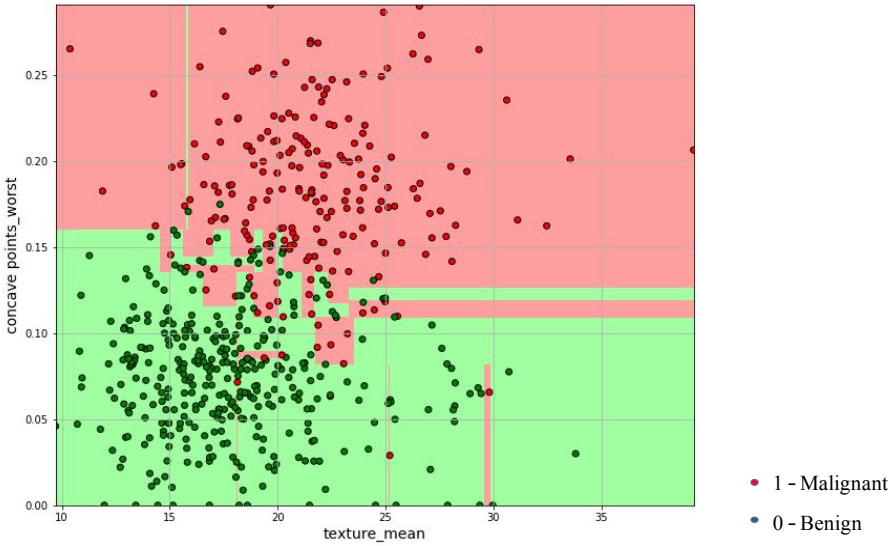
Since the new attributes generated by PCA are logical, there is no inherent association and logic between the two new attributes, and the values of the two attributes will also appear negative. These negative numbers are also incomprehensible. Non-PCA approach uses primitive attributes and values, so these dimensions and values are understandable. The most important thing is that we can find the most accurate combination of attributes, which also means that there is a strong relationship between these two factors and the disease diagnosis result. The scatter plots Figure 10 and Figure 11 shows the visualization for classification based on Gini and

information gain for the breast cancer data set:



concavity_mean + concave points_worst, Testset accuracy: 94.74%

Figure 10. Breast cancer data set: the two attributes with highest accuracy (Gini)



texture_mean + concave points_worst, Testset accuracy: 94.74%

Figure 11. Breast cancer data set: the two attributes with highest accuracy (information gain)ß

Unlike the breast cancer data set, the other two data sets both contain categorical data and numerical data. Taking credit card data set as an example: Figure 12 shows the correlation between two categories, not numerical values. So we can see that all of the points on the coordinates fall on integers because each integer represents a category.

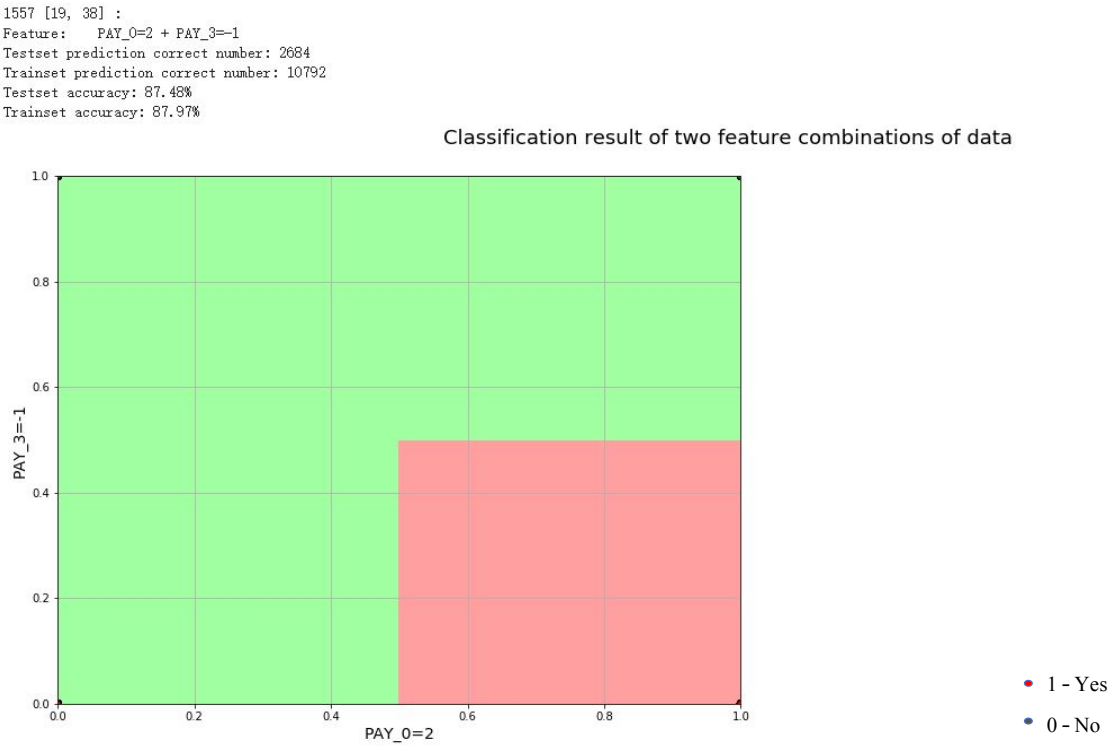


Figure 12. Credit card data set: the two attributes with highest accuracy (Gini)

From the comparison and analysis above, we can find that two attribute combinations can be used not only for numerical values but also for training and prediction of classes. However, PCA can only be used for numerical values. The most important thing is that the two attribute combination is not only more accurate than PCA, but also reflects the importance of attributes. More figures on the correlation are presented in the Appendix.

5. Results

By calculating the accuracy rate for all cases, we generate an accuracy matrix, as shown in Table 1:

Table 1 Accuracy Matrix

Measurement / Databases	PCA		NON-PCA			
	Gini	Information Gain	Gini		Information Gain	
			All Attributes	Two Highest Attributes	All Attributes	Two Highest Attributes
Heart Disease	77.0%	75.4%	75.4%	86.9%	78.6%	86.9%
Breast Cancer	88.6%	86.8%	90.4%	94.7%	87.7%	94.7%
Credit Card	79.5%	80.0%	79.2%	87.5%	78.1%	87.5%

For the two small data sets (heart disease data set contains 303 instances, breast cancer data set has 569 instances) and one larger data set (credit card data set has 15336 instances after the pre-processing), our first analysis is to compare the approach of PCA approach and non-PCA approach. We use PCA and decision tree with Gini and information gain to compare with non-PCA and decision tree with Gini and information gain (all attributes):

1. In Heart Disease data set, the difference of accuracy between PCA approach and Non-PCA approach is small: under Gini criteria, 1.6%; under information gain criteria, is -3.2%.
2. In Breast Cancer data set, PCA approach has accuracy slightly lower than non-PCA approach (the differences are -1.8% under Gini and -0.9% under information gain).
3. In Credit Card data set, the difference between these two approaches is even smaller, despite that PCA is relatively higher than non-PCA (0.3% under Gini and 1.9% under information gain).

The second analysis is to compare the “two highest attributes” with PCA approach:

Either for small size data set or for larger data set, the approaches with only two selected attributes have the highest accuracy rates.

6. Limitations and Possible Extensions

Due to the limitation of time and resources, in further research, we can conduct a more profound and broader study on the PCA impact:

In the present research, we only test on three data sets. In order to exclude the influence of data size on the results, we may test on more data sets, including several small sizes and several relatively larger size data sets.

Decision tree is the classification algorithm that has been discussed in the present research. To extend the investigation of PCA impact, we can further test with more classification models such as random forest.

Only two principal components and two attributes are selected and set-up in this research. To be more objective, more dimensions could be tested for the comparison.

As experimental research, we could consider more factors that might explain the results. For instance, using statistical methods to verify if the data sets are Gaussian distribution. Then discuss PCA on Gaussian and non-Gaussian distribution data set.

7. Conclusions

To conclude, this has been a research report aiming to find the impact of PCA on the decision tree algorithm. From the evaluation results, we find that PCA is not a data preprocessing method that could improve the accuracy of all data sets. Moreover, in this research, using the selected attributes from the original data can make better predictions than principal components. For a data set with all numerical variables, PCA could have a better performance than the data sets with both numerical and categorical variables. From decision tree perspective, in some cases, Gini and information gain do not have significant differences in accuracy.

In a nutshell, although PCA is an importance preprocessing method for machine learning algorithms, choosing reducing dimension with PCA should also consider many factors in the experiment. Other methods such as selecting attributes from the data set may perform better than PCA method.

References

1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART. Retrieved 28 November 2019, from https://scikit-learn.org/stable/modules/tree.html

Breast Cancer Wisconsin (Diagnostic) Data Set. (1995). Retrieved 24 November 2019, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Default of credit card clients Data Set. (2016). Retrieved 24 November 2019, from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining, 17*(3), 41.

Heart Disease Data Set. (1988). Retrieved 24 November 2019, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

I Smith, L. (2002). A tutorial on Principal Components Analysis. *Computer Science Technical Report*. doi: OUCS-2002-12

Nasution, M Z F, O S Sitompul, and M Ramli. 2018. "PCA Based Feature Reduction To Improve The Accuracy Of Decision Tree C4.5 Classification". *Journal Of Physics: Conference Series* 978 012058. doi:10.1088/1742-6596/978/1/012058.

Howley, T., Madden, M., O'Connell, M., & Ryder, A. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, *19*(5), 363-370. doi: 10.1016/j.knosys.2005.11.014

Sharda, R., Delen, D., & Turban, E. (2019). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support* (11th ed.). Pearson, 227.

Shlen, J. (2014). A Tutorial on Principal Component Analysis. *Machine Learning (Cs.LG)*. doi: arXiv:1404.1100

sklearn.tree.DecisionTreeClassifier. Retrieved 28 November 2019, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Appendix A

```
87 [5, 13] :
Feature:    ca + exang_Yes
Testset prediction correct number: 53
Trainset prediction correct number: 182
Testset accuracy: 86.89%
Trainset accuracy: 75.21%
```

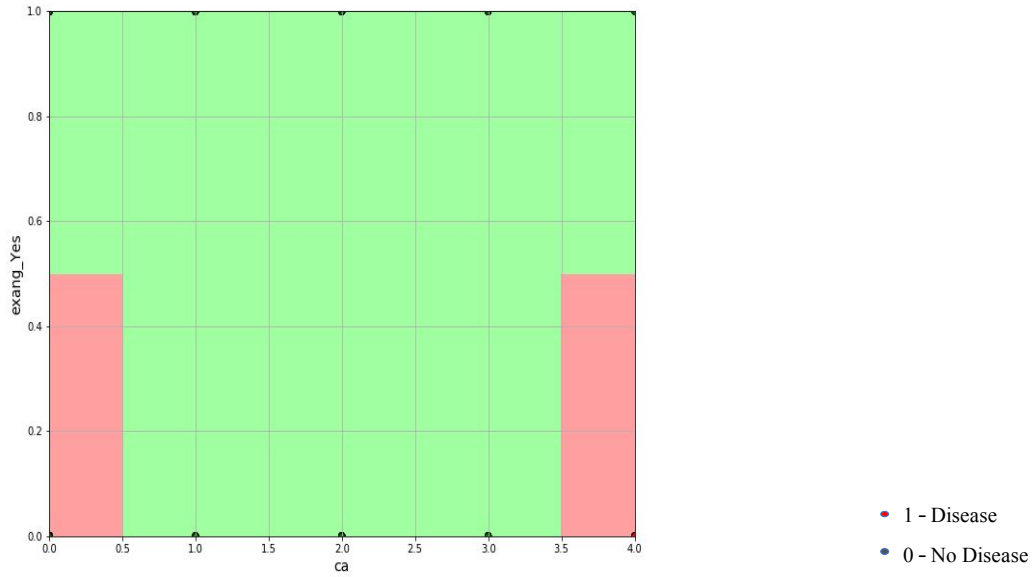


Figure 13. Heart Disease data set: the two attributes with the highest accuracy (Gini)

```
87 [5, 13] :
Feature:    ca + exang_Yes
Testset prediction correct number: 53
Trainset prediction correct number: 182
Testset accuracy: 86.89%
Trainset accuracy: 75.21%
```

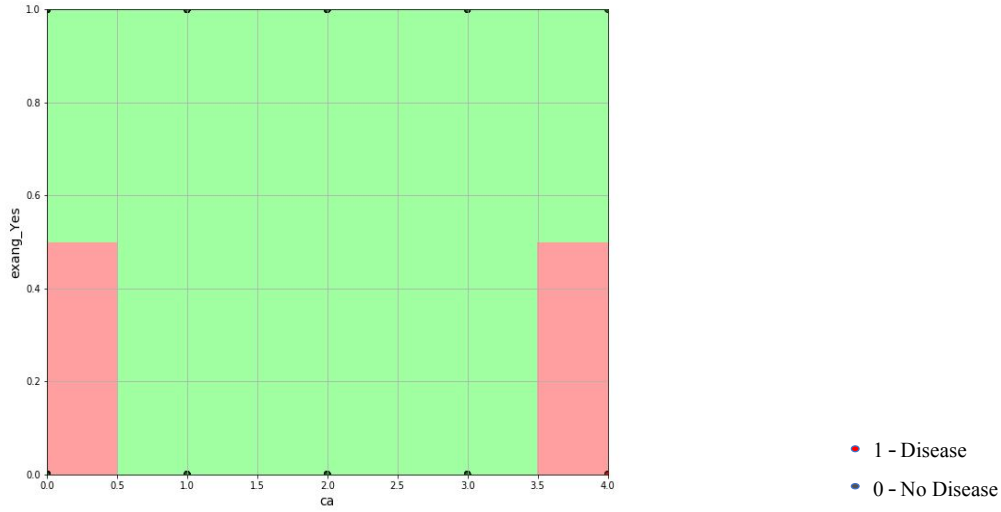


Figure 14. Heart Disease data set: two attributes with the highest accuracy (information gain)

Appendix B

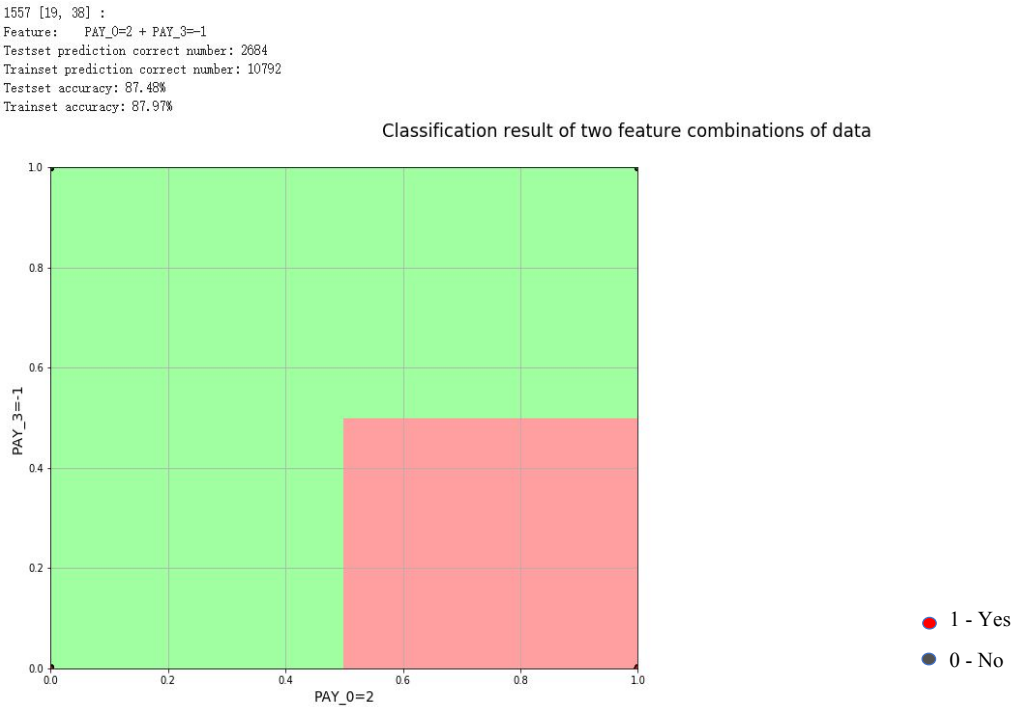


Figure 15. Credit card data set: two attributes with the highest accuracy (information gain).