南京大学数据挖掘报告

对发展中国家发展潜力的评估报告

Assessment report on the development potential of developing countries

Name	Student_id	Class
Kangxin Liu(Team Leader)	MF1832098	1
Xiang Li	MF1832088	1
Wenyao Li	MF1832087	1
Yinzhong Liang	MF1832092	2

南京大学

NanJing University

Contents

1 Introduction

1.1 Project Background

There are nearly 200 countries in the world, but the number of developed countries is only 10% of the total number of countries in the world. For each country, they are expecting their country to break through the current development bottleneck and become the leading country in the world. Therefore, it is necessary to observe and analyze the potential of those countries in the world to become developed countries.

Therefore, the experimental group used the “countries of the world” above Kaggle as the experimental data set, and discovered and analyzed the unified characteristics of developed countries through data mining algorithms, and calculated the development potential of developing countries based on these characteristics, thus obtaining development. The result of the development potential of the Chinese state. Through analysis, it can remind countries to focus on the development of those industries,

1.2 Data sets and tasks

The data set we selected is "countries of world" above the website of “kaggle”. In this data set, we include data on 20 attributes of such as population density, service industry, industry, GDP, etc. in 227 countries around the world. Through these data, our data mining task is to analyze the data and find out the common characteristics of all 24 developed countries as a basic indicator to measure and analyze whether other non-developed countries have the potential to become developed countries, and finally to be tested. Identify and list those countries with the potential to become developed countries in developed countries, and give suggestions on the development of corresponding attributes of the country.

2 Data Mining Algorithm

The whole process of experimental is shown as Figure2.1:

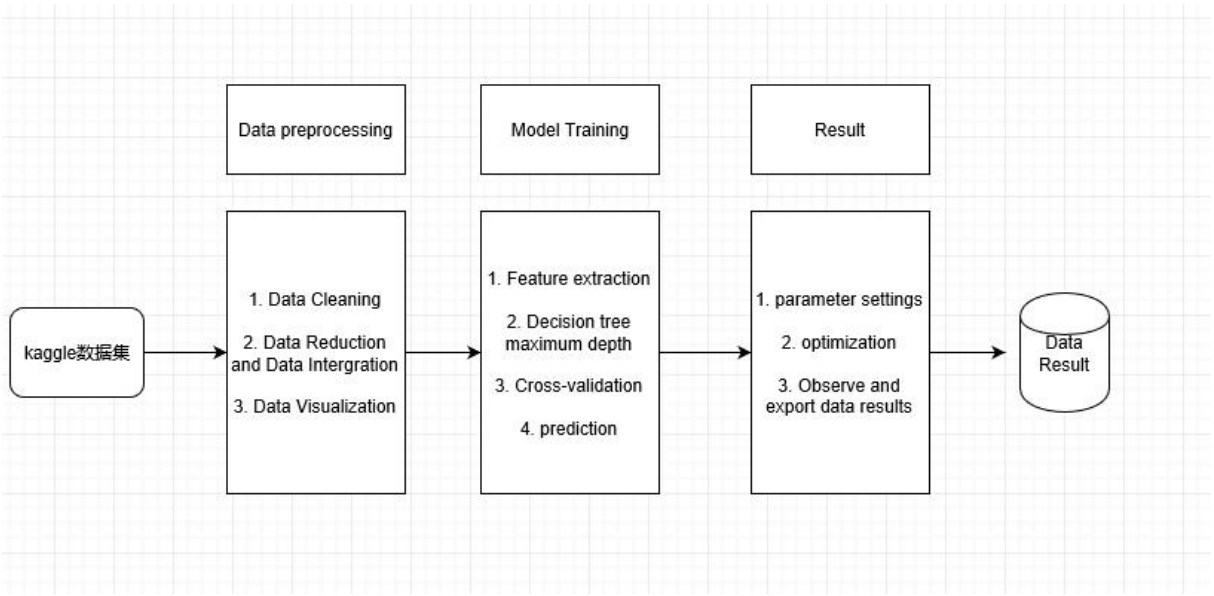

Figure 2.1 Flow Chart

2.1 Data preprocessing

2.1.1 Data Cleaning

In the process of data cleaning, we examined all the attributes, related processing of abnormal, missing and difficult to calculate attribute values, making the data set easy to calculate while ensuring data integrity.

1. Processing abnormal data

The format of the attribute value of the floating point number type in the data set is abnormal. A comma appears in the value. According to the attribute label, the original data should be a floating point number. Therefore, the author replaces all the commas in the floating point number with a decimal point.

2. Handling missing data

If the country name in the dataset is missing, delete the row record;

The "Population", "InfantMortality", "Literacy", "Birthrate", and "Deathrate" attribute values are missing and filled with the world average; look up the data to fill in the missing "Area", "Coastline", and "GDP" attribute values.

3. Delete attributes

"Region", "Phones", and "Climate" have little effect on measuring whether a country is developed, so the three columns of attribute data are deleted.

2.1.2 Data Integration & Visualization

In order to find out the main factors affecting the development of a country, the author has done a lot of data exploration, such as drawing a histogram of the relevant attributes of all countries, taking the per capita GDP of all countries as an example, as shown in Figure 2.2.

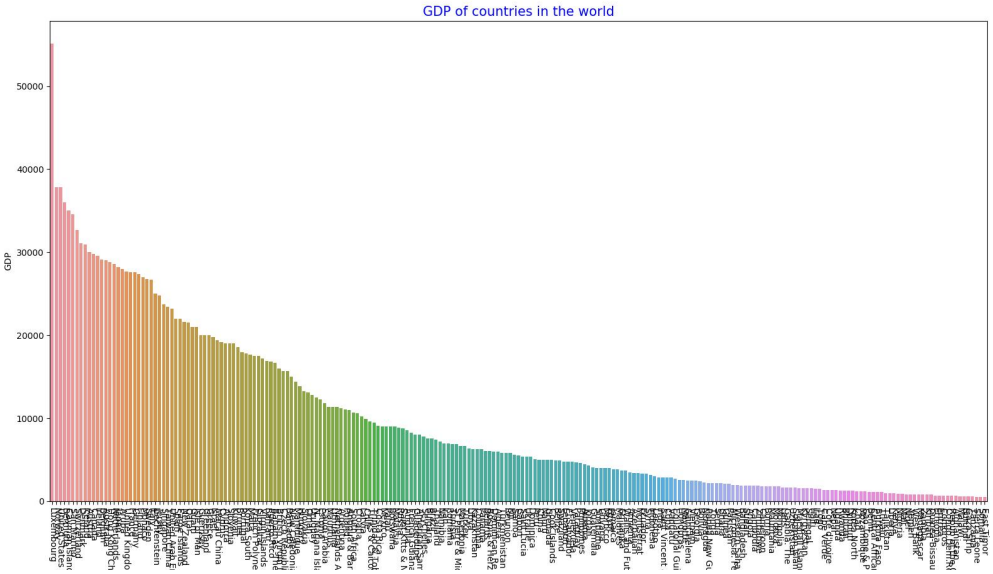

Figure 2.2 GDP of countries in the world

It can be seen from Figure 2.2 that the gap between GDP per capita is very large, and the gap between the highest-ranking Luxembourg on the left and the lowest East Timor on the right is more than ten times. When the United Nations certifies developed countries, per capita GDP is an important indicator to measure whether a country is developed. The huge gap between the countries shown in Figure 1 is also explained from the side, and some important indicators are closely related to the development of the country.

To this end, the author needs to select the attributes that are essential to the development of the country from the many attributes of the data set. Accurate data is very beneficial to any classification algorithm. In order to unearth the commonality between developed countries, the differences between the various attributes of developed countries are investigated. If a

certain attribute has little difference between developed countries, then this attribute is likely to be an important attribute that affects or reflects the level of national development. . Calculating the variance of each attribute between developed countries, and then through the line chart can clearly reflect which attributes are commonality between developed countries.

The author first removes all developing countries from the data set and makes a histogram of the attributes of developed countries, or taking per capita GDP as an example, as shown in Figure 2.3.

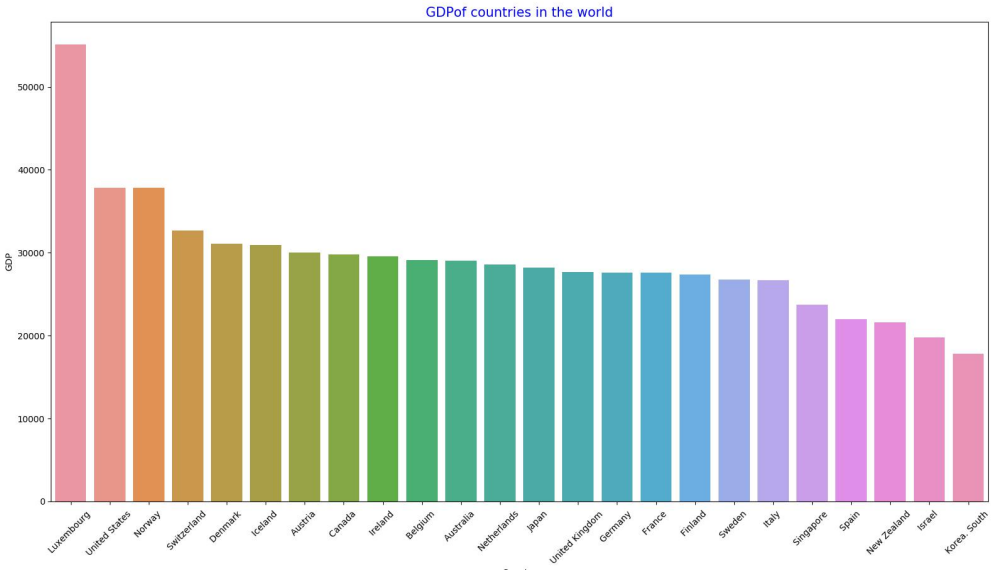

Figure 2.3 Bar plot of developed countries' GDP

Figure 2.3 shows the GDP per capita in developed countries. The highest on the left is still Luxembourg, and the lowest on the right is South Korea, and the range is narrowed to a relatively small extent. It can be clearly seen from the figure that the GDP per capita of Luxembourg is much higher than that of other developed countries. Can you think that GDP per capita is not an important indicator for measuring the level of national development? The answer is no. In order to eliminate the influence of these outlier data on the calculation of variance, it is necessary to find the outlier data of each attribute through the box diagram, so that the data is more representative. Figure 2.4 is a box plot of GDP per capita in developed countries. The data in the figure indicates which data is abnormal. According to the comments in the figure, these data can be easily removed from the dataset of developed countries.

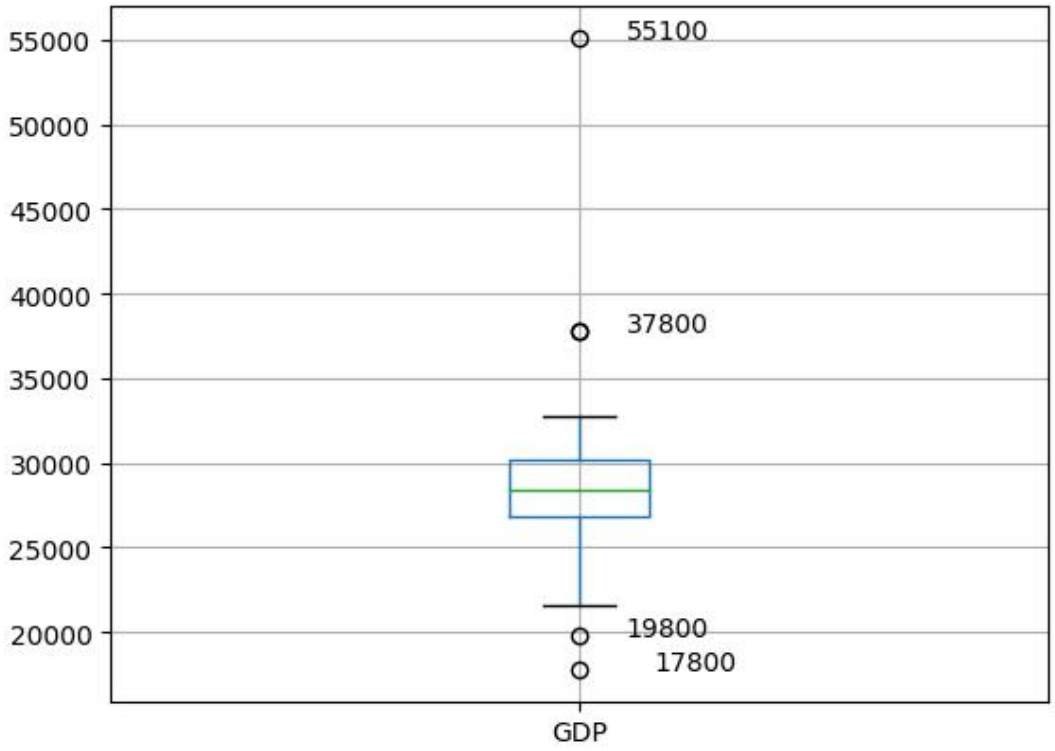

Figure 2.4 Box plot of developed countries' GDP

After finding the outliers in each attribute value through the box plot, you can start to calculate the variance of each attribute. The key code is shown in Figure 2.5. The flt in the code is a dictionary type, which contains the outliers of each attribute, and the outliers contained in the dictionary are eliminated when calculating the variance. The return type is the dictionary type, the key is the individual attribute, and the value is the variance of the attribute. In addition, in order to facilitate comparison of the variance of each attribute, the data is linearly normalized before the variance is calculated.

After calculating the variance of each attribute of developed countries, it is also necessary to determine which attributes are the attributes required by the training set according to the variance. Therefore, the author made a line chart with the variance of each attribute. As shown in Figure 2.6, the gap between the various attributes of developed countries can be clearly seen through the line chart. It can be seen from the figure that the

variance of "Coastline" and "Crops" is relatively large, so it can be inferred that the difference between the two attribute values of developed countries is large, and their effect on measuring the level of national development is small. In addition to the above two attributes, the variances of "PopDensity", "NetMigration", "Arable" and "Agriculture" are basically average, and it is difficult to speculate on their role in measuring national development levels.

```
def calculateVariance(dataset,flt = {}):
    Variance = dict()
    for feature in dataset.columns:
        print("process %s:" % feature)
        tmp = dataset.loc[:,[feature]]    #筛选列
        countries = list(tmp.index)
        if len(flt) != 0:
            countries = list(set(tmp.index).difference(set(flt[feature])))
            if len(countries) != len(dataset) - len(flt[feature]):
                print("Error:the outliers' cnt is wrong!")
                print("%d - %d != %d" % (len(dataset),len(flt[feature]),len(countries)))

        d = tmp.loc[countries]      #筛选行
        arr = np.asarray(list(d[feature]))
        m_sum = sum(arr)
        nor = []
        # print("     normalize:")
        for x in arr:
            x = x / m_sum
            # x = float(x - np.min(arr)) / (np.max(arr) - np.min(arr))
            # print(x)
            nor.append(x)
        Variance[feature] = np.asanyarray(nor).std()
        # Variance[feature] = arr.std()
        print("标准差:%f" % Variance[feature])
    return Variance
```

Figure 2.5 Calculate variance

The variances of "InfantMortality","GDP","Literacy","Birthrate","Deathrate","Industry" and "Service" are very small, so these seven attributes can be used as a basic feature subset and will be in The attribute of the variance average is added to the basic feature subset to form a new feature subset. Finally, all the attributes are used as a feature subset. The classification algorithm training data set is used for the three feature subsets respectively. After verifying the accuracy rate, the basis is obtained. The classification result of the feature subset is the most accurate, and the accuracy rate is 90%. Therefore, it is concluded that the attributes of

the basic feature subset are the most important indicators for measuring the level of national development.

According to the above conclusions,the training set and the prediction set are constructed. The training set is used to train the classification model, and the prediction set is used to predict the development potential of the country.

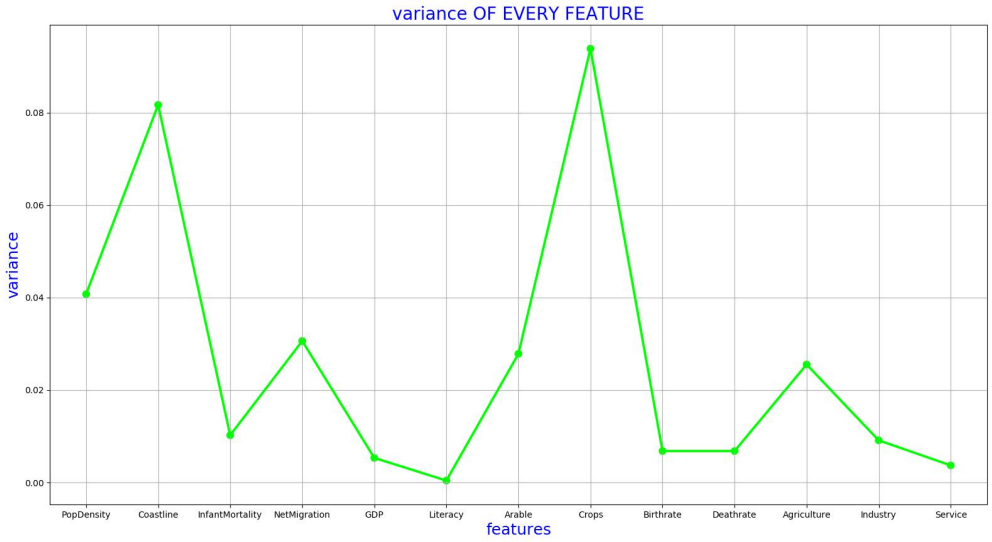

Figure 2.6 Variance of developed countries' feature

2.2 Model Training

The countries in the data set are divided into developed countries and developing countries,so the classification model can be used to train the data. In the classification method, the decision tree classification method is a simple and widely used classification method with good classification effect.

2.2.1 Decision tree maximum depth

From the perspective of the fitting effect, the maximum depth of the decision tree is an important parameter that affects the performance of the classifier. The author tested the performance of the classifier when the decision tree depth was between 3-10, as shown in Figure 2.7.

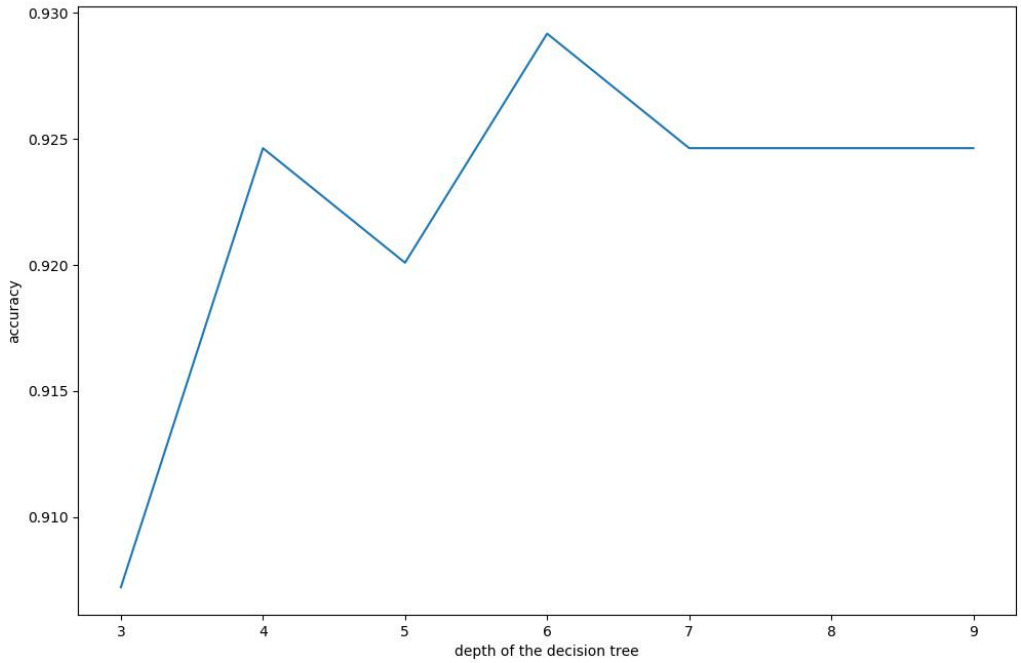

Figure 2.7 Accuracy on depth of the decisiion tree

It can be seen from Figure 2.7 that when the maximum depth of the decision tree is 6, the accuracy of the classifier is the highest. When the maximum depth of the decision tree is greater than or equal to 7, the performance of the classifier gradually becomes stable. Therefore, when the maximum depth of the decision tree is 6, the fitting effect is optimal.

2.2.3 Cross-validation

In order to test the effect of model classification, the author used the method of cross-checking, as shown in the code in Figure 2.8.

```
clf = tree.DecisionTreeClassifier(max_depth=max_depth)
cv_result = cross_val_score(clf,x,y,cv = n)
print("cv scores:",cv_result)
avg = np.sum(cv_result) / n
print("cv scores average",avg)
cv_avg.append(avg)
```

Figure 2.8 Trainning model

The results of the cross-check showed that the average prediction accuracy of the classifier reached 93%.

2.2.4 Prediction

For the prediction set, the author chooses the top countries in the attributes corresponding to the basic feature subsets in the developing countries to form the prediction set of this topic. After the classification model training is completed, the predict() interface is called, and the parameters are prediction sets. The result returned is whether the countries in the forecast set have reached the development level of developed countries, as shown in Figure 2.9.

As can be seen from the returned results，"Bermuda"，"Cayman Islands"，"San Marino"，"Aruba"， "Liechtenstein" and "Jersey" have reached the level of development in developed countries.

```
[0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0]
DevelopedCountries:
Bermuda
Cayman Islands
San Marino
Aruba
Liechtenstein
Jersey
```

Figure 2.9 Predict result

In fact, the per capita GDP of these countries is very high, as shown in Figure 2.10. The three data on the left side of the figure use the GDP of United States, United Kingdom, and Japan as a reference.

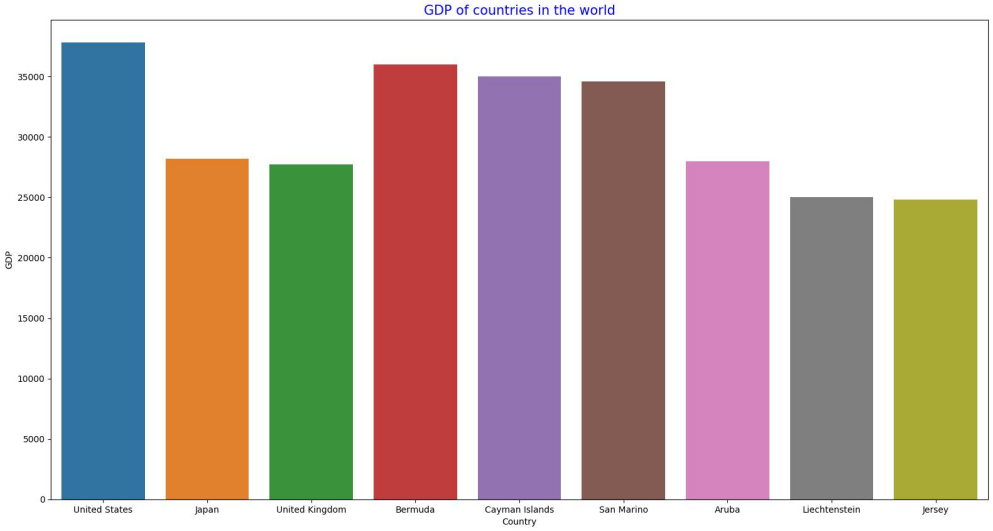

Figure 2.10 GDP of predicted countries

3 Result Analysis

3. 1 Decision Tree Analysis

In order to further verify the accuracy of the prediction results, the author deeply explored the classification strategy of the decision tree, as shown in Figure 3.1.

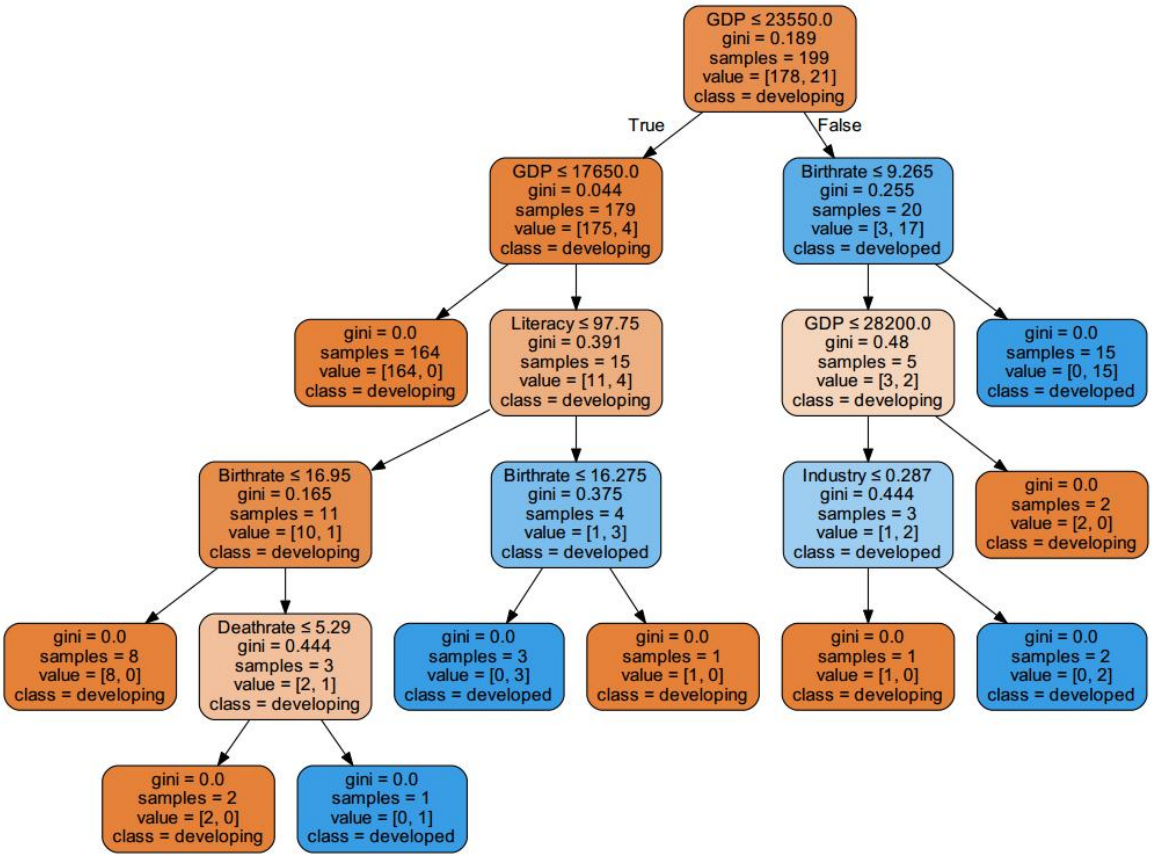

Figure 3.1 Decision Tree

Figure 3.1 shows the internal classification strategy of the decision tree model. It can be seen that per capita GDP is the primary feature used in the classification of the model, followed by birth rate, cultural rate and industrial level as the classification indicators of the decision tree. More importantly, These characteristics are important indicators for the evaluation of national development levels by international agencies, including the United Nations. In addition, the author notices that the "InfantMorality" and "Service" have no effect on the decision tree model. Although their variance is small, considering that their attribute

values are small in the original data, they are difficult to measure the level of national development and decision-making. It is also reasonable to delete these two attributes in the tree model. Therefore, the decision tree model shown in Figure 3.11 has high credibility.

3. 2 Confidence interval for estimating prediction accuracy

According to the verification result of 2.2.3, the accuracy of the classifier is 93%, that is, the empirical accuracy of the classification is acc = 0.93. The training data set used by the author has a total of 202 records, that is, N = 202, and the acc and N are substituted into the formula (3-1).

$$\frac{2 \times N \times acc + Z^2_{\alpha/2} \times \left(Z^2_{\alpha/2} + 4 \times N \times acc - 4 \times N \times acc^2\right)^{1/2}}{2 \times \left(N + Z^2_{\alpha/2}\right)} \tag{3-1}$$

Table 3-1 shows the value of Z at different confidence interval levels：

Table 3-1

1 - α	0.99	0.98	0.95	0.9	0.8	0.7	0.5
Z	2.58	2.33	1.96	1.65	1.28	1.04	0.67

In the 95% confidence level, what is the confidence interval of the true accuracy of the classification model, and Z is substituted into the formula (3-1), and the confidence interval is between 88.6% and 95.7%, so it can be concluded that the model classification The probability of accuracy between 88.6% and 95.7% is 95%。

3. 3 Conclusion

In this report, research on the development potential of some developing countries has led to the following conclusions:：

At 95% confidence level, you can think，"Bermuda"，"Cayman Islands"，"San Marino"，"Aruba"， "Liechtenstein" and "Jersey" have reached the level of development in developed countries.

Reference

[1] Pang-Ning Tan,Michael Steinbach,Vipin Kumar.Introduction to Data Mining[M].2011.1