CERTIFICATION

This is to certify that the project entitled "**Bank direct marketing prediction**", submitted to "**Department of Electronics and Communication, University of Allahabad**", embodies the record of study carried under the supervision of **Dr. Richa Mishra**. This project is a part of fulfillment criteria for the award of the degree of science (B. Sc.). It is an original work completed by **VARTIKA JAISWAL**.

I certify that this work is an original work and is not copied from or submitted in any other institution.

Signature of Student	**Signature of Supervisor**
(Vartika Jaiswal)	(Dr. Richa Mishra)

ACKNOWLEDGEMENT

With immense please, I would like to thank my project supervisor, **Dr. RICHA MISHRA** for her guidance and unending support. I express my gratitude to my supervisor for providing me an opportunity to embark on this project entitled “Bank direct marketing prediction”, which helped me out to analyze various concepts and go in deep with the related topics.

Also, I would like to sincerely thank to my classmate Hemant Singh for his help and accordance in the completion of this project. Above all, I thank to each and everyone who helped me to complete this project within time. I would sincerely carry out with this project in my further studies.

DECLARATION

I, **VARTIKA JAISWAL**, hereby declare that this project work entitled “Bank direct marketing prediction”, submitted to “Department of Electronics and Communication, University of Allahabad” is an original work and is not copied from somewhere else. This study carried under the supervision of **Dr. RICHA MSHRA.**

The information and data given in the report is authentic to the best of my knowledge. This project is not submitted in any other institution for the commencement of any degree as per my knowledge.

ABSTRACT

With the launching of various products and sales in banking sector, it becomes a tough task for the banks and institutions to review sales results, measure the effectiveness of sales, advertising tactics and easily make adjustments to the campaigns in the favor of launches. Direct marketing enables banks and other financial organizations to concentrate on those customers who present the possibility of subscribing to their products, offers and other packages. This work proposes machine learning approach to predict the success of telemarketing calls for selling the bank's long term deposits (success in the subscription of deposits due to campaign). Extracting most benefits out of a dataset is a challenging task as it requires the evaluation of its different attributes and corresponding values. This job is usually done by presenting data in a visual format to reveal hidden patterns. Here, a Portuguese retail bank dataset is assessed with data collected from 2008 to 2013, thus including the effect of recent financial crisis. Work introduces analysis and applications of various machine learning techniques such as logistic regression (LR), k Nearest Neighbor (kNN), Gaussian NB, support vector machine (SVM), etc.

The aim of this study is to predict whether a client is going to subscribe a long-term deposit or not, by examining the performance of logistic regression, decision tree and other techniques. Different plots and metrics are examined on all the afore mentioned techniques. We then compared all the models i.e. LR, kNN, Gaussian NB, SVM on the basis of accuracy report. Accuracy report portrays that LDA with an accuracy of 85.80% is most prolific classifier in terms of predictive ability and then LR with accuracy of 84.82% follows through the result with others with lower accuracy percentage. Both the classifiers are checked on the original dataset. Stepping forward a new dataset with some removed attributes gives better accuracy score. LDA, LR and SVM now give an accuracy score same as the old dataset and two other classifiers (kNN and Gaussian NB) give noticeable increase in accuracy.

INDEX

LIST OF TABLES

LIST OF FIGURES

1. Introduction

It has become relevant for corporate organizations to explore online and direct marketing as part of marketing strategy in order to increase sales, enhance business and keep a tab on customers. Marketing campaigns generally constitute a technique of outsourcing by organizations with the expectation of improving the financial outcome of their business and also having a competitive advantage over their peers. Quantitative analysis for forecasting in business and marketing especially in consumer behavior and consumer choice model has become more prominent in business practices. The ability to understand and accurately predict a consumer decision can lead to more effectively targeting products, cost effectiveness in marketing strategies, increasing sales and result in substantial improvement in the overall profitability of the firm. Organizations utilize direct marketing when focusing on fragments of clients by reaching them to meet a particular objective. Organizations utilize various methodologies to contact and get idea about client's interest. Follow on different communication styles and manipulative aspects to attract the users to get them devoted with the products of organization. Reaching out to client through remote communication centers facilitates operational administration of campaigns. Different call centers are taken under operation for carrying out client interaction task. Such call centers permit speaking with clients through their contact details by different mediums for instance by using phones, telecommunication services, generally it is mobile. The advertisement of a product undertaken via contact center is called telemarketing because its nature of remoteness. Contact can be divided into inbound and outbound depending on which side triggered the contact (client or contact center), with each case posing different challenges. There are two methodologies by which corporate institutions promote their products through mass crusades, which considers overall population and direct campaign, which targets a limited or just a particular group of individuals. Formal review as discussed in other papers depict that there is low efficiency of mass campaigning, typically very low percentage of entire populace will have a positive response towards the designation of campaign. While looking on the aspect of direct campaign, it concentrates just on a little group of individuals who are thought to be attracted and show interest in the product being advertised and in this manner is substantially be more productive to engage. Technology enables rethinking marketing by concentrating on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing us to build longer and tighter relations in alignment with business demand. The task of selecting the prospective customers i.e. that are more likely to subscribe to the product pose a dire challenge of classification in machine learning which builds by matching and comparing customer attributes e.g. age, marital status, educational level, etc. and other characteristics (e.g. loan request, repayment etc.) to different outputs. The goal is to build a data driven model that learns an unknown underlying function that maps several inputs variables which characterize an item. Here, bank client, with the labeled output target (e.g. success or failure of deposit sell).Given a Portuguese retail bank dataset to carry out different classification task to come up with an effective predictive model. There are several classification models, here it is the various machine learning algorithms which are taken into consideration. The algorithms used are LR, Gaussian NB, Linear Discriminant analysis (LDA), KNN and SVM. Logistic regression and decision tree have an advantage of fitting models that tend to be easily understood by humans, while also provide good prediction in classification tasks. NN and SVM have no priori restriction imposed and are more flexible when compared with LR. Presenting learning capabilities that range from linear to complex nonlinear mappings. Due to such flexibility, NN and SVM provide accurate prediction, but the obtained model is not generic for humans to understand. Difference in the performance and result depict the impact of problem context and the ways involved, giving a reason to test several other techniques when addressing a problem before making a choice between the solutions. Main contribution to this work is as follows, firstly we have tested all the afore mentioned algorithms on the Portuguese bank direct marketing dataset from the UCI website. The experimental results reveal the set of features improves the accuracy of model. Further chapters discussed the related work, dataset description, proposed methodology and concluded with identifying the relevant features. We then compared several algorithms on the dataset by calculating its accuracy score and concluded the result. The result was such that the highest accuracy lied between 84% to 86%. Secondly, the project has emphasized on whether the selection features have an impact on the models.

2. Related Works

Digitalization in the era has evoked a lot of work in the preferred section. In the field of information technology, the ability and accurate analysis of a large set of data is essential. Considering the aspects many researchers have paid attention towards the subfield of artificial intelligence, i.e. machine learning. Nowadays, the datasets are easily available and some potentially useful information can be extracted to attain the objective for decision making and prediction.

With respect to marketing campaign, several studies have been delved to enhance the telemarketing success, machine learning helps to construct various models to solve the problems about bank direct telemarketing. Elsalamony [1] used four statistical measures: classification, accuracy, sensitivity and specificity on the bank dataset (17 features and 45,211 instances). He compared and evaluated the classification performance of four different data mining models: Multilayer perception neural network (MLPNN), Tree Augmented Naïve Bayes (TAN), LR and C5.0 Decision tree classifier. He reported that C5.0 model achieved slightly better performance than the MLPNN, LR and TAN. Prusty [8] applied the Naïve Bayes and decision tree algorithm to the dataset. Main thrust of Prusty's study was to compare the results obtained when the data was unbalanced with results obtained using some classification algorithms when the data was balanced with equal selections of "yes" and "no" in the response class. Results showed that the area under curve value improved after balancing the response class.

In one of the other works, researchers undertook a research with the objective of targeting customers through direct marketing to outsource the long-term deposits. LR, DT, Neural network (NN), SVM were employed. Amongst these NN obtained the best result with AUC of 0.8 with 81% prediction accuracy. In a similar study, LR and DT were applied to a dataset regarding an Iran bank. Moro et al. [7] proposed a datamining approach to analyze the probability of success. For extracting the key information, feature selection was emphasized by employing neural network. The study aimed to extract additional knowledge, according to the benefit from past contacts history. By using neural networks, they found that the presence of the historical data could improve the accuracy for targeting the deposit subscribers. Other details are that by using naïve Bayes, decision tree and SVM classifier, Moro et al. [7] used 29 features and 45,211 instances to model the bank data. Results from analysis after application of classification algorithms revealed that call duration is most important feature, followed by month of contact. They found that SVM to be most reliable predictive algorithm with an AUC value of 0. 938. Many other works are also done in this field. Portuguese bank dataset is considered in this project and trying to extract the maximum value of accuracy with minimum use of attributes.

3. Proposed Methodology

3.1 Algorithms Description

Machine learning algorithms are classified as supervised learning algorithms and unsupervised learning algorithms. Here in this project, supervised learning algorithms are used for prediction purpose. The used methodologies are:

Logistic Regression (LR)

LR discovers association between nominal dependent variables and one or more continuous explanatory variables. In this, initial values of the predicted parameters are used and the probability of the sample coming from a population with those parameters is calculated. Transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more classes. The sigmoid function is as follows:

$$S(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Maps any real value into another value between 0 and 1.in machine learning we use sigmoid to map predictions to probabilities.

Gaussian Naïve Bayes classifier algorithm (Gaussian NB)

This algorithm is based on Bayes theorem, which describes the probability of an event, based on prior knowledge of conditions. Equation (2) represents Bayes formula,

$$P(A|B) = (P(B|A)P(A)) \,/\, P(B) \tag{2}$$

Where, P(A) is the probability of occurring an event A, P(B) is the probability of occurring of an event B, P(A|B) is the probability of occurring A when A is occurred, and P(B|A) is the probability of occurring B when A is occurred. It assumes all the features are interdependent to each other. Even if the features depend on each other or upon the existence of the other features. A Gaussian naïve bayes algorithm is special type of naïve bayes. It is specifically used when the features have continuous values.it is assumed that all the features are following a Gaussian distribution i.e. normal distribution.

Linear Discriminant Analysis (LDA)

It is a dimensionality reduction technique which is used for supervised classification problems. Used for modeling differences in groups i.e. separating two or more classes.it is used to project the features in higher dimension space into a lower dimension space by creating new axis and projects data onto a new axis in a way to maximize the separation of two categories. Two criteria used by LDA to create a new axis they are maximize the distance between means of the two classes and minimize the variation within each class.

K nearest neighbor's classifier (KNN)

KNN algorithm uses feature similarity to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training dataset. For implementing this algorithm we need to choose the value of K i.e. the nearest datapoints. Then for each point distance between test data and each row of training data is calculated and based on it, points are sorted. Next it chooses the top K rows from sorted array and then assigns a class to test point based on most frequent class of these rows.

Support vector machine(SVM)

It works by classifying the data in different classes by finding a line (hyperplane) which separates the training dataset into classes. There are many such linear hyperplanes, SVM algorithm helps to maximize the distance between the various class that are involved and this is referred as margin maximization.

3.2 Dataset Description

The bank direct marketing dataset employed here is downloaded from the University of California at Irvine (UCI)Machine Learning Repository. The dataset used was collected by S. Moro, R. Laureano and P. Cortez. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The characteristics of data set composed of two kinds: nominal and numerical attributes as described in the Table1.1. This table shows that three kinds of attributes: Numerical, Categorical, and Binary. Numerical which are in the range type for all of them like (Age, Balance, Day, Duration, Campaign, Pdays and previous), Categorical are in the set type as the attributes (Job, Marital, Education, Contact, Month, Poutcome), and Binary categories are all the attributes that represented as yes or no in their classes: the attributes(Default,housing,Loan,Output i.e.y).

The table contains column headed attributes, kind and attribute description. The column headed Attributes description is presenting the number classes for each attribute and the relation with its name. The first attribute age gives the numerical value of the age of customers. The second attribute named Job. There exists many kinds of job belonging to this, such as admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, and services. The attribute Marital can be illustrated in the classes as (married, divorced, and single) where the divorce means divorced and widowed. The education classes are divided into unknown, secondary, primary, and tertiary and followed attributes Default, Housing, Loan, and output attribute (y) has only two values (i.e. yes or no). The contact communication in column attribute are unknown, telephone and cellular. The attribute duration presents the description of last contact duration, in seconds (numeric), this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Month attribute has 'Jan','Feb', etc. The attributes Poutcome presents the outcome of the previous marketing campaign like failure, success and non-existent, attribute campaign consists number of contacts performed during this campaign and for this client following pdays gives number of days that passed by after the client was last contacted from a previous campaign. And lastly previous illustrates number of contacts performed before this campaign and for this client.

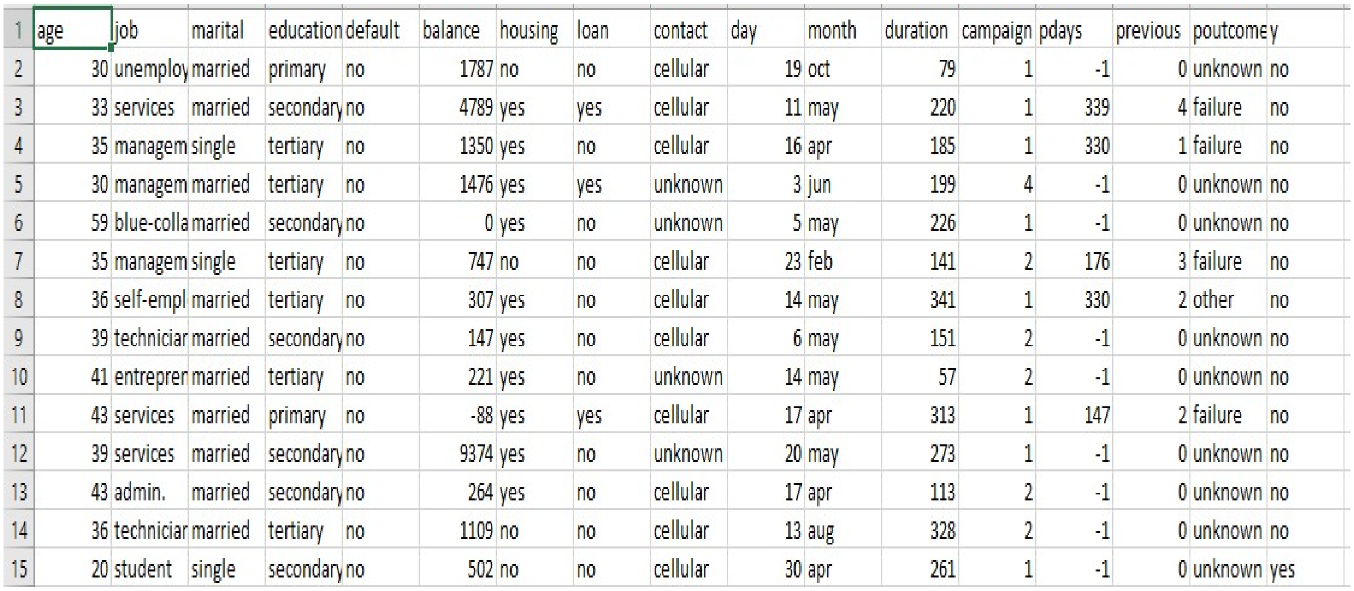
Datasets for prediction can be categorized in two types supervised and unsupervised learning. The supervised learning requires input data that contains both independent and dependent attributes, while in unsupervised learning all attributes are treated equally that is, there is not dependent variable, the prediction is made by searching out for patterns, groupings which may lead data relation understanding. Here the bank direct marketing dataset is supervised algorithm having some independent attributes and one output dependent attribute i.e. y for prediction.

Table 3.1 Dataset description of original dataset

Sr.No.	Attributes	Kind	Attribute Description
1.	Age	Numeric	NaN
2.	Job	Categorical	'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
3.	Marital	Categorical	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4.	Education	Categorical	Unknown, secondary, primary, tertiary
5.	Default	Binary(categorical)	has credit in default? 'no','yes','unknown')
6.	Balance	Numeric	Average yearly balance, in euros
7.	Housing	Binary(categorical)	has housing loan? ('no','yes','unknown')
8.	Loan	Binary(categorical)	has personal loan? (‘no','yes','unknown')
9.	Contact	Categorical	contact communication type ('cellular’, ‘telephone')
10.	Day	Numeric	Last contact day of the month
11.	Month	Categorical	last contact month of year (categorical: 'Jan', 'feb', 'mar', ..., 'nov', 'Dec')
12.	Duration	Numeric	last contact duration, in seconds
13.	Campaign	Numeric	number of contacts performed during this campaign and for this client (includes last contact)
14.	Pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign
15.	Previous	Numeric	number of contacts performed before this campaign and for this client
16.	Poutcome	Categorical	outcome of the previous marketing campaign ('failure','nonexistent','success')
17.	Y	Binary(categorical)	has the client subscribed a term deposit? ('yes’, ‘no')

3.2.1 PRE- PROCESSING OF DATASET

Bank marketing dataset downloaded from UCI has the following format shown in figure 3.1.



1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
2	30	unemploy	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
3	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
4	35	managem	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
5	30	managem	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
6	59	blue-colla	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
7	35	managem	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
8	36	self-empl	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
9	39	techniciar	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
10	41	entreprer	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
11	43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
12	39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
13	43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
14	36	techniciar	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
15	20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes

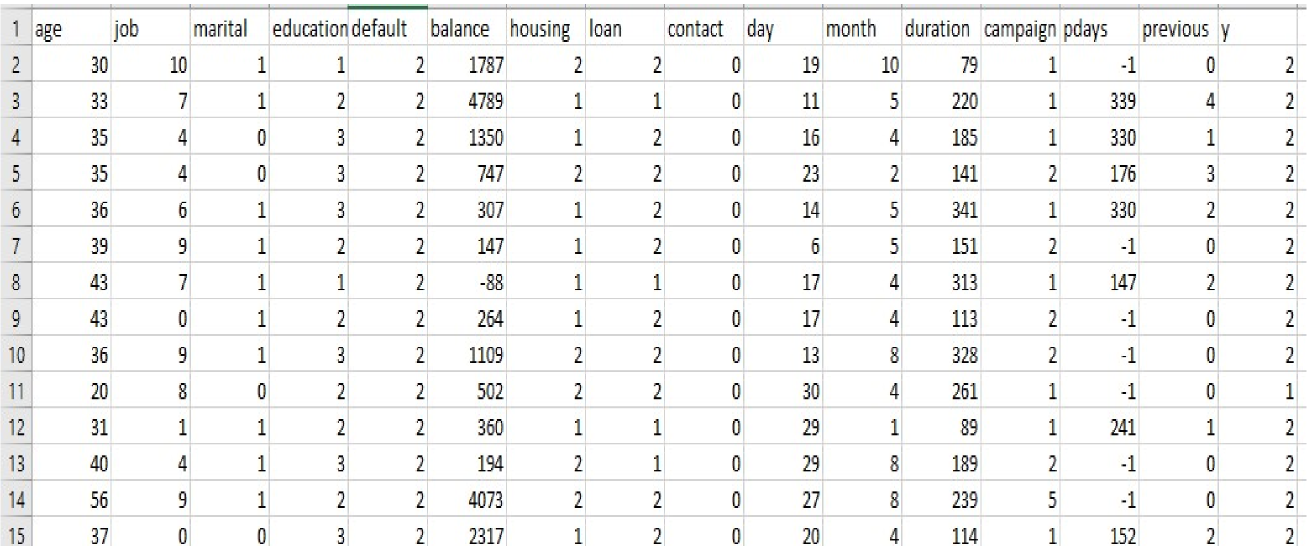
Figure 3.1 Bank marketing dataset

The dataset used in this work is first processed by converting some classes to int, we have dropped a column Poutcome, deleted the rows which had missing values. The attributes which has values converted to int are as detailed in the table below describing the character which is converted with the respective values. The final labeling of attribute values has shown in table 3.2. The sample dataset(dataset 1) format after preprocessing step is shown in table 3.3.

The work is performed on this dataset, firstly. The machine learning algorithms are tested on the dataset1 and the observation is noted.

Table 3.2 Labeling of attribute values

Attributes	**Classes in attributes**	**Values assigned**
Marital	Single	0
	Married	1
	Divorced	3
Job	Admin	0
	Blue-collar	1
	Entrepreneur	2
	Housemaid	3
	Management	4
	Retired	5
	Self-employed	6
	Services	7
	Student	8
	Technician	9
	Unemployed	10
Contact	Cellular	0
	Telephone	1
Month	Jan-Dec	1-12
Education	Primary	1
	Secondary	2
	Tertiary	3
Loan, Housing	Yes	1
	No	2

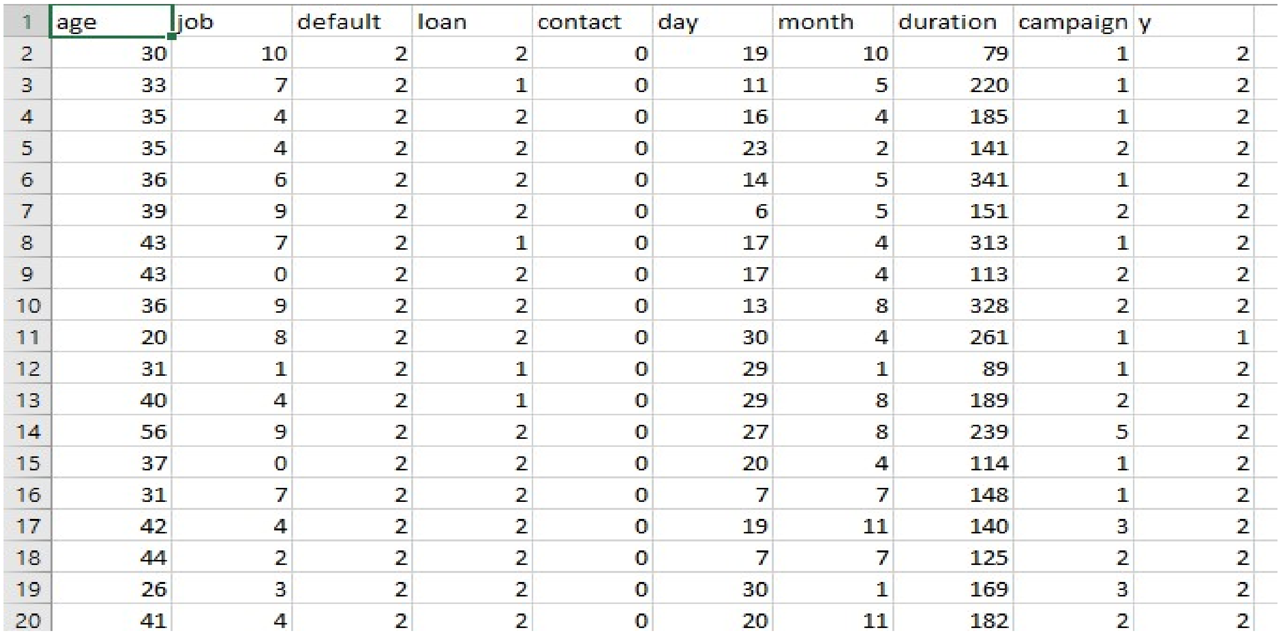


1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	y
2	30	10	1	1	2	1787	2	2	0	19	10	79	1	-1	0	2
3	33	7	1	2	2	4789	1	1	0	11	5	220	1	339	4	2
4	35	4	0	3	2	1350	1	2	0	16	4	185	1	330	1	2
5	35	4	0	3	2	747	2	2	0	23	2	141	2	176	3	2
6	36	6	1	3	2	307	1	2	0	14	5	341	1	330	2	2
7	39	9	1	2	2	147	1	2	0	6	5	151	2	-1	0	2
8	43	7	1	1	2	-88	1	1	0	17	4	313	1	147	2	2
9	43	0	1	2	2	264	1	2	0	17	4	113	2	-1	0	2
10	36	9	1	3	2	1109	2	2	0	13	8	328	2	-1	0	2
11	20	8	0	2	2	502	2	2	0	30	4	261	1	-1	0	1
12	31	1	1	2	2	360	1	1	0	29	1	89	1	241	1	2
13	40	4	1	3	2	194	2	1	0	29	8	189	2	-1	0	2
14	56	9	1	2	2	4073	2	2	0	27	8	239	5	-1	0	2
15	37	0	0	3	2	2317	1	2	0	20	4	114	1	152	2	2

Figure 3.2 Dataset 1 sample

Again, modifying the dataset1 to check whether the results are getting improved or not, some of the attributes of above dataset is dropped. The dropped attributes are marital, education , balance, housing and pdays.

The dataset then obtained is shown in figure 3.3, named as dataset 2:



1	age	job	default	loan	contact	day	month	duration	campaign	y
2	30	10	2	2	0	19	10	79	1	2
3	33	7	2	1	0	11	5	220	1	2
4	35	4	2	2	0	16	4	185	1	2
5	35	4	2	2	0	23	2	141	2	2
6	36	6	2	2	0	14	5	341	1	2
7	39	9	2	2	0	6	5	151	2	2
8	43	7	2	1	0	17	4	313	1	2
9	43	0	2	2	0	17	4	113	2	2
10	36	9	2	2	0	13	8	328	2	2
11	20	8	2	2	0	30	4	261	1	1
12	31	1	2	1	0	29	1	89	1	2
13	40	4	2	1	0	29	8	189	2	2
14	56	9	2	2	0	27	8	239	5	2
15	37	0	2	2	0	20	4	114	1	2
16	31	7	2	2	0	7	7	148	1	2
17	42	4	2	2	0	19	11	140	3	2
18	44	2	2	2	0	7	7	125	2	2
19	26	3	2	2	0	30	1	169	3	2
20	41	4	2	2	0	20	11	182	2	2

Figure 3.3 Dataset 2 sample

3.3 Flowchart

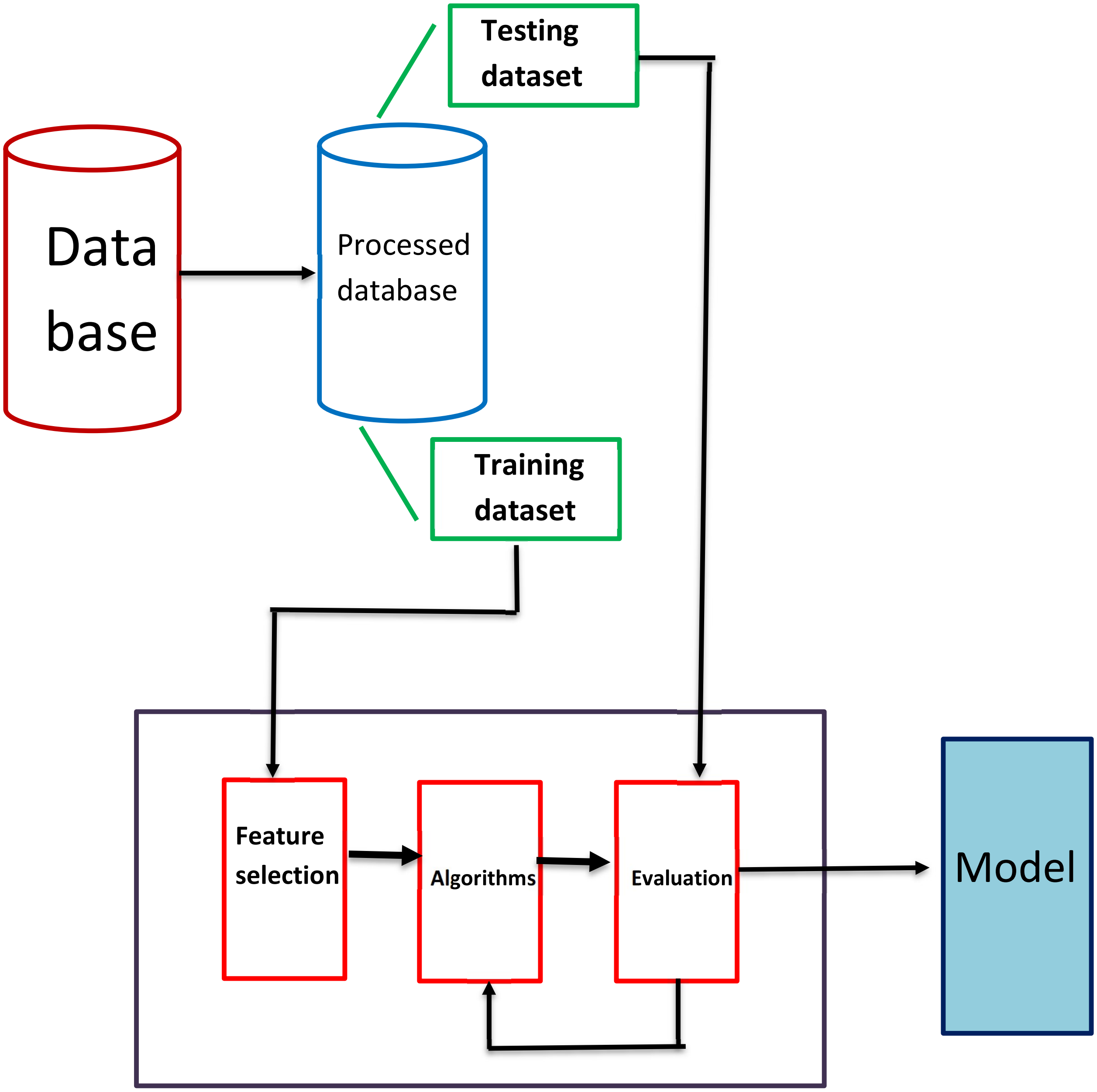


Figure 3.4 Flowchart of the proposed approach

3.4 Code Implementation

```
#import python libraries
import sys
print('Python: {}'.format(sys.version))
# scipy
import scipy
print('scipy: {}'.format(scipy.__version__))
# numpy
import numpy
print('numpy: {}'.format(numpy.__version__))
# matplotlib
import matplotlib
print('matplotlib: {}'.format(matplotlib.__version__))
# pandas
import pandas
print('pandas: {}'.format(pandas.__version__))
# scikit-learn
import sklearn
print('sklearn: {}'.format(sklearn.__version__))
```

```
#import python libraries

import seaborn as sns

import matplotlib.pyplot as plt

from pandas.plotting import scatter_matrix

from sklearn import cross_validation

from sklearn.metrics import classification_report

from sklearn.metrics import confusion_matrix

from sklearn.metrics import accuracy_score

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn.naive_bayes import GaussianNB

from sklearn.svm import SVC

#Resd dataset from the directory

dataset = pandas.read_csv("bankr1.csv")

print(dataset.shape)

print(dataset.head(20))

print(dataset.describe())

print(dataset.groupby('y').size())

dataset.hist()
```

```
plt.show()

sns.set(style="ticks")
sns.pairplot(dataset, hue="y")
array = dataset.values
X = array[:,0:15]
Y = array[:,15]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, Y, test_size=validation_size, random_state=seed)
num_folds = 10
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('SVM', SVC()))
models.append(('NB', GaussianNB()))
models.append(('KNN', KNeighborsClassifier()))
```

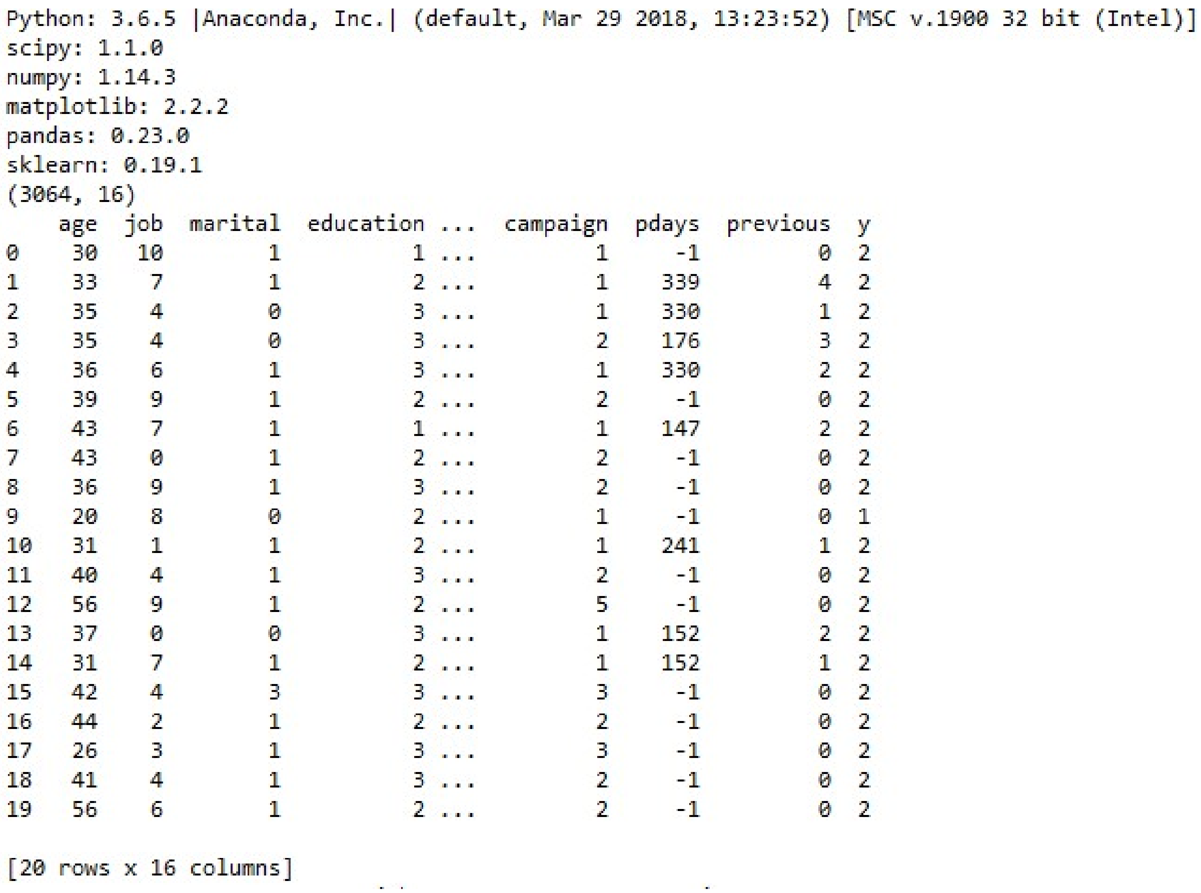
```
results = []

names = []

for name, model in models:

    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)

    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)

    results.append(cv_results)

    names.append(name)

    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())

    print(msg)

 # Compare Algorithms

fig = plt.figure()

fig.suptitle('Algorithm Comparison')

ax = fig.add_subplot(111)

plt.boxplot(results)

ax.set_xticklabels(names)

plt.show()


# Make predictions on validation dataset

knn = KNeighborsClassifier()

knn.fit(X_train, Y_train)
```

```
predictions = knn.predict(X_validation)
print('Accuracy score KNN',accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
#prediction of svm
svm = SVC()
print('SVM')
svm.fit(X_train, Y_train)
predictions=svm.predict(X_validation)
print('Accuracy score SVM',accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
#prediction of gaussian nb
nb=GaussianNB()
nb.fit(X_train, Y_train)
predictions = nb.predict(X_validation)
print('Accuracy score NB',accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
#prediction of LDA
lda=LinearDiscriminantAnalysis()
```

```
lda.fit(X_train, Y_train)
predictions = lda.predict(X_validation)
print('Accuracy score LDA',accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
#prediction of logistic regression
model=LogisticRegression(penalty='l2', max_iter=1000)
model.fit(X_train, Y_train)
predictions=model.predict(X_validation)
print('Accuracy score LR',accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

4. Experimental Results and Discussion

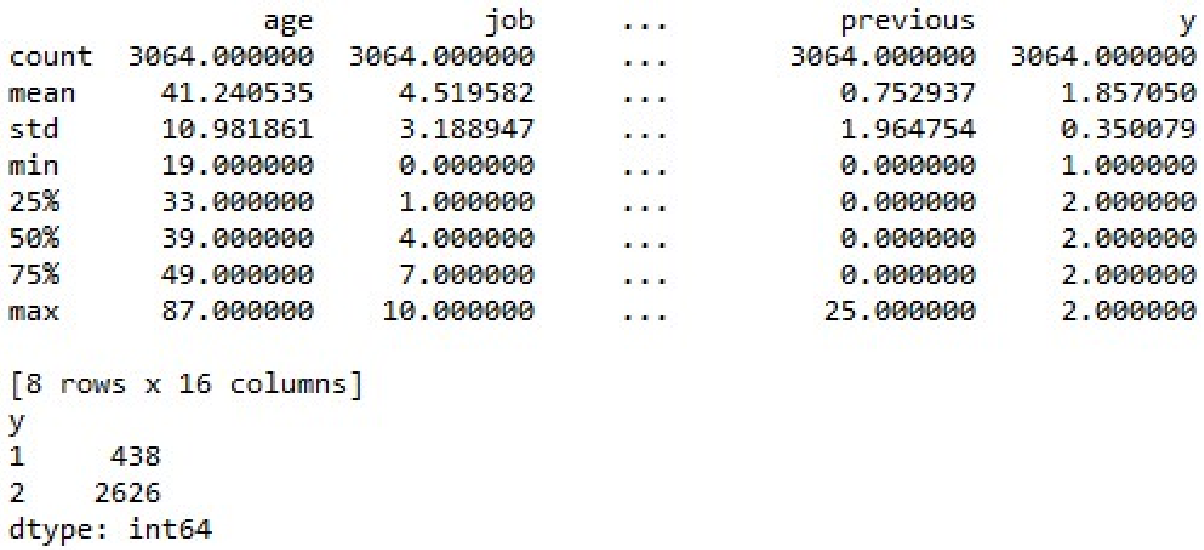
Initialize the work on the original Dataset 1 with version checking of the python libraries and ending with algorithm prediction results. Here, Spyder platform is used for the purpose. The version and dataset description as received by “dataset. shape “(represents dimensionality of dataset) and “dataset. head (20)” [returns the first 20 rows of the dataset] is depicted in figure 4.1:



```
Python: 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:23:52) [MSC v.1900 32 bit (Intel)]
scipy: 1.1.0
numpy: 1.14.3
matplotlib: 2.2.2
pandas: 0.23.0
sklearn: 0.19.1
(3064, 16)
    age  job  marital  education ...  campaign  pdays  previous  y
0    30   10        1          1 ...         1     -1         0  2
1    33    7        1          2 ...         1    339         4  2
2    35    4        0          3 ...         1    330         1  2
3    35    4        0          3 ...         2    176         3  2
4    36    6        1          3 ...         1    330         2  2
5    39    9        1          2 ...         2     -1         0  2
6    43    7        1          1 ...         1    147         2  2
7    43    0        1          2 ...         2     -1         0  2
8    36    9        1          3 ...         2     -1         0  2
9    20    8        0          2 ...         1     -1         0  1
10   31    1        1          2 ...         1    241         1  2
11   40    4        1          3 ...         2     -1         0  2
12   56    9        1          2 ...         5     -1         0  2
13   37    0        0          3 ...         1    152         2  2
14   31    7        1          2 ...         1    152         1  2
15   42    4        3          3 ...         3     -1         0  2
16   44    2        1          2 ...         2     -1         0  2
17   26    3        1          3 ...         3     -1         0  2
18   41    4        1          3 ...         2     -1         0  2
19   56    6        1          2 ...         2     -1         0  2

[20 rows x 16 columns]
```

Figure 4.1 Output showing version and description of dataset

Further we have used pandas describe () function to generate a descriptive statistic that summarizes the central tendency, dispersion and shape of a dataset’s distribution for the given dataset. The output of the function is displayed in figure 4.2. Histograms for different attributes of the dataset1 is plotted which gives an accurate graphical representation of the distribution of numerical data along with the estimate of the probability distribution of continuous variable. The histograms obtained is shown in figure 4.3.



```
               age          job     ...        previous            y
count  3064.000000  3064.000000     ...     3064.000000  3064.000000
mean     41.240535     4.519582     ...        0.752937     1.857050
std      10.981861     3.188947     ...        1.964754     0.350079
min      19.000000     0.000000     ...        0.000000     1.000000
25%      33.000000     1.000000     ...        0.000000     2.000000
50%      39.000000     4.000000     ...        0.000000     2.000000
75%      49.000000     7.000000     ...        0.000000     2.000000
max      87.000000    10.000000     ...       25.000000     2.000000

[8 rows x 16 columns]
y
1     438
2    2626
dtype: int64
```

Figure 4.2 Statistical description of dataset

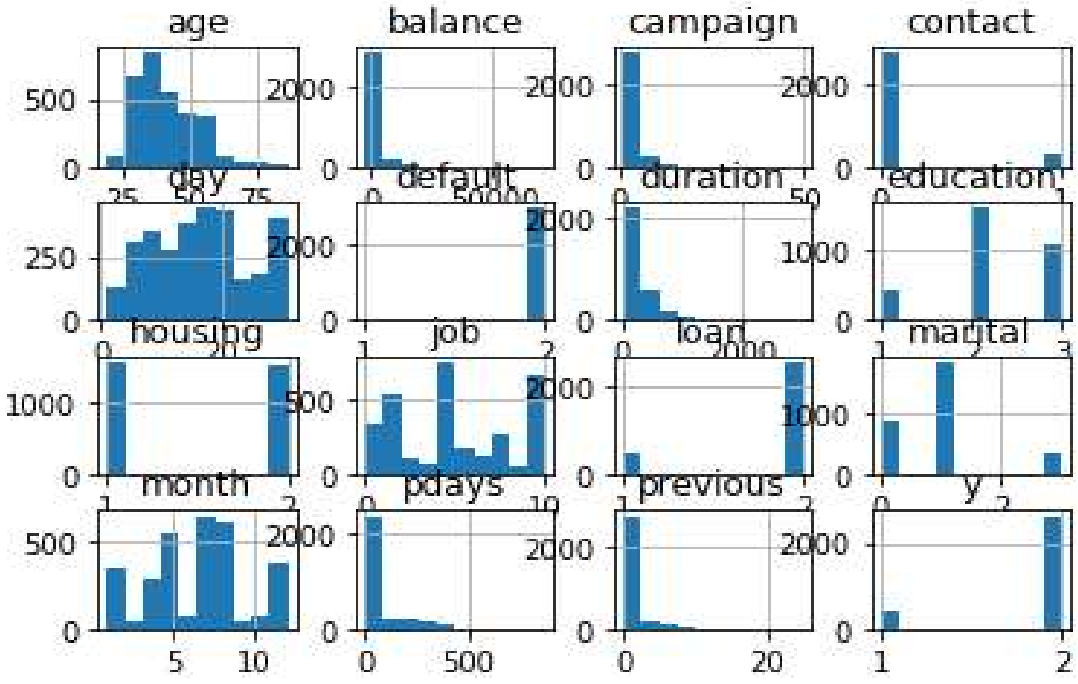


Figure 4.3 Histograms

In this work the dataset used is splitted in two parts, training and validation. Where training set contains 80% of the data and validation set has 20% of data. We have 2452 instances in the training set and 614 instances in validation set.

Carrying on, five machine learning algorithms i.e. support vector machine, Gaussian NB, Linear discriminant analysis, logistic regression and knearest neighbor are analyzed and their mean and std is received as mention below:

LR: 0.867815 (0.016192)

LDA: 0.866590 (0.012591)

SVM: 0.860056 (0.017613)

NB: 0.849056 (0.022750)

KNN: 0.860063 (0.010079)

Further scatter pairplot is analyzed, it is used to see relationships and spread of each data point .It takes each numerical column puts them on both the axes and plot a scatter plot where they meet .Where the same variables meet ,we get a histogram that shows the distribution of the variables. Figure 4.4 shows the pair plot sample.

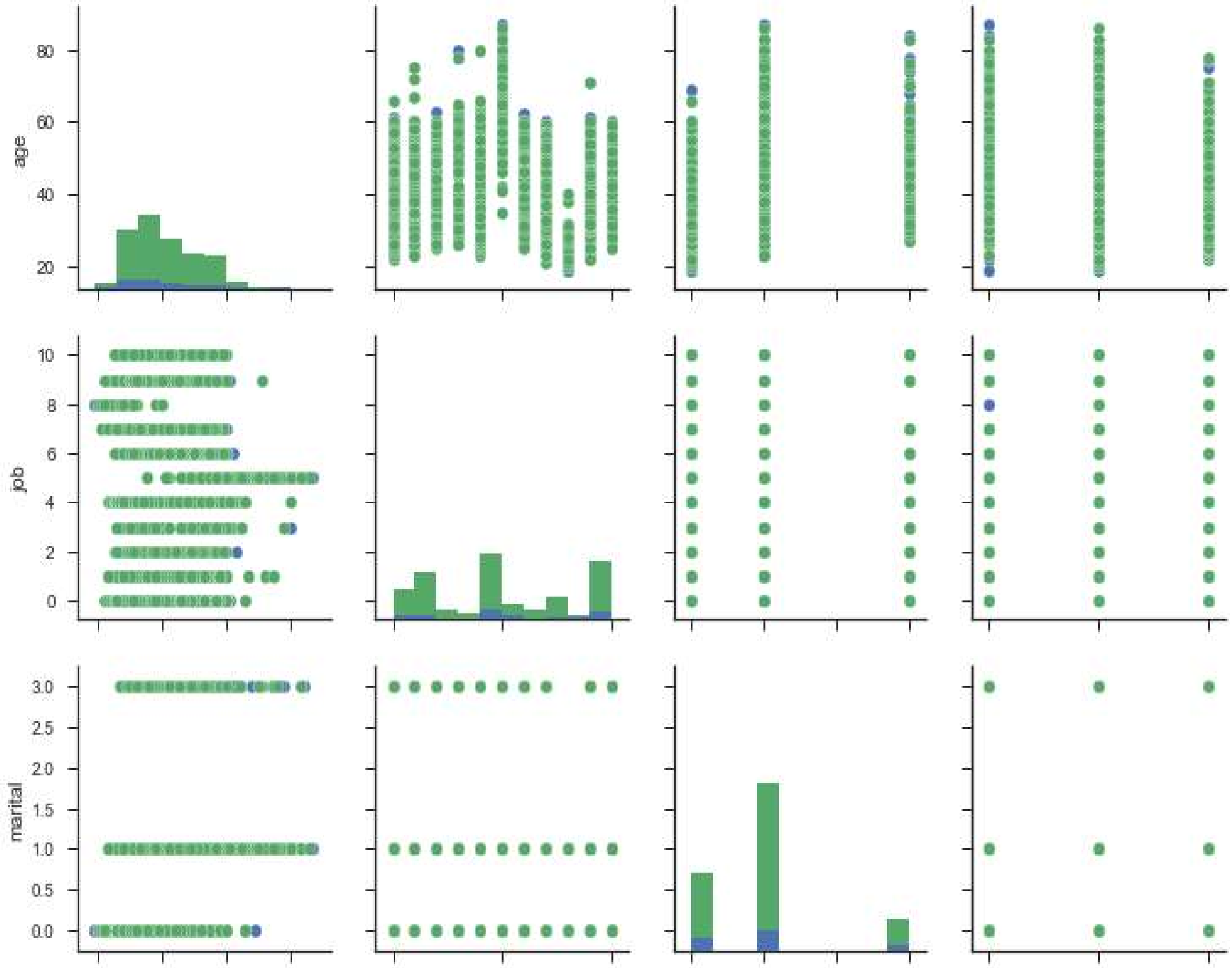


Figure 4.4 Pair plot sample

Last section of scatter plot shows the output variable correlation with the different attributes age, marital, job, etc. that how many of with the respective age has subscribed for deposit or not, the deposit subscription is related all the attributes. Figure 4.5 depict the mentioned section.

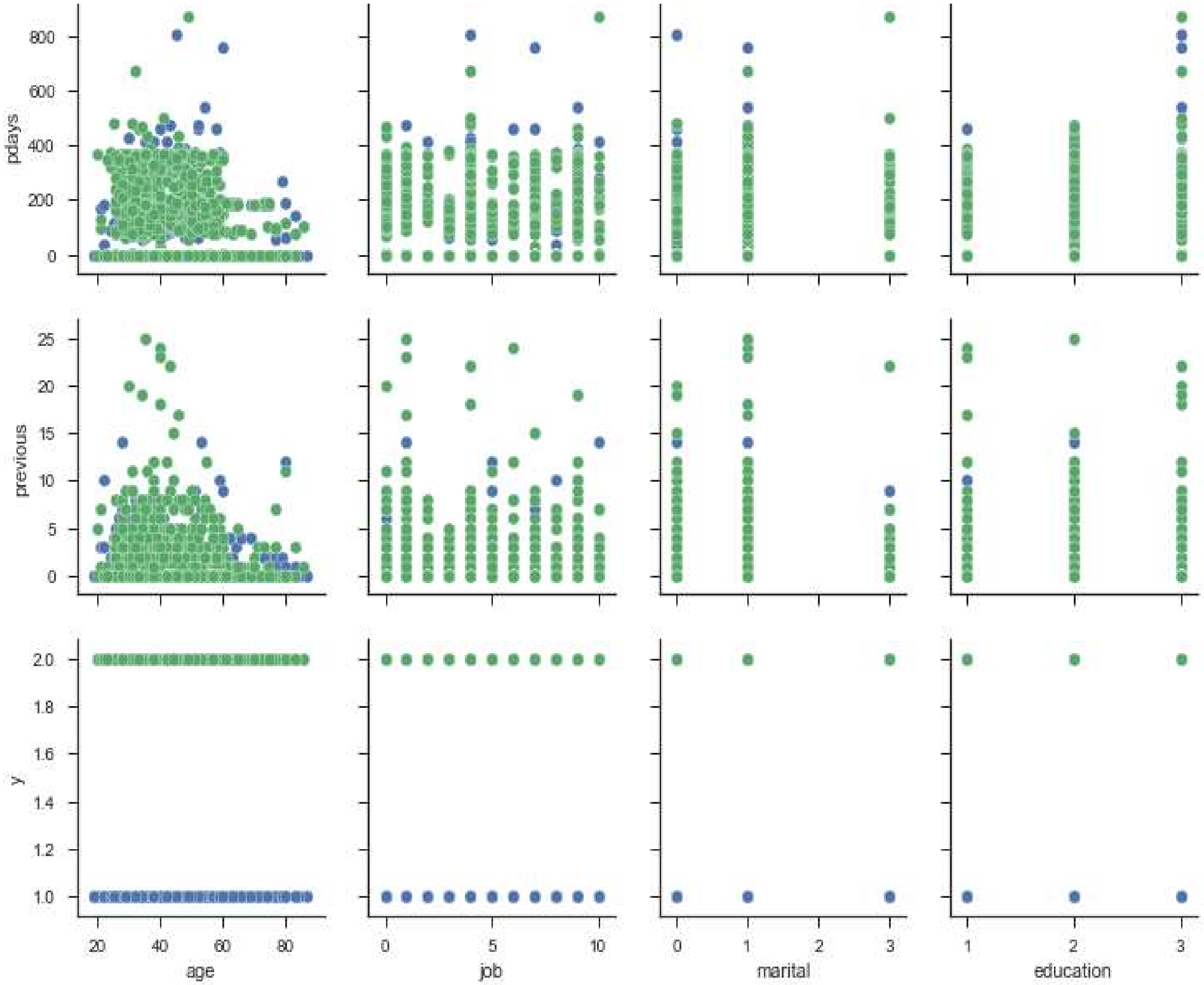


Figure 4.5 Last section of scatter plot

Boxplots are as standardized way of displaying the distribution of data based on a five number summary (“minimum “, first quartile (Q1), median, third quartile (Q3), and “maximum”). Here, boxplot is used for algorithm comparison (as shown in figure 4.6), which shows that LDA is the best performing algorithm with least errors while NB is the worst performing algorithm.

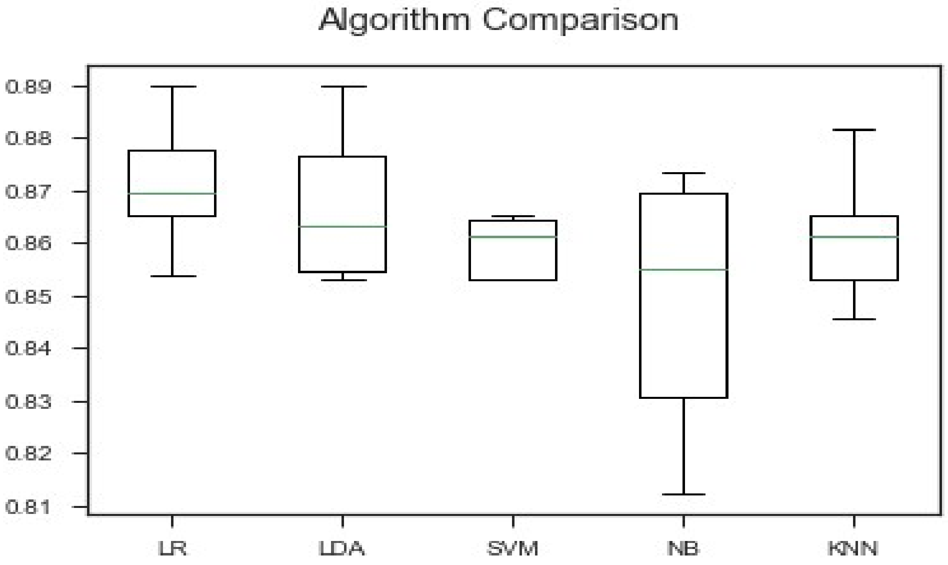
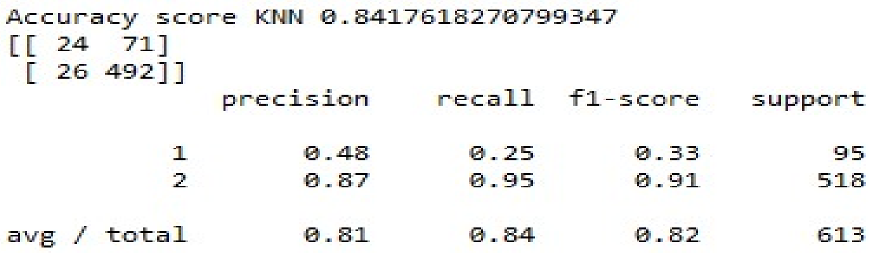


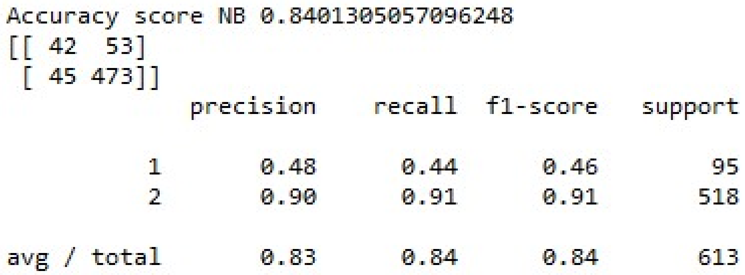
Figure 4.6 Boxplot of different algorithm of dataset 1

In the process of performance evaluation of each classification model, they are evaluated using three statistical measures; classification accuracy, sensitivity and specificity. These measures are defined as confusion matrix. Confusion matrix along with accuracy score for various algorithms is depicted below in figure 4.7.



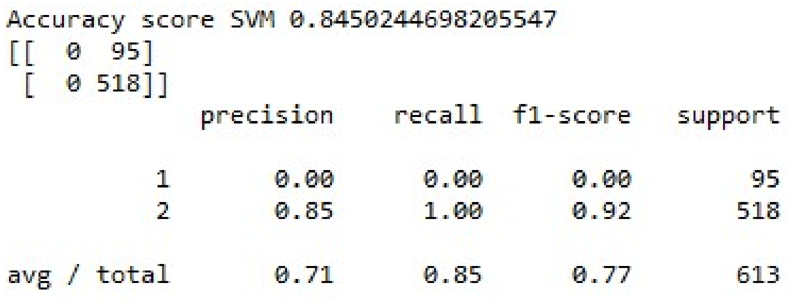
```
Accuracy score KNN 0.8417618270799347
[[ 24  71]
 [ 26 492]]
             precision    recall  f1-score   support

          1       0.48      0.25      0.33        95
          2       0.87      0.95      0.91       518

avg / total       0.81      0.84      0.82       613
```

(a)



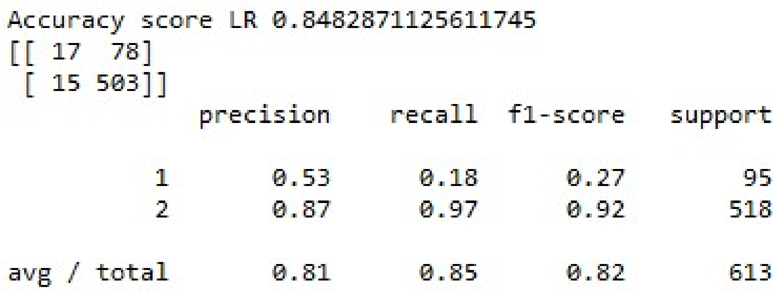
```
Accuracy score NB 0.8401305057096248
[[ 42  53]
 [ 45 473]]
             precision    recall  f1-score   support

          1       0.48      0.44      0.46        95
          2       0.90      0.91      0.91       518

avg / total       0.83      0.84      0.84       613
```

(b)



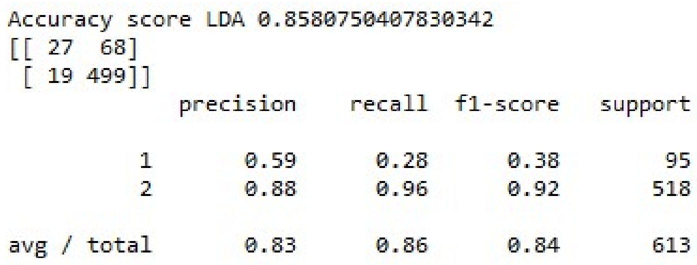
```
Accuracy score SVM 0.8450244698205547
[[  0  95]
 [  0 518]]
             precision    recall  f1-score   support

          1       0.00      0.00      0.00        95
          2       0.85      1.00      0.92       518

avg / total       0.71      0.85      0.77       613
```

(c)



```
Accuracy score LR 0.8482871125611745
[[ 17  78]
 [ 15 503]]
             precision    recall  f1-score   support

          1       0.53      0.18      0.27        95
          2       0.87      0.97      0.92       518

avg / total       0.81      0.85      0.82       613
```

(d)



```
Accuracy score LDA 0.8580750407830342
[[ 27  68]
 [ 19 499]]
             precision    recall  f1-score   support

          1       0.59      0.28      0.38        95
          2       0.88      0.96      0.92       518

avg / total       0.83      0.86      0.84       613
```

(e)

Figure 4.7 Accuracy score and confusion matrices of dataset 1 for (a) Knn, (b)NB, (c)SVM, (d)LR, (e)LDA

To enhance the performance of the algorithms, we have worked on Dataset 2 which is modified form of Dataset 1 obtained by dropping some of the attributes i.e. marital, education, balance , housing, pdays and previous. Table 4.1 illustrates the attributes used in the datasets mentioned.

Table 4.1 Attributes used in the datasets

Attributes	Attributes used in dataset 1	Attributes used in dataset 2
Age	yes	yes
Job	yes	yes
Marital	yes	no
Education	yes	no
Default	yes	yes
Balance	yes	no
Housing	yes	no
Loan	yes	yes
Contact	yes	yes
Day	yes	yes
Month	yes	yes
Duration	yes	yes
Campaign	yes	yes
Pdays	yes	no
Previous	yes	no
Poutcome	no	no

The same process is carried out with dataset 2 as with dataset 1. Plotted the histograms, boxplot, and confusion matrix, and did the algorithm comparison. Table 4.2 depicts the algorithm comparison between both the datasets.

Table 4.2 Accuracy score comparison between two datasets

Algorithms	Accuracy scores for Dataset 1	Accuracy scores for Dataset 2
SVM	0.8450	0.8450
LR	0.8482	0.8482
Gaussian NB	0.8401	0.8629
LDA	0.8580	0.8580
KNN	0.8417	0.8548

Boxplot, Confusion matrices of all the five algorithms along with the accuracy scores is represented below in figure 4.8 and figure 4.9 respectively when evaluated on dataset 2.

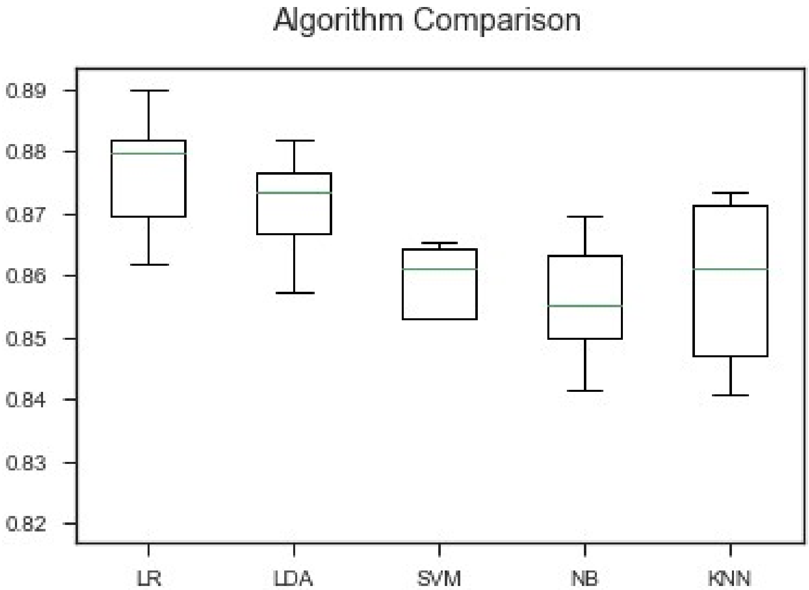
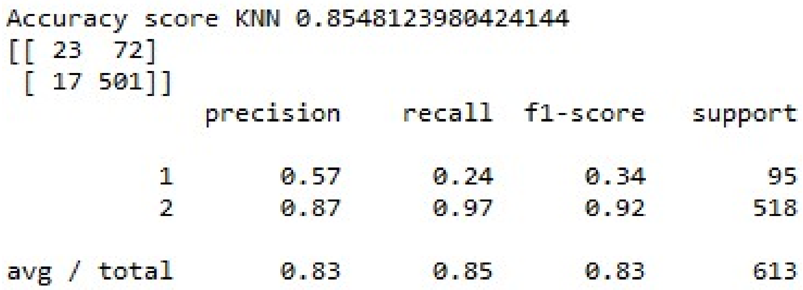
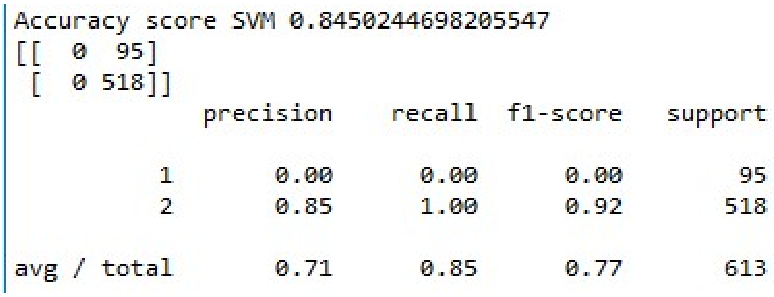


Figure 4.8 Boxplot of algorithm comparison of dataset 2



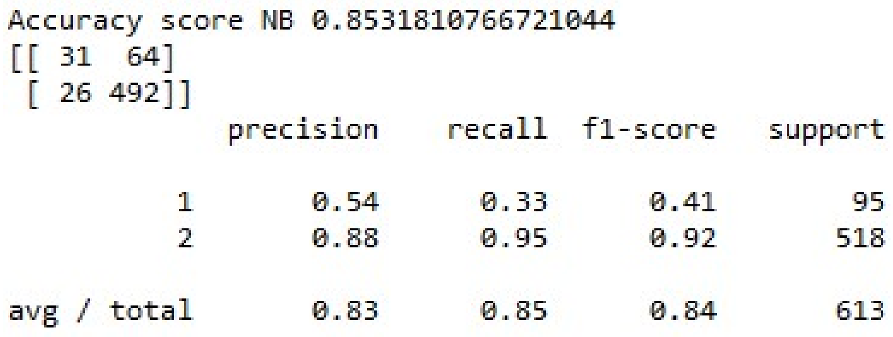
```
Accuracy score KNN 0.8548123980424144
[[ 23  72]
 [ 17 501]]
             precision    recall  f1-score   support

          1       0.57      0.24      0.34        95
          2       0.87      0.97      0.92       518

avg / total       0.83      0.85      0.83       613
```

(a)



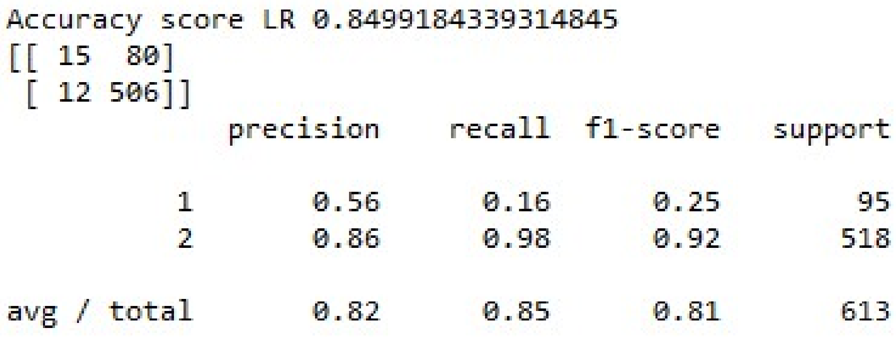
```
Accuracy score SVM 0.8450244698205547
[[  0  95]
 [  0 518]]
             precision    recall  f1-score   support

          1       0.00      0.00      0.00        95
          2       0.85      1.00      0.92       518

avg / total       0.71      0.85      0.77       613
```

(b)



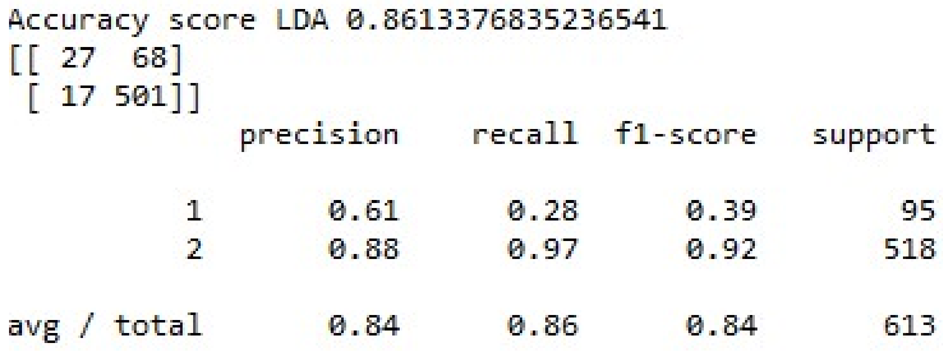
```
Accuracy score NB 0.8531810766721044
[[ 31  64]
 [ 26 492]]
             precision    recall  f1-score   support

          1       0.54      0.33      0.41        95
          2       0.88      0.95      0.92       518

avg / total       0.83      0.85      0.84       613
```

(c)



```
Accuracy score LR 0.8499184339314845
[[ 15  80]
 [ 12 506]]
             precision    recall  f1-score   support

          1       0.56      0.16      0.25        95
          2       0.86      0.98      0.92       518

avg / total       0.82      0.85      0.81       613
```

(d)



```
Accuracy score LDA 0.8613376835236541
[[ 27  68]
 [ 17 501]]
             precision    recall  f1-score   support

          1       0.61      0.28      0.39        95
          2       0.88      0.97      0.92       518

avg / total       0.84      0.86      0.84       613
```

(e)

Figure 4.9 Accuracy score and confusion matrices of dataset 2 for (a) kNN, (b)SVM, (c) NB, (d)LR, and (e)LDA

From above interpretations it is seen that when working with dataset 2, the accuracy score of Gaussian NB and KNN increase and scores of SVM, LDA and LR remain same as the dataset 1. So, it can be concluded that 9 attributes are sufficient to predict the deposit subscription with high accuracy percent.

5. Conclusion And Future Work

This work is evaluating the performance of five machine learning algorithms on bank direct marketing dataset to classify for bank deposit subscription. Secondly it focuses on extracting the least attributes sufficient for increasing the accuracy score rather than using all the attributes. Taking all the attributes in consideration we have LDA to be best performing algorithm with accuracy score 85.80% and Guassian NB with accuracy score 84.01% performs the worst. After modification in dataset we conclude that Gaussian NB algorithm performs best with accuracy score of 86.29% and worst performing algorithm is SVM with accuracy 84.50%. This study also emphasis that 9 attributes are sufficient for prediction purpose.

Dataset used in this work is just an extract of a large dataset, in future, the dataset that used for evaluation can be spanned for large instances and more attributes. Further, more machine learning algorithms can be used for comparison purpose and predicting result.

Bibliography

[1] Elsalamony, Hany A. "Bank direct marketing analysis of data mining techniques." *International Journal of Computer Applications* 85.7 (2014): 12-22.

[2] Asare-Frempong, Justice, and Manoj Jayabalan. "Predicting customer response to bank direct telemarketing campaign." *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*. IEEE, 2017.

[3] Moro, Sérgio, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing." *Decision Support Systems* 62 (2014): 22-31.

[4] Jiang, Yiyan. "Using Logistic Regression Model to Predict the Success of Bank Telemarketing." *International Journal on Data Science and Technology* 4.1 (2018): 35.

[5] Ghosh, Soumadip, et al. "A comparative study to the bank market prediction." *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, Cham, 2018.

[6] Apampa, Olatunji. "Evaluation of classification and ensemble algorithms for bank customer marketing response prediction." *Journal of International Technology and Information Management* 25.4 (2016): 6.

[7] Moro, Sergio, Raul Laureano, and Paulo Cortez. "Using data mining for bank direct marketing: An application of the crisp-dm methodology." *Proceedings of European Simulation and Modelling Conference-ESM'2011*. EUROSIS-ETI, 2011.

[8] Prusty, S. "Data mining applications to direct marketing: identifying hot prospects for banking product." *Web data mining (ECT 584), Spring. DePaul University, Chicago* (2013).

[9] www.medium.com

[10]https://www.kaggle.com/janiobachmann/bank-marketing-dataset

[11]www.kaggle.com