Predicting the Downfall of Non-Profit Organizations Using Machine Learning

Rubén A. Vázquez Rodríguez
Master of Engineering in Computer Engineering
Advisor: Dr. Othoniel Rodríguez Jiménez
Electrical & Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract — *Machine learning can be applied to finances of non-profit organizations taken from IRS Tax Forms 990ez to determine if an organization will be dissolved. This is useful to determine if a cause is viable. Data stored on an online database is extracted, formatted, parsed and segregated using Python. The code selects the attributes used to predict the organization's downfall. Finances were compared and attributes that were critical were identified. Three supervised predictive algorithms, Decision Tree, K Nearest Neighbors and Naïve Bayes, were used. Results from the algorithm's predictions for organizations that were dissolved and non-dissolved are presented in this paper and discussed. This study also determined the average duration of non-profit organizations based on the current financials.*

Key Terms —*Algorithms, Analytics, Big Data, Prediction, Machine Learning.*

INTRODUCTION

In modern days the amount of data available for analysis is vast enough to allow us to predict the behavior of almost anything, including image and speech recognition, medical diagnosis, traffic conditions, financial services and so on. Big data, which describes extremely large data sets, is widely being used nowadays for research and analytics. Traditional databases are not capable of handling big data. Machine learning is an interdisciplinary research area which focuses on the development of fast and efficient learning algorithms which can make predictions on data [1]. This article presents an application of machine learning related to the finances of non-profit organizations. The goal in this work is to determine, using machine learning, if an organization will be dissolved or not, based on their finances as reported in their IRS tax form 990ez, which is located in a public database. Knowing if the non-profit organization will be dissolved could help investors decide if it is viable to support that specific cause. Taking into consideration the expenses and revenues is not enough to determine the success or failure of an organization. Other factors that can influence the outcome will also be examined in this paper.

Machine Learning

Handling big data is an extremely difficult task to carry out using conventional data processing applications. It usually involves finding on it the relevant information, modeling the elements composing it, and transforming it into useful information and knowledge. For such goals Machine Learning techniques are used. These techniques provide methods to treat and extract information from data automatically, where human operators and experts are not able to deal with because of the level of complexity or the volume to be treated per time unit [2].

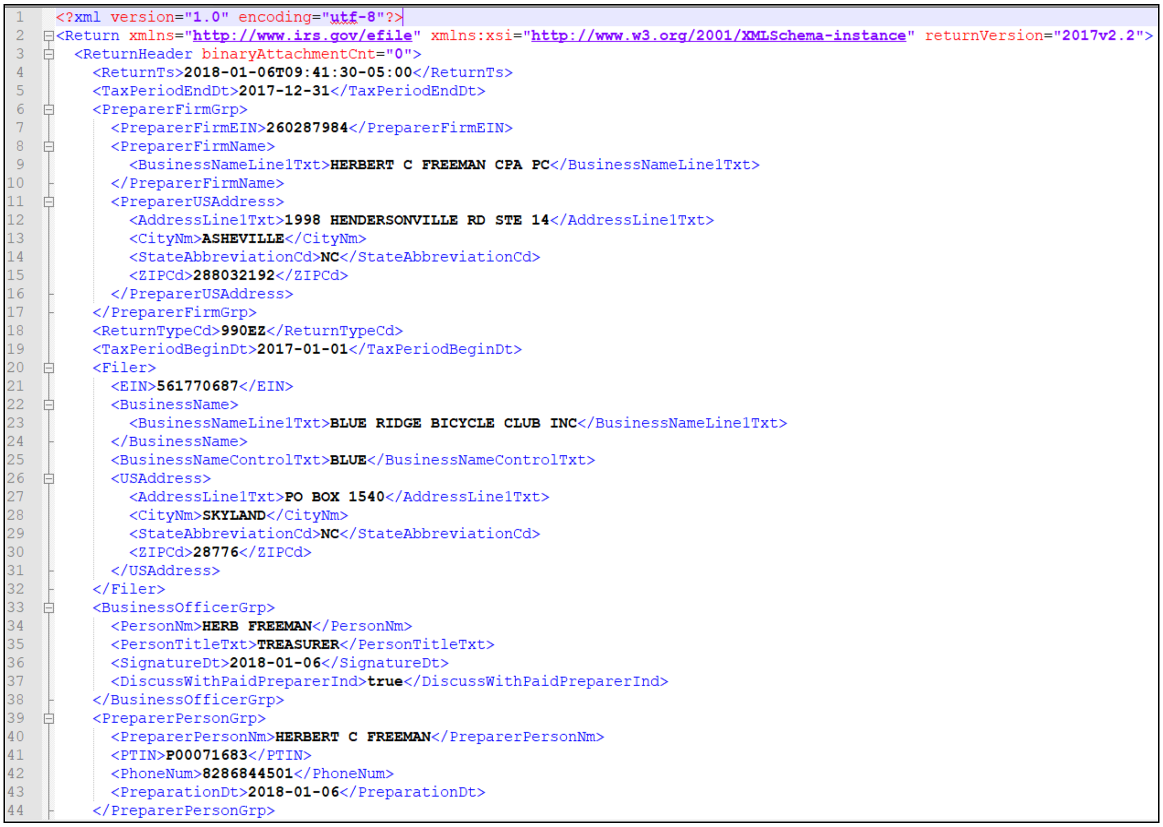
Machine learning tasks are grouped into three categories: supervised, unsupervised and reinforcement learning. Supervised machine learning requires training with labeled data, each consisting of input value and a desired target value. The supervised learning algorithm analyzes the training data and makes an inferred function. In unsupervised machine learning, hidden insights are drawn from unlabeled data sets. Reinforcement learning allows a machine to learn its behavior from feedback received through the interactions with an external environment [1]. From a data processing point of view, supervised and unsupervised learning techniques are preferred for data analysis, and reinforcement techniques are preferred for decision making [3].

Methods

Data Retrieval and Parsing

The first step to begin the analysis is to retrieve the data from the Amazon Web Services S3, an online database. The data is stored as an XML format. It then needs to be parsed into tables so the predicting algorithms can iterate over the rows. It should be pointed out that the database contains all the type of forms that can be filled out by the non-profit organizations. These include forms 990, 990ez and 990pf. The 990 form is used for organization with gross receipt greater than $200k or total assets greater than $500k. The 990ez form is used for organizations with gross receipt less than $200k and total assets less than $500k and the form 990pf is used for private foundations regardless of the financial status. The 990ez form contains 27 attributes related to the cash flow of the organization while the form 990 and 990pf contains more than 120 attributes of cash flow depending on the size of the organization. This study will be limited to the data for 990ez due to hardware limitations. The retrieval of the data from AWS S3 took approximately 5 days to download 2,959,695 files with a total of 91.4Gb. Parsing the XML files to XLSX took 5 days even though the Python algorithm was filtering only by 990ez form.

To retrieve the data from AWS S3 using AWS Command Line Interface the command *aws s3 ls s3://irs-form-990/*./Form990xml –recursive* was used. After the command has been run, the following form of XML as shown in Figure 1 is downloaded into the desired location.



```xml
<?xml version="1.0" encoding="utf-8"?>
<Return xmlns="http://www.irs.gov/efile" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" returnVersion="2017v2.2">
  <ReturnHeader binaryAttachmentCnt="0">
    <ReturnTs>2018-01-06T09:41:30-05:00</ReturnTs>
    <TaxPeriodEndDt>2017-12-31</TaxPeriodEndDt>
    <PreparerFirmGrp>
      <PreparerFirmEIN>260287984</PreparerFirmEIN>
      <PreparerFirmName>
        <BusinessNameLine1Txt>HERBERT C FREEMAN CPA PC</BusinessNameLine1Txt>
      </PreparerFirmName>
      <PreparerUSAddress>
        <AddressLine1Txt>1998 HENDERSONVILLE RD STE 14</AddressLine1Txt>
        <CityNm>ASHEVILLE</CityNm>
        <StateAbbreviationCd>NC</StateAbbreviationCd>
        <ZIPCd>288032192</ZIPCd>
      </PreparerUSAddress>
    </PreparerFirmGrp>
    <ReturnTypeCd>990EZ</ReturnTypeCd>
    <TaxPeriodBeginDt>2017-01-01</TaxPeriodBeginDt>
    <Filer>
      <EIN>561770687</EIN>
      <BusinessName>
        <BusinessNameLine1Txt>BLUE RIDGE BICYCLE CLUB INC</BusinessNameLine1Txt>
      </BusinessName>
      <BusinessNameControlTxt>BLUE</BusinessNameControlTxt>
      <USAddress>
        <AddressLine1Txt>PO BOX 1540</AddressLine1Txt>
        <CityNm>SKYLAND</CityNm>
        <StateAbbreviationCd>NC</StateAbbreviationCd>
        <ZIPCd>28776</ZIPCd>
      </USAddress>
    </Filer>
    <BusinessOfficerGrp>
      <PersonNm>HERB FREEMAN</PersonNm>
      <PersonTitleTxt>TREASURER</PersonTitleTxt>
      <SignatureDt>2018-01-06</SignatureDt>
      <DiscussWithPaidPreparerInd>true</DiscussWithPaidPreparerInd>
    </BusinessOfficerGrp>
    <PreparerPersonGrp>
      <PreparerPersonNm>HERBERT C FREEMAN</PreparerPersonNm>
      <PTIN>P00071683</PTIN>
      <PhoneNum>8286844501</PhoneNum>
      <PreparationDt>2018-01-06</PreparationDt>
    </PreparerPersonGrp>
```

Figure 1

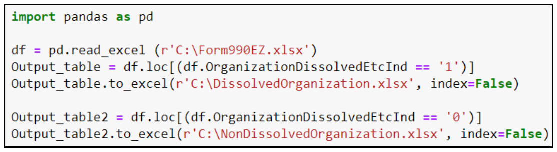
Example of an XML Structure

The development was done using Python, a programming language that contains multiple libraries for Analytics. In this study, a code was generated for the data parsing which iterates over the XML using the ElementTree library. The code selects the attributes that describe the organization and that could be used for predicting the downfall of the non-profit organizations. The code produces a XLSX file as shown in Table 1 with all the 990ez selected content from organizations between the years 2009 to 2019.

Data Segregation and Cleaning

After the data is changed to a legible format, it is segregated into dissolved organizations and non-dissolved organizations in order to analyze their behaviors. The data extracted from AWS S3 contains "OrganizationDissolvedEtcInd", an attribute that determines if the organization was dissolved at the end of the year**.** This attribute had to be standardized since some forms contain a 0, FALSE, 1 or TRUE value. The attribute was transformed to either 0 or 1: 0 meaning the organization is still operating and 1 the organization was dissolved. The predictive algorithms need the data of the features in the same format and size in order to function accurately.

The file generated is approximately 186Mb in size and contains 807,828 rows. Each row contains the information of an organization in a determined year. The non-dissolved organizations sum 798,551 whereas the dissolved organizations total 9,277. The Python algorithm in Figure 2 was run to create two distinct files, one for dissolved organizations and the second for the non-dissolved ones. This will allow focus of the analysis on each category by separate, to determine patterns and behaviors.



```
import pandas as pd

df = pd.read_excel (r'C:\Form990EZ.xlsx')
Output_table = df.loc[(df.OrganizationDissolvedEtcInd == '1')]
Output_table.to_excel(r'C:\DissolvedOrganization.xlsx', index=False)

Output_table2 = df.loc[(df.OrganizationDissolvedEtcInd == '0')]
Output_table2.to_excel(r'C:\NonDissolvedOrganization.xlsx', index=False)
```

Figure 2

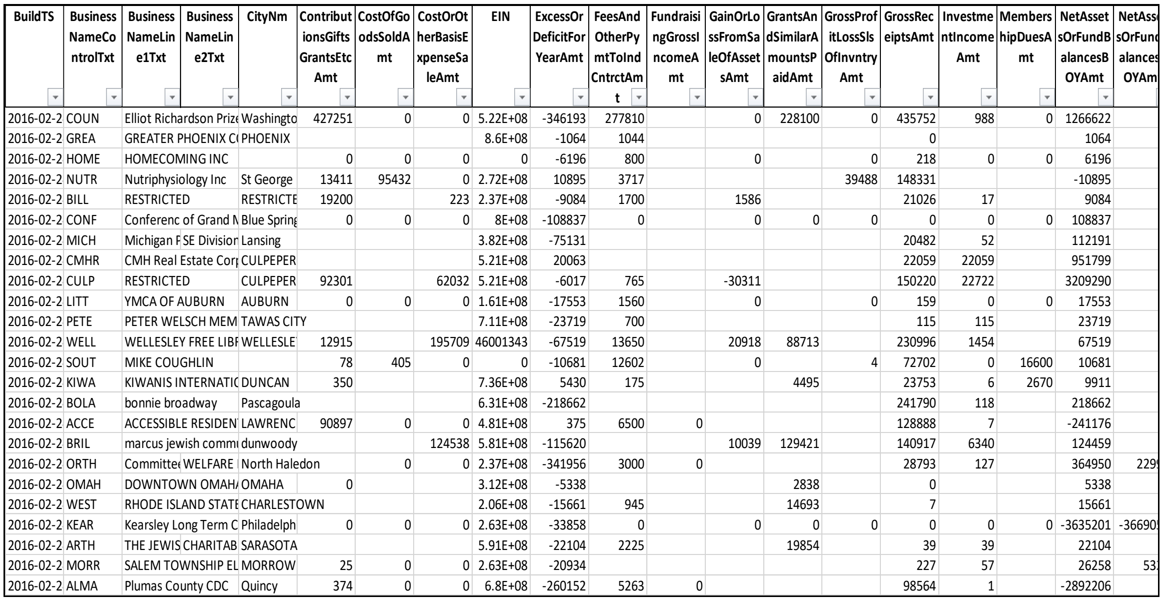
Code to Create Files for Dissolved and Non-Dissolved Organizations Separately

Attribute Selection

The platform Qlik, a visual analytics tool, is highly impactful in providing data analytics solutions [1]. The tool was used to obtain an overview of the data and helps to identify which attributes are significant in determining the downfall of an organization.

Table 1

Example of a Truncated XLSX File With All the 990ez Organizations



BuildTS	BusinessNameControlTxt	BusinessNameLine1Txt	BusinessNameLine2Txt	CityNm	ContributionsGiftsGrantsEtcAmt	CostOfGoodsSoldAmt	CostOrOtherBasisExpenseSaleAmt	EIN	ExcessOrDeficitForYearAmt	FeesAndOtherPymtToIndCntrctAmt	FundraisingGrossIncomeAmt	GainOrLossFromSaleOfAssetsAmt	GrantsAndSimilarAmountsPaidAmt	GrossProfitLossSlsOfInvntryAmt	GrossReceiptsAmt	InvestmentIncomeAmt	MembershipDuesAmt	NetAssetsOrFundBalancesBOYAmt	NetAsssOrFundalancesOYAm
2016-02-2	COUN	Elliot Richardson Prize		Washingto	427251	0	0	5.22E+08	-346193	277810		0	228100	0	435752	988	0	1266622	
2016-02-2	GREA	GREATER PHOENIX C(		PHOENIX				8.6E+08	-1064	1044					0			1064	
2016-02-2	HOME	HOMECOMING INC			0	0	0	0	-6196	800		0		0	218	0	0	6196	
2016-02-2	NUTR	Nutriphysiology Inc		St George	13411	95432	0	2.72E+08	10895	3717				39488	148331			-10895	
2016-02-2	BILL	RESTRICTED		RESTRICTE	19200		223	2.37E+08	-9084	1700		1586			21026	17		9084	
2016-02-2	CONF	Conferenc of Grand N		Blue Spring	0	0	0	8E+08	-108837	0		0	0	0	0	0	0	108837	
2016-02-2	MICH	Michigan F	SE Division	Lansing				3.82E+08	-75131						20482	52		112191	
2016-02-2	CMHR	CMH Real Estate Corp		CULPEPER				5.21E+08	20063						22059	22059		951799	
2016-02-2	CULP	RESTRICTED		CULPEPER	92301		62032	5.21E+08	-6017	765		-30311			150220	22722		3209290	
2016-02-2	LITT	YMCA OF AUBURN		AUBURN	0	0	0	1.61E+08	-17553	1560		0		0	159	0	0	17553	
2016-02-2	PETE	PETER WELSCH MEM		TAWAS CITY				7.11E+08	-23719	700					115	115		23719	
2016-02-2	WELL	WELLESLEY FREE LIBR		WELLESLE	12915		195709	46001343	-67519	13650		20918	88713		230996	1454		67519	
2016-02-2	SOUT	MIKE COUGHLIN			78	405	0	0	-10681	12602		0		4	72702	0	16600	10681	
2016-02-2	KIWA	KIWANIS INTERNATIC		DUNCAN	350			7.36E+08	5430	175			4495		23753	6	2670	9911	
2016-02-2	BOLA	bonnie broadway		Pascagoula				6.31E+08	-218662						241790	118		218662	
2016-02-2	ACCE	ACCESSIBLE RESIDEN		LAWRENC	90897	0	0	4.81E+08	375	6500	0				128888	7		-241176	
2016-02-2	BRIL	marcus jewish commu		dunwoody			124538	5.81E+08	-115620			10039	129421		140917	6340		124459	
2016-02-2	ORTH	Committe	WELFARE	North Haledon		0	0	2.37E+08	-341956	3000	0				28793	127		364950	229
2016-02-2	OMAH	DOWNTOWN OMAH/		OMAHA	0			3.12E+08	-5338				2838		0			5338	
2016-02-2	WEST	RHODE ISLAND STATE		CHARLESTOWN				2.06E+08	-15661	945			14693		7			15661	
2016-02-2	KEAR	Kearsley Long Term C		Philadelph	0	0	0	2.63E+08	-33858	0		0	0	0	0	0	0	-3635201	-36690
2016-02-2	ARTH	THE JEWIS	CHARITAB	SARASOTA				5.91E+08	-22104	2225			19854		39	39		22104	
2016-02-2	MORR	SALEM TOWNSHIP EL		MORROW	25	0	0	2.63E+08	-20934						227	57		26258	53
2016-02-2	ALMA	Plumas County CDC		Quincy	374	0	0	6.8E+08	-260152	5263	0				98564	1		-2892206	

Starting with the non-dissolved organization file the following behavior is observed: on average the revenues are greater than the expenses thru the years. In Figure 3, the red line represents the revenues and the blue line represents the expenses. In Figure 4 the maximum average deficit of all non-dissolved organizations did not exceed $1K dollars per year.

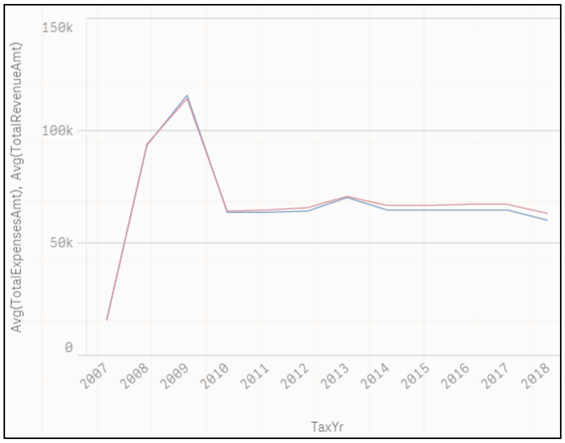


Figure 3
Non-Dissolved Organizations Expenses and Revenues per Year

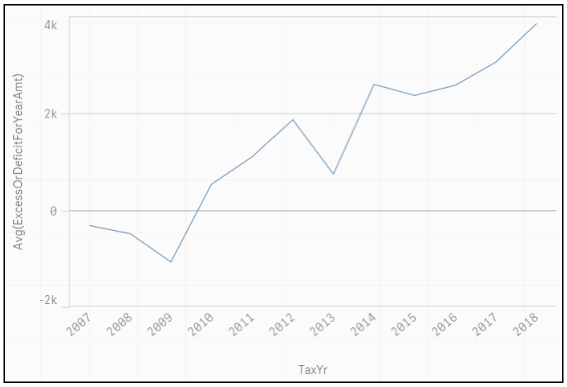


Figure 4
Non-Dissolved Organizations Deficit Per Year

The average net assets of all the non-dissolved organizations at the end of the year was greater than the net asset at the beginning of the year. The red line in Figure 5 is related to the net asset at the end of the year and the blue line is related to net assets at the beginning of the year.

Plotting the same attributes for the dissolved organizations yields the following results in Figure 6. This plot evidences that the expenses in an organization that were dissolved are greater than the revenues received in a determined year. This behavior is the opposite of the expenses and revenues of a non-dissolved organization.

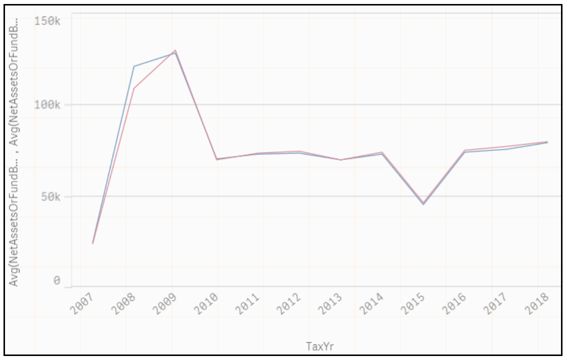


Figure 5
Non-Dissolved Organizations Balance BOY and EOY per Year

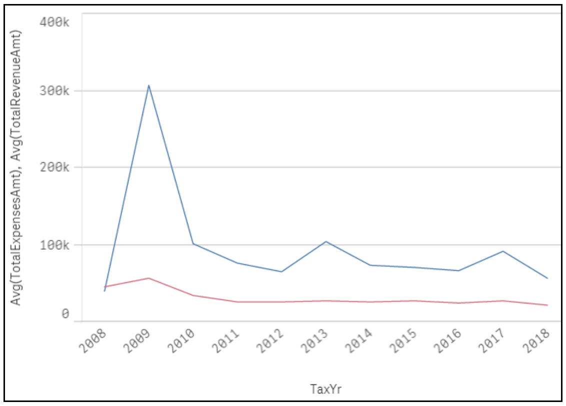


Figure 6
Dissolved Organizations Expenses and Revenues per Year

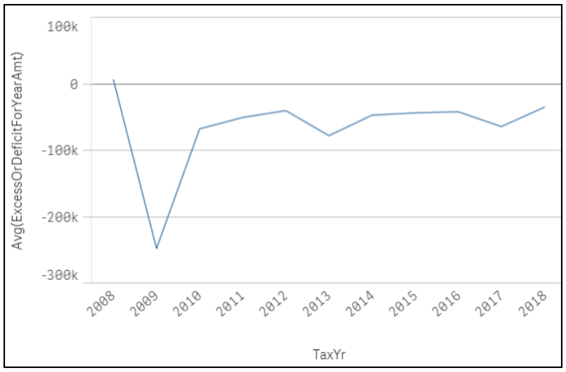


Figure 7
Dissolved Organizations Deficit Per Year

Figure 7 shows the deficit of a dissolved organization is approximately $55k dollars per year. In 2009 all organizations, both dissolved and non-dissolved, suffered losses due to the downfall in the market however the effect was most notable on the dissolved organizations as can be seen on plot below. Looking at the comparison between the assets at the beginning of the year and those at the end of the year of the dissolved organizations in Figure 8, it is observed that on average the organizations end the year with almost zero dollars in their balances.

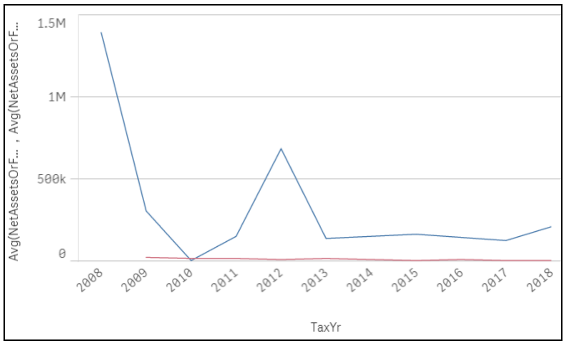


Figure 8

Dissolved Organizations Balance BOY and EOY Per Year

This quick insight provided by Qlik helped identify the attributes that were critical between a non-dissolved organization and a dissolved one. Five attributes were selected for the prediction: Total Expenses, Total Revenues, Total Deficit, Total Asset Balance at the beginning of the year and the Total Asset Balance at the end of the year.

Algorithms and Approaches

Since organizations that were dissolved and non-dissolved were identified because of the attribute "OrganizationDissolvedEtcInd", supervised algorithms for the prediction were used. Python allows users to access the SciKit-Learn library, which provides a wide sort of algorithms for classification, regression and clustering [2]. Due to the nature of the data the following predicting algorithms were selected: Decision Tree, K-Nearest Neighbors and Naïve Bayes.

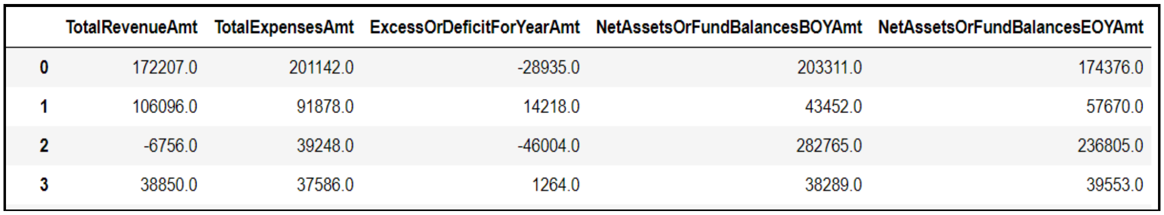
- **Decision Tree:** A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each brand represents an outcome of the test, class label is represented by each leaf node (or terminal node). Given a tuple X, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to the leaf node which holds the class prediction for the tuple. In this tree structure, leaves represent class labels and branches represent conjunction of features that lead to those class labels [4].
- **K-Nearest Neighbors:** Another set of algorithms consist on memorizing examples and comparing new observations to the ones in memory, like k-Nearest Neighbors [5], where the training examples are kept in the model, and when a new observation arrives the nearest k examples are used to vote its class or to average its expected value.
- **Naïve Bayes:** The Naïve Bayes algorithm uses the theorem of Bayes to compute the probability of a new example belonging to each class conditioned to its features, using the examples to compute the probabilities of each having a specific value on each feature according to each class [2].

These three algorithms will be tested to see which results in a higher percentage of accuracy in predicting the downfall of an organization. Depending on the resulting value, the best algorithm that best fits the data can be chosen.

As mentioned earlier a subset of the data that was relevant for the prediction was created as displayed below in Table 2. This subset contains the Total Revenue, Total Expenses, Excess or Deficit for Year, Net Assets or Fund Balances BOY and Net Assets or Fund Balances EOY. This data set contains the independent variables called features that will be used for the prediction.

Another subset was created that contains what is called the label which is the dependent value that will be produced when features have certain behavior. Table 3 contains a portion of that subset. To explain this relationship better, the row with index 0 in Table 2 will produce the value for index 0 in Table 3. For example, if an organization has

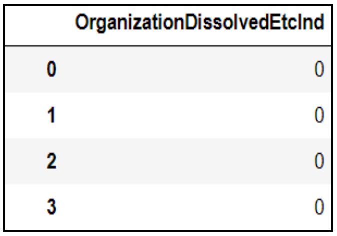
Table 2
Selected Features



	TotalRevenueAmt	TotalExpensesAmt	ExcessOrDeficitForYearAmt	NetAssetsOrFundBalancesBOYAmt	NetAssetsOrFundBalancesEOYAmt
0	172207.0	201142.0	-28935.0	203311.0	174376.0
1	106096.0	91878.0	14218.0	43452.0	57670.0
2	-6756.0	39248.0	-46004.0	282765.0	236805.0
3	38850.0	37586.0	1264.0	38289.0	39553.0

Total Revenues = $172,207, Total Expenses = $201,142, Excess or Deficit = -$28,935, Net Assets or Fund Balances BOY = $203,311 and Net Assets or Fund Balances EOY = $174,376, the organization will not be dissolved at the end of the year.

Table 3
Labels



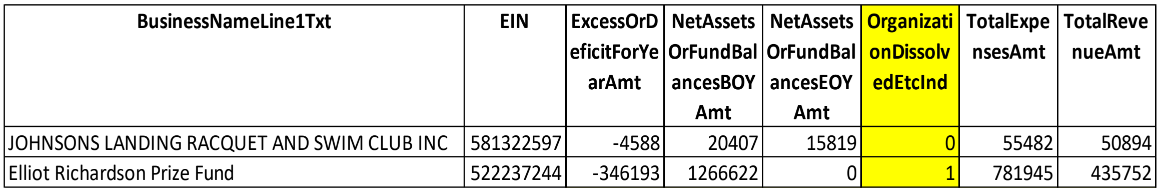
	OrganizationDissolvedEtcInd
0	0
1	0
2	0
3	0

All the data set of the 990ez form was used for the training and testing of the algorithms. A small sample of 2 organizations (Table 4), one known to be dissolved and one which is still operating, were used to determine if the algorithms predicted what was being looked for.

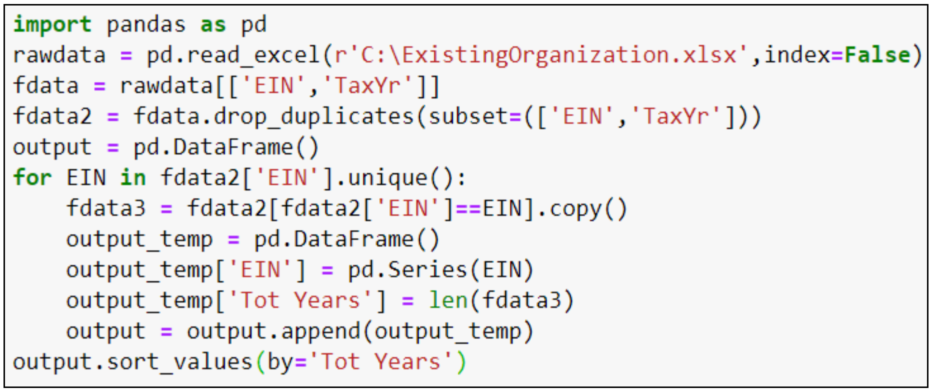
RESULTS

Using the algorithm in Figure 9 below allows to count how many years a specific organization existed. The same algorithm was run for the dissolved organizations and non-dissolved separately. Table 5 shows the results of this algorithm for organizations by using its EIN or Employer Identification Number which is unique for each organization. This way the amount of years an organization has existed can be known.

Table 4
Test Sample



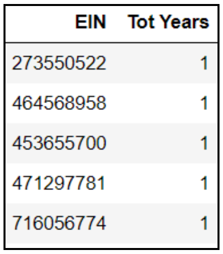
BusinessNameLine1Txt	EIN	ExcessOrDeficitForYearAmt	NetAssetsOrFundBalancesBOYAmt	NetAssetsOrFundBalancesEOYAmt	OrganizationDissolvedEtcInd	TotalExpensesAmt	TotalRevenueAmt
JOHNSONS LANDING RACQUET AND SWIM CLUB INC	581322597	-4588	20407	15819	0	55482	50894
Elliot Richardson Prize Fund	522237244	-346193	1266622	0	1	781945	435752



```
import pandas as pd
rawdata = pd.read_excel(r'C:\ExistingOrganization.xlsx',index=False)
fdata = rawdata[['EIN','TaxYr']]
fdata2 = fdata.drop_duplicates(subset=(['EIN','TaxYr']))
output = pd.DataFrame()
for EIN in fdata2['EIN'].unique():
    fdata3 = fdata2[fdata2['EIN']==EIN].copy()
    output_temp = pd.DataFrame()
    output_temp['EIN'] = pd.Series(EIN)
    output_temp['Tot Years'] = len(fdata3)
    output = output.append(output_temp)
output.sort_values(by='Tot Years')
```

Figure 9
Algorithm to Determine Year Count

Table 5
Example of the Amount of Years an Organization has Existed

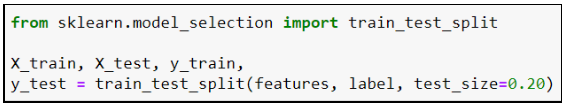


EIN	Tot Years
273550522	1
464568958	1
453655700	1
471297781	1
716056774	1

Using simple mathematics, the average years that a dissolved organization lasts and the average time a non-dissolved organization has lived can be determined. The total years are added together and then divided by the distinct count of organizations. For the dissolved organizations the numbers were a total of 8,599 organizations with a total of 8,841 years. Dividing the total amount of years by the total count of distinct organization gives an average of 1.03 years. For the non-dissolved organizations there is a total of 238,096 organizations with a total of 792,401 years, which results in a total average of 3.33 years. This means organizations that have less than 3 years are most likely to be dissolved and organizations that have existed for more than 3 years will remain existing.

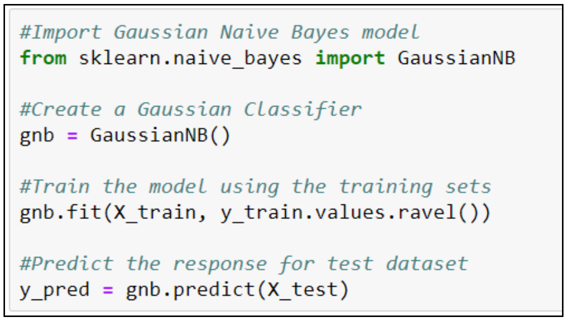
The data for the training and the testing was separated equally for the three algorithms, 80% of the data set was used for training and 20% for testing. The code in Figure 10 separated the data into training and testing.

First, the Naïve Bayes model was trained using the data split in Figure 10. The training was done using the code in Figure 11. Once the model is trained, its accuracy can be known using the test data. The accuracy of the model can be determined comparing the training data to the test data. The accuracy of the model resulted in 99.28% as shown in Figure 12. A case where the organization is still operating, and the data was not part of the data set used for the training was used next to see what would happen. As shown in Figure 13, the model accurately predicted that the organization was not dissolved as it was already known. Afterwards the same was done for an organization known to be dissolved (Figure 14). The algorithm was able to predict the status of the dissolved organization as expected. With the Naïve Bayes the 2 cases were predicted with success.



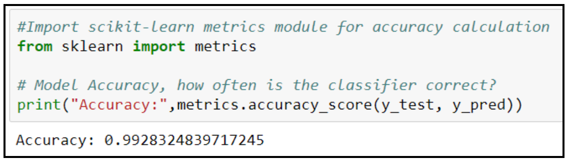
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train,
y_test = train_test_split(features, label, test_size=0.20)
```

Figure 10
Data Split Code for Training



```
#Import Gaussian Naive Bayes model
from sklearn.naive_bayes import GaussianNB

#Create a Gaussian Classifier
gnb = GaussianNB()

#Train the model using the training sets
gnb.fit(X_train, y_train.values.ravel())

#Predict the response for test dataset
y_pred = gnb.predict(X_test)
```

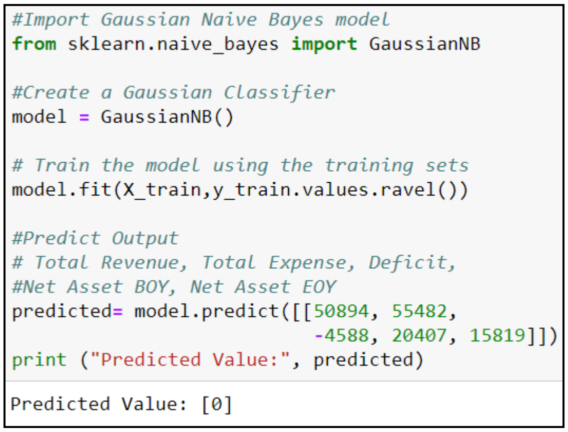
Figure 11
Naïve Bayes Training



```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.9928324839717245
```

Figure 12
Naïve Bayes Model Accuracy

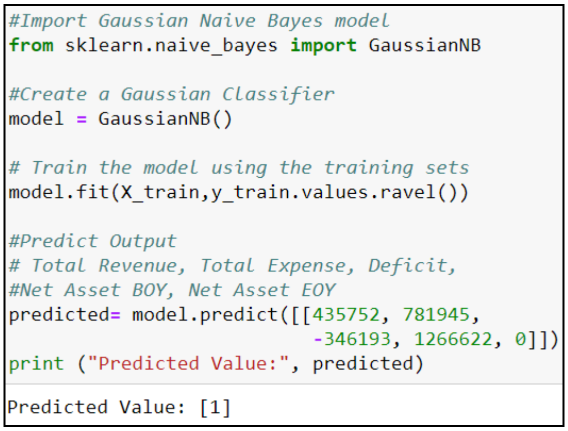


```
#Import Gaussian Naive Bayes model
from sklearn.naive_bayes import GaussianNB

#Create a Gaussian Classifier
model = GaussianNB()

# Train the model using the training sets
model.fit(X_train,y_train.values.ravel())

#Predict Output
# Total Revenue, Total Expense, Deficit,
#Net Asset BOY, Net Asset EOY
predicted= model.predict([[50894, 55482,
                           -4588, 20407, 15819]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [0]
```

Figure 13
Naïve Bayes Classifier Applied to Sample Data Set of Non-Dissolved Organization

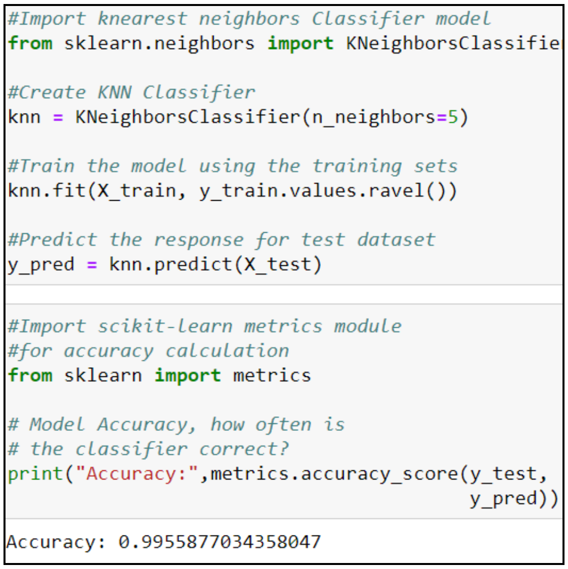
Next, the K-Nearest Neighbors model was trained using the data split in Figure 10. The training was done using the code in Figure 15. Once the model is trained, its accuracy can be known using the test data. Comparing the training data to the test data we can determine the accuracy of the model. The accuracy of the model resulted in 99.56 %. A case where the organization is still operating, and the data was not part of the data set used for the training was used next to see what will happen. As shown in Figure 16, the model accurately predicted that the organization was not dissolved, as was already known. Afterwards the same was done for an organization known to be dissolved. However, the algorithm was not able to predict the status of the dissolved organization as expected as show in Figure 17.



```
#Import Gaussian Naive Bayes model
from sklearn.naive_bayes import GaussianNB

#Create a Gaussian Classifier
model = GaussianNB()

# Train the model using the training sets
model.fit(X_train,y_train.values.ravel())

#Predict Output
# Total Revenue, Total Expense, Deficit,
#Net Asset BOY, Net Asset EOY
predicted= model.predict([[435752, 781945,
                           -346193, 1266622, 0]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [1]
```

Figure 14
Naïve Bayes Classifier Applied to Sample Data Set of a Dissolved Organization

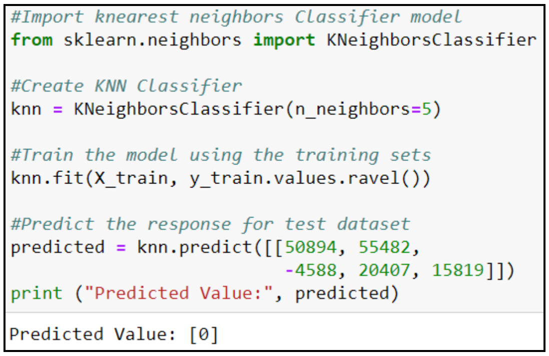


```
#Import knearest neighbors Classifier model
from sklearn.neighbors import KNeighborsClassifie

#Create KNN Classifier
knn = KNeighborsClassifier(n_neighbors=5)

#Train the model using the training sets
knn.fit(X_train, y_train.values.ravel())

#Predict the response for test dataset
y_pred = knn.predict(X_test)
```

```
#Import scikit-learn metrics module
#for accuracy calculation
from sklearn import metrics

# Model Accuracy, how often is
# the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test,
                                          y_pred))
```

```
Accuracy: 0.9955877034358047
```

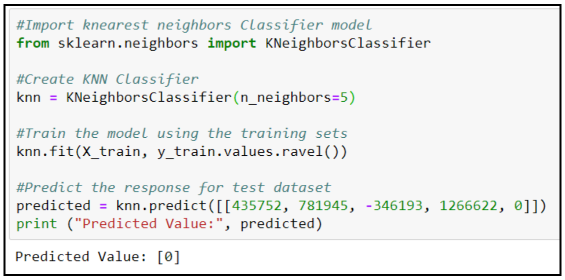
Figure 15
K-Nearest Neighbors Training and Model Accuracy



```
#Import knearest neighbors Classifier model
from sklearn.neighbors import KNeighborsClassifier

#Create KNN Classifier
knn = KNeighborsClassifier(n_neighbors=5)

#Train the model using the training sets
knn.fit(X_train, y_train.values.ravel())

#Predict the response for test dataset
predicted = knn.predict([[50894, 55482,
                         -4588, 20407, 15819]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [0]
```

Figure 16
K-Nearest Neighbor Classifier Applied to Sample Data Set of Non-Dissolved Organization

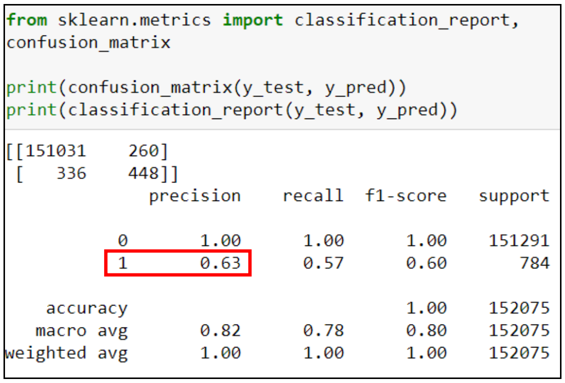


```
#Import knearest neighbors Classifier model
from sklearn.neighbors import KNeighborsClassifier

#Create KNN Classifier
knn = KNeighborsClassifier(n_neighbors=5)

#Train the model using the training sets
knn.fit(X_train, y_train.values.ravel())

#Predict the response for test dataset
predicted = knn.predict([[435752, 781945, -346193, 1266622, 0]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [0]
```

Figure 17
K-Nearest Neighbor Classifier Applied to Sample Data Set for Dissolved Organization

For this algorithm in particular, it is difficult to determine how many neighbors are required for the algorithm to accurately predict the failure of an organization. Figure 18 below demonstrates the algorithms precision in predicting this. It shows a 100% precision for non-dissolved organizations and a 63% precision for the dissolved one.

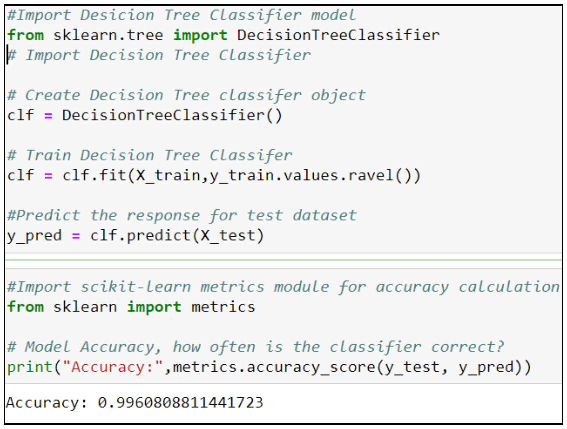


```
from sklearn.metrics import classification_report,
confusion_matrix

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[151031    260]
 [   336    448]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    151291
           1       0.63      0.57      0.60       784

    accuracy                           1.00    152075
   macro avg       0.82      0.78      0.80    152075
weighted avg       1.00      1.00      1.00    152075
```

Figure 18
K-Nearest Neighbors Model Precision

Last, the Decision Tree model was trained using the data split in Figure 10. The training was done using the code in Figure 19. Once the model is trained, its accuracy can be known using the test data. Comparing the training data to the test data the accuracy of the model can be determined. The accuracy of the model resulted in 99.61%. Next, a case where the organization was known to be still operating and the data was not part of the data set used for the training was used to see what would happen. As shown in Figure 20, the model accurately predicted that the organization was not dissolved as was already known. Afterwards the same was done for an organization known to be dissolved (Figure 21). The algorithm was able to predict the status of the dissolved organization as expected. With the Decision Tree the 2 cases were also predicted with success.



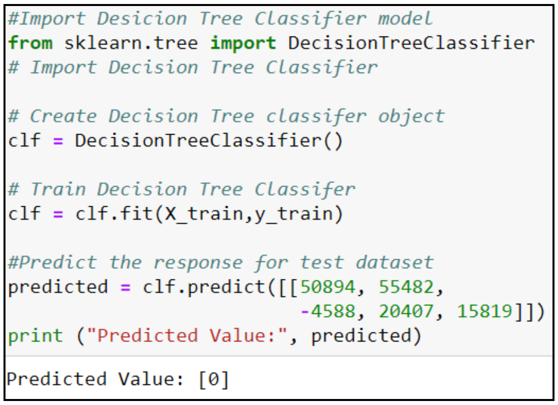
```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train.values.ravel())

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.9960808811441723
```

Figure 19

Decision Tree Training and Model Accuracy

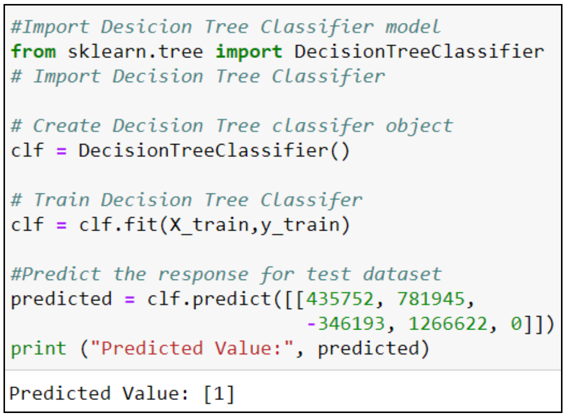


```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
predicted = clf.predict([[50894, 55482,
                          -4588, 20407, 15819]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [0]
```

Figure 20

Decision Tree Classifier Applied to Sample Data Set of Non-Dissolved Organization



```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
predicted = clf.predict([[435752, 781945,
                          -346193, 1266622, 0]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [1]
```

Figure 21

Decision Tree Classifier Applied to Sample Data Set for Dissolved Organization

CONCLUSION

It was proved throughout this work that Machine Learning can be used for predicting behaviors in different fields if there is sufficient data. In the majority of the cases the data needs to be parsed and wrangled to be able to use tools and see trends. Predicting the downfall of a Non-Profit Organization was possible using certain attributes: Total Revenue, Total Expenses, Excess or Deficit for Year, Net Assets or Fund Balances BOY and Net Assets or Fun Balances EOY. There are classifiers and regression models that are more suitable for a determined type of data. As showed here the Naïve Bayes, K-Nearest Neighbors and Decision Tree classifiers resulted in different accuracies. In this case, the Decision Tree was the most accurate at predicting if the organization was going to be dissolved or not with a 99.61%. It is important to point out that the precision of the models will depend on the amount of data that is used for the training and how clean the data is. The K-Nearest Neighbors classifier will not be the best predictor for this type of data since varying the number of neighbors has little to no effect on improving the outcome for the dissolved scenario. On average, organizations that are dissolved only last 1.03 years, on the other hand organizations that keep running last more than 3.33 years.

FUTURE WORK

The work can be improved by expanding the scope of the study to include all IRS tax forms including 990 and 990pf instead of only using the 990ez form. This will need the use of a cluster since the amount of data available to sort and analyze exceeds the capabilities of a personal computer and thus will not have the processing capacity to complete this analysis in a reasonable time. It took 5 days to process only the 990ez form data without all the attributes available. In scenarios like this, in order to make the process easier, it would be best to employ an execution framework such Apache Hadoop, Spark, Tensor Flow or Azure-ML. Also, more attributes can be used in the analysis to determine if an organization will be dissolved or not instead of using the attribute selected in this work.

The current work determines if the organization was going to be dissolved or not using the finances in a determined year. Future work can include a functionality that will determine in what moment the organization will be dissolved. For example, if the current algorithm determined that an organization was not going to be dissolved now, it can be modified to specify if an organization continues the same trend whether it will be dissolved and in how many years.

REFERENCES

[1] S. Athmaja, M. Hanumanthappa and V. Kavitha, "A survey of machine learning algorithms for big data analytics," *2017 International Conference on Innovations in Information,* Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-4. DOI: 10.1109/ICIIECS.2017.8276028

[2] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", *18th International Conf. on Transparent Optical Networks*, pp. 1-4, 2016.

[3] J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng, "A survey of machine learning for big data processing", *EURASIP Journal on Advances in Signal Processing*, Springer, vol. 2016:67, pp. 1-16,2016. DOI: 10.1186/s13634-016-0355-x.

[4] K. Sunil and S. Himani, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research*, vol. 5, no. 4, pp. 2094-2097, 2016.

[5] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.