Build a Rural Development Classification Model in Python with Scikit-learn

Arvind Kumar
Research Scholar, Department of Computer Science
Baba Mast Nath University, Rohtak,

Dr. Satpal
Professor, Department of Computer Science
Baba Mast Nath University, Rohtak,

ABSTRACT

Rural data mining concerns with developing methods for discovering knowledge from Rural datasets. Data Mining is the analysis step of the KDD, a process of extracting new patterns from large data sets involving methods from statistics, machine learning and artificial intelligence. This paper focuses on the building a Rural Development Classification Model in Python with Scikit-learn. In this work the classification Techniques of Data mining is used to efficiently analyse and predict rural development level and better manage rural economy. This technology can offer vast opportunities and capabilities in rural areas which will in turn help in effectively solving many problems associated with rural areas. Information on rural development is not only crucial but the maintenance and analysis of data is considered one of the main subjects in future decision making and improvement in rural life. Strengthening data collection and analysis provides perhaps the greatest opportunity for researchers and policy makers in rural areas. Utilization of such data is considered a driving force that improves the economic dynamism while creating a new type of knowledge-based economy. Many countries are adopting Data mining applications to Rural Development.

Keywords— Rural Development (RD), Classification, Clustering, Data Mining (DM); **Knowledge Discovery in Databases *(*KDD*);***

1. INTRODUCTION

The aim of this paper is to develop a system for data mining classification model in rural development. The system performs classification of data points given as input to this system. The system takes inputs, data points to be classified and number of classes to be made of these input data points. The data points are prepared into data preparation phase of data mining process before using this system. The raw data is pre-processed, normalized and then data points are classified using Random Forest Classification technique. The selection of the attributes is done carefully. As an example education, standard of living, agriculture and livestock, and economic status are used as input attributes to make the data point which is given Classification algorithm as input. Classification of these attributes is successfully performed. In this way this paper will show the applicability of data mining classification tools and techniques in rural development in next forth coming sections.

Rural development: Rural development is conceived as strategy aimed at finding ways to improve the rural lives with participation of the rural people themselves so as to meet the required need of the rural area. According to World Bank (2005), rural development is the process of rural modernization and the monetization of the rural society leading to its transition from traditional isolation to integration with the national economy. Also, rural development is perceived as a process of not only increasing the level of per capital income in the rural areas but also the standard of living of the rural population measured by food and nutrition level, health education, housing, recreation and security. Haryana is a leading contributor to the country's production of food grain and milk. This classification model will assist in minimizing the rural-urban gap in terms of basic infrastructure facilities essential for sustainable rural development [4]. This classification model will also assist in the planning process.

Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information. Technically, Data mining as a term used for the specific classes of six activities or tasks as a) Classification b) Estimation c) Prediction d) Association rules e) Clustering. Out of these six activities here in this work the focus is given on classification.

Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under [1].

Few of the terminologies encountered in classification are:

Classifier: An algorithm that maps the input data to a specific category.

Classification model: A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.

Feature: A feature is an individual measurable property of a phenomenon being observed.

Binary Classification: Classification task with two possible outcomes. eg: Gender classification (Male / Female)

Multi class classification: Classification with more than two classes. In multi class classification each sample is assigned to one and only one target label. Eg: An animal can be cat or dog but not both at the same time

Multi label classification: Classification task where each sample is mapped to a set of target labels (more than one class). Eg: A news article can be about sports, a person, and location at the same time.

In this work random forest machine learning algorithm is used for classification and regression problem. Random forest applies the technique of bagging (bootstrap aggregating) to decision tree learners. Random forest is a good option for regression and best known for its performance in classification problems. Furthermore, it is a relatively easy model to build and doesn't require much hyper-parameter tuning. This is because the main hyper-parameters are the number of trees in the forest and the number of features to split at each leaf node. Random forest has numerous business use cases for classification and regression problems. Some of the business use cases are a) Fraud prediction b) Cancer detection c) Stock market predictions d) Spam filter e) News classification etc. [2].

2 PREREQUISITES

Here implementation of Random Forest Classification Model is performed using a well known programming language Python 3.7. Following table shows the hardware and software requirements for this work.

Requirements	Name	Version
Programming language	Python	3.7
Operating System	Windows 7	18.04.1 LTS
Library Functions	Pandas	0.23.4
	Numpy	1.15.3
	Scipy	1.1.0
	Scikit-Learn	0.20.0
	Matplotlib	3.0.1
	Seaborn	0.9.0
	Plotly	3.4.1rc1
Python IDE	Jupyter	3.3.1

Table 2.1

3. DATASET AND EXPLANATORY VARIABLES

The present study has been carried out in the Kaithal district. It is one of the 26 districts of Haryana, state in northern India. Kaithal, the north eastern district of Haryana State with a total geographical area of 2317 sq. km (approx 228000 hac). Kaithal was previously a part of Karnal District and later, Kurukshetra District. Kaithal came in to existence as a district of Haryana in 1 November 1989. The boundaries of Kaithal district are touching the three district of Haryana namely Karnal, Jind and Kurukshetra and the Patiala district of Punjab. The district is under control of Ambala division. The Kaithal city, occupies an area of 43.76 sq. km within the municipal council.

It is having a total population of 1072861 as per 2011 census. Out of this population 78 percent people live in the rural areas. Sex ratio in the district is 880 females per 1000 males as compared to 877 in the state and the population density is 463 per sq.km as compared to 573 of the state. The literacy rate is 76.4 percent as compared to 76.6 of the state. The district is having six blocks and 263 villages. There are seven blocks in Kaithal district namely Kaithal, Kalayat, Rajound, Pundari, Dhand, Guhla and Siwan that has been selected for the present study. Ten percent of villages from each block have randomly been selected for the study. In this district rural poor are primarily reliant on agriculture and animal husbandry. 4% of households are selected from each randomly selected village. The distribution of the sample, List of Blocks, List of villages in each block is given in appendix.[5]

Thus the ultimate dataset consists of 4521 samples from 28 randomly selected villages. The data has been collected from primary as well as secondary sources. Secondary data has been taken up because of the easy availability. Primary data has been collected from the borrowers in the field by personnel interview method through an interview schedule designed for the purpose. To collect the primary data we have visited the villages of the study area and has interacted with the peoples extensively by putting the questions in the local language. Secondary data has been collected from various sources including Department of Economic and Statistical Analysis, Haryana and District Rural Development Agency.

4. STEPS INVOLVED IN BUILDING A CLASSIFICATION MODEL

The following are the steps involved in building a classification model:

Step 1: Load Python Packages

Let's begin by installing the Python module Scikit-learn, one of the best and most documented machine learning libaries for Python. To begin implementation, let's activate Python 3.7 programming environment. Make sure you're in the directory where your environment is located, and run the following command:-

my_env/bin/activate

With our programming environment activated, check to see if the Scikit-learn module is already installed:-

python -c "import sklearn"

If sklearn is installed, this command will complete with no error otherwise install it.

```
#import libraries
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
#import plotly.graph_objs as go
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.metrics import precision_recall_curve
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report
```

Step 2 Import and Pre-Process The Dataset

Now we have sklearn imported in jupyter notebook, we can begin working with the dataset for our machine learning model. The dataset we are working with in this work is the rural development dataset. The dataset includes various information about rural peoples, as well as classification labels. The dataset has 4521 *instances*, or data, from 28 randomly selected villages. and includes information on 30 *attributes*, or features. Let's load the dataset and begin exploring these parameters.

```
#Read the data set

data = pd.read_excel('C:\\Users\\Dell\\Desktop\\coding\\Source_code1\\Source_code\\Data_set\\classification.xlsx','Sheet1')

df = pd.DataFrame(data)

corr = df.corr(method='pearson')
```

Before we build the model, we need to make some changes to the data in order to make it ready for the model. Let's begin with lowercasing and one-hot encoding the categorical variables so that we can turn the categorical variables to numeric.

Step3 Split the data into train and test sets / Organizing data into sets

We will build the model on the training set and check the accuracy of the model by using it on the testing set. So to evaluate how well a classifier is performing, we test the model on unseen data. Therefore, before building a model, split your data into two parts: a *training set* and a *test set*. Fortunately, Scikit-learn has a function called train_test_split(), which will divide data into two data sets. Import this function and then use it to split the data:

```
#splitting the data

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

print("Training data set\n")

train = pd.concat([X_train, y_train], axis = 1)

print(train.head(25))

print("\n")

print("Test data set\n")

test = pd.concat([X_test], axis = 1)

print(test.head(25))
```

Step 4 Build A Random Forest Classifier and Predict

There are many models for machine learning, and each model has its own strengths and weaknesses. In this work, we will focus on a simple machine learning algorithm that usually performs well in classification tasks, namely Random forest algorithm. First, import the RandomForestClassifier module. Then initialize the model with the RandomForestClassifier() function, then train the model by fitting it to the data using clf.fit(). After we train the model, we can then use the trained model to make predictions on our test set, which we do using the predict() function. The predict() function returns an array of predictions for each data instance in the test set. We can then print our predictions to get a sense of what the model determined.

```
clf = RandomForestClassifier(n_estimators=10000, random_state=0, n_jobs=-1)

# Train the classifier

clf.fit(X_train, y_train)

#pre = clf.predict([[7,3,3,3,3,3]])

#print(pre)

y_pred = clf.predict(X_test)

#classification report

a= (metrics.accuracy_score(y_test, y_pred))

cr = (classification_report(y_test, y_pred))
```

Step 5 Evaluating the Model's Accuracy

Using the array of true class labels, we can evaluate the accuracy of model's predicted values by comparing the two arrays (test_labels vs. preds). We will use the sklearn function accuracy_score() to determine the accuracy of our machine learning classifier.

```
#classification report
a= (metrics.accuracy_score(y_test, y_pred))
cr = (classification_report(y_test, y_pred))
#confusion matrix
cm = (metrics.confusion_matrix(y_test, y_pred))
print("Confusion Matrix:\n\n",cm,"\n")
#print(type(cm))
cm1 = list(metrics.confusion_matrix(y_test, y_pred))
#print(type(cm1))
print(classification_report(y_test, y_pred))
print("Precision:\n")
#Accuracy score
print("Accuracy score:",round(a,1))
```

Now, to check the accuracy of the model, we will check how the predictions stack up against the actual test set values. A confusion matrix is one of the methods used to check the accuracy of a classification model.

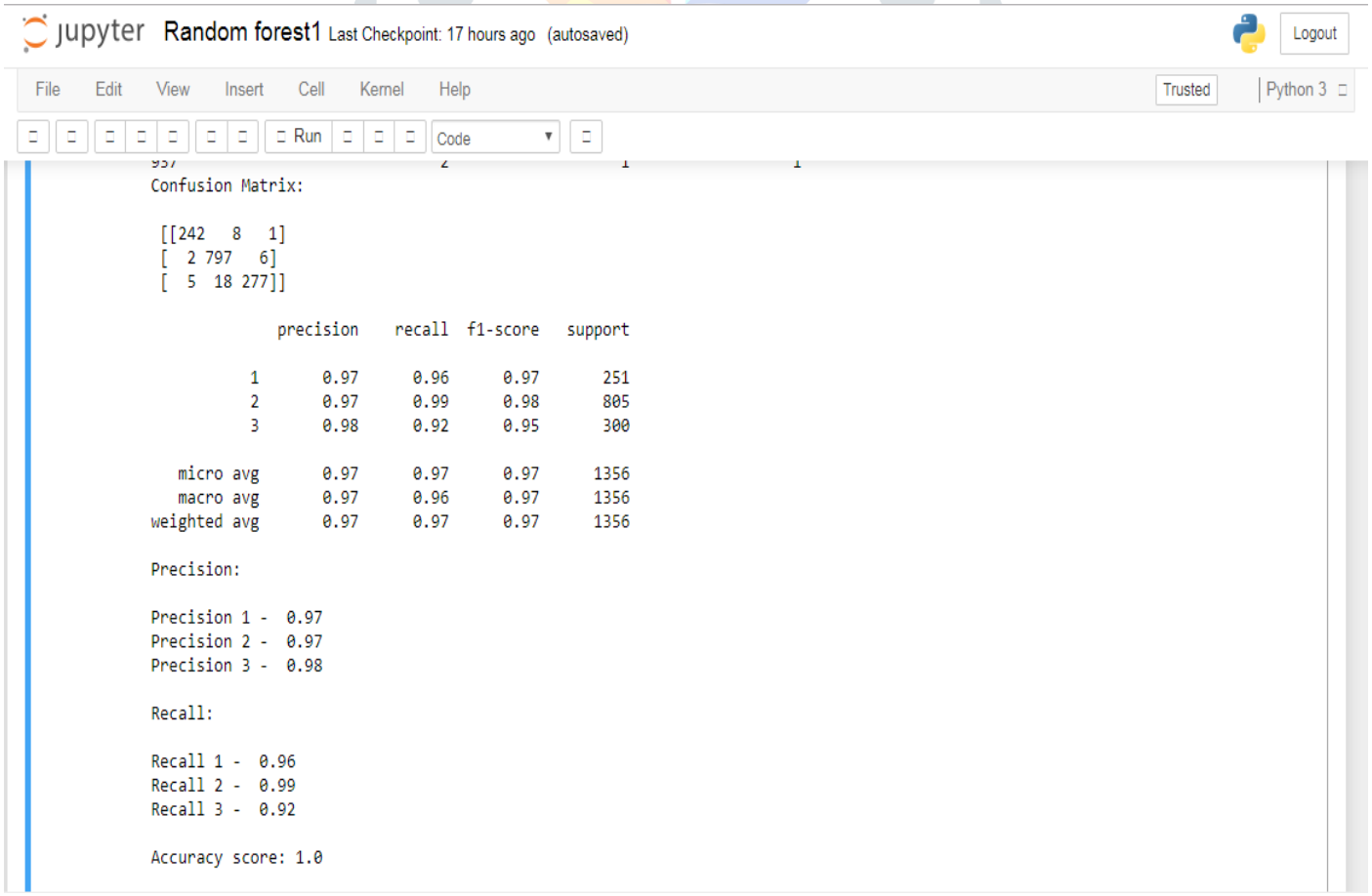

Figure 1

As we can see the output in above figure, the model did pretty well! As shown in the output, the Random Forest Classifier accuracy score is 1.0. This means that 100 percent of the time the classifier is able to make the correct prediction. These results suggest that our feature set of 30 attributes are good indicators of rural development class. In this way this work successfully built machine learning classification model.

5. CONCLUSIONS

It is clear from this work that random forest classifier is an efficient classifier and easy to build. It can produce quite impressive results. The data used in this random forest model classification is a large dataset and but does not require too much pre-processing. Readers should be cautious in interpreting the results, since the findings are based on data collected from the survey. As previously stated, strengthening data collection and analysis is a fundamental necessity for rural areas policymakers such that efficient utilization of the resulting opportunities will entail society's economic dynamism while creating a knowledge-based economy. Data collection and adequate data analysis has many applications in various areas including market management, agricultural products, immigration management, rural products distribution methods etc. Such approach is employed because data mining applications in rural development are still rarely described in journal articles. However, I feel that even such a survey can describe the current state in data mining applications in rural development. The most often used methods are classification and prediction, concept/class description and evolution analysis. It can be concluded that data mining methods and other related techniques of knowledge discovery in databases and intelligent data analysis are indispensable in rural development.

6. REFERENCES

[1] https://www.analyticsindiamag.com/7-types-classification-algorithms/
[2] https://www.datascience.com/blog/classification-random-forests-in-python
[3] https://www.digitalocean.com/community/tutorials/how-to-build-a-machine-learning-classifier-in-python-with-scikit-learn
[4] Kanu Raheja, Rural Development in Haryana, International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015 1 ISSN 2250-3153.
[5] Arvind Kumar, Dr. Satpal, " A Study of Application of Data Mining in Rural Development of Kaithal Region" © 2019 JETIR June 2019, Volume 6, Issue 6 www.jetir.org (ISSN-2349-5162)
[6] Abolfazl Shahbazi, Maryam Karambeygi, Application of data mining in Rural Planning, ISSN: 2277-3754 ISO 9001:2008 Certified International Journal of Engineering and Innovative Technology (IJEIT) Volume 5, Issue 1, July 2015
[7] Kush R. Varshney,1,2, George H. Chen,3 Brian Abelson,1,4 Kendall Nowocin,3 Vivek Sakhrani,5 Ling Xu,6 and Brian L. Spatocco7 ,Targeting Villages for Rural Development Using Satellite Image Analysis, Big Data Volume 3 Number 1, 2015 Mary Ann Liebert, Inc. DOI: 10.1089/big.2014.0061.
com/wpcontent/uploads/crisp-dm-4-problems-fig1.png