الجامعة الإسلامية العالمية ماليزيا

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

يونيورسيتي إسلام انتارابغسا ملايسيا

Kulliyyah of Information and Communications Technology
Department of Computer Science
International Islamic University Malaysia

Semester II 2016/2017

PRINCIPLES OF ARTIFICIAL INTELLIGENCE
CSC 3301
SECTION 2
LECTURER'S NAME: DR AMELIA RITAHANI BINTI ISMAIL

CIGARETTE CONSUMPTION CLASSIFICATION USING BACKPROPAGATION ALGORITHM

GROUP PROJECT ASSESSMENT #3

NAME	MATRIC NUMBER
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

SUBMISSION DATE: 17TH MAY 2017

Table of Contents

1.0 Introduction

Machine learning is one of the many types of artificial intelligence. It works by providing a computer the ability to learn without being programmed. This is also known as self-learning. When they are exposed to new data, the computer program will learn about the data and later will change according to what the machine was learning previously. This machine learning algorithm manages to learn and make a prediction based on the data given. Cigarette Consumption is the dataset that has been selected. This project is about how the cigarette consumption among people depends on the minimum price in adjoining states per pack of cigarettes from year 1963 until 1992.

2.0 Project Objectives

- ✓ To collect the data about cigarette consumption.
- ✓ To study the dataset from year 1963 until 1992 about cigarette consumption.
- ✓ To predict the actual output of cigarette consumption and the actual minimum price in adjoining states per pack of cigarettes.
- ✓ To analyze the data on how the minimum price in adjoining states per pack of cigarettes will affect the cigarette consumption.

3.0 Expected Output/Results

Based on the Cigarette Consumption dataset, we will analyze and visualize the dataset by using backpropagation algorithm. After evaluating the model, we expect the get the evaluation results that consist of true negative (TN), true positive (TP), false negative (FN) and false positive (FP). The classifiers that predict correctly are called true positives (TP) and true negatives (TN), respectively. Similarly, the incorrectly classified instances are called false positives (FP) and false negatives (FN). The accuracy of the model is also to be expected. With that, we can predict the minimum price of cigarette per state for the next year based on the results.

4.0 Literature Review

4.1 Cigarette Consumption and Its Causes

Cigarette consumption is becoming an issue more and more every day. Nowadays, cigarettes are not only consumed by adults, but by young boys and girls as well. In order to know more about cigarette consumption, we first need to know what is cigarette. According to Cambridge Dictionary, cigarette gives a meaning “a small paper tube filled with cut pieces of tobacco that people smoke”. Also in Cambridge Dictionary, tobacco is “a substance smoked in cigarettes, pipes, etc. that is prepared from the dried leaves of a particular plant”. It is found that the basic components of most cigarettes are tobacco, chemical additives, a filter, and a paper wrapping. Chemical additives here include nicotine which is a type of drug which acts as an addictive agent in cigarettes [1]. Studies shows that nicotine changes the way your brain works and causes people to crave more and more [2]. Therefore, cigarette consumption causes both physical and mental addiction.

Currently cigarette consumption by women exceeds men, in which a lot of health-related problems arise. These health problems do not only limit to the smoker which in this case women, but others as well. Others here include new born babies where they could develop health defects if their mother was smoking during pregnancy. Another issue is that kids nowadays start consuming cigarette from a very young age. Adolescence ages 15 – 18 are also among the heaviest to consume cigarettes [3]. This is very dangerous especially for kids as smoke-related diseases does not only affected by how they smoke, but also by how long they smoke. As kids start smoking at an early age, they will likely to smoke the longest.

There are several causes in what leads to cigarette consumption. Most would say that peer influence is the biggest cause. While this is true, there are other causes that we would never have comprehended. It is found that cigarette advertisements lead to sudden interest

towards smoking, which caused affected people to seek out information about smoking [4]. This sudden interest does not only peak by cigarette promotion advertisements, but also advertisements about anti-smoking. Most of these affected people primarily do not have any intentions into divulging themselves to cigarette consumption, but the psychological effect from the sudden interest based on humans' curiosity gives an unintended result.

Another cause of cigarette consumption is childhood experience. As most smokers start smoking from a very young age, it is possible that bad childhood experience would cause this to happen. Consuming cigarette will likely be a coping mechanism for most adolescent that is having a rough experience growing up. Therefore, it is important to prevent adverse childhood experiences and improving treatment of exposed children in order to reduce smoking, not only among adolescents but also adults [5]. If this is made possible, then overall cigarette consumption may slowly be reduced.

4.2 Machine Learning – Backpropagation Algorithm

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web searches, and a vastly improved understanding of the human genome [6]. And nowadays, with the increase of effective machine learning technique many people probably use it dozens of times a day without knowing it. Machine learning is the field that concentrates on induction algorithms and on other algorithms that can be said to "learn", usually through datasets.

Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data [7]. Unsupervised algorithms can draw inferences from datasets [8]. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. There are tens of thousands of machine learning algorithm. And the algorithm that we use for our machine learning is backpropagation algorithm.

The backpropagation algorithm is a supervised learning method for multilayer feed-forward networks from the field of Artificial Neural Networks [9]. The principle of the backpropagation approach is to model a given function by modifying internal weightings of input signals to produce an expected output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state. Technically, the backpropagation algorithm is a method for training the weights in a multilayer feed-forward neural network. As such, it requires a network structure to be defined of one or more layers where one layer is fully connected to the next layer. A standard network structure is one input layer, one hidden layer, and one output layer.

4.3 Current Issues

A lot of people all around the world categorized as active smokers. No matter what kind of advertisements that we use to prevent or to discourage them from being a smoker, they do not work at all. Cigarette consumption or we also refer as smoking is one of the unhealthy habits that most of the people around the world could not manage to avoid.

We can see that not only men tend to smoke but also women. Even though we make a rule that at age above 18-year old is the legal age to smoke but there are also teenagers that already start to smoke during their elementary school. This issue is a concern to parents especially as they are exposed at such an early age. For teenagers, they usually got influenced by their friends who smoke. Even though they do not have money, they tend to get ask from their parents and in extreme cases, lie to in order to get it. The percentage of the dataset for teenagers that are smoking under 16 years old is quite dangerous. In school, maybe the teacher can give them punishments if they are caught smoking in school. But at home, most of the parents are only at home late at night as they are busy with work, neglecting their children in the process.

Smoking is addicting to those who consume it, the same as taking drugs and having sexual intercourse. It is not only that smoking will harm the smoker but the consequences are far worst for the second-hand smokers. There cases reported that a baby died just a couple of days after he or she has been born. Usually, people think that if they are not smoking in front of others, then the rest will be okay but they forget about the scent of the cigarette is still on their shirt and that later on would have direct contact with other people, in this case the baby. It will cause the new-born to get a disease that is far worse which will lead to death as the immune system of the baby is still weak. Even adults who are non-smokers also suffer if they get in touch with active smokers for a long time.

For the smoker itself, they only think that the best way to solve problems is by smoking, but the truth is they are only calling for a disease. If they are a heavy smoker, they do not have to wait for a long time before the cigarette affected their body system. Some of them are having lung cancer, asthma, unpleasant smell on their mouth and stroke [10]. Some of the diseases can be treated but some are not. For smokers who are pregnant, their habits will affect their infants. Their baby will get diseases or they will become disable because they were exposed to the smell of cigarette from they are still in the womb of their mothers. This case does not only happen if the mother is a smoker but also if the father is a smoker as the mother has been exposed to it. For the worst-case scenario, some infants cannot even manage to survive and die during pregnancy.

If the smokers save their money from buying the cigarette, they can save a lot and that money can be used for emergency. They do not realize the value of money until they stop smoking and saving up the money that they usually spend to buy cigarettes. When they finally calculate all the money that they already save, they will be surprised with the amount that they usually spend only to buy cigarettes.

Smoking also will pollute the environment [11]. If the demands for cigarette increase, the government or any company will cut the trees to plant the tobacco. It will cause global warming and land slide. They only think about the profits but they forget about the effects on what they are currently doing.

4.4 Cigarette Consumption and Its Effects

Smoking in our community has been very common nowadays because of the addictive nature of the product. An amount of studies has contributed in providing the evidences of how far tobacco-use gives impact to the users. According to some other researches, majority of the countries have shown an increase in the rate of cigarette consumption based on cigarette sales per capita. However, the increase in price per pack which is depicted from the minimum price in adjoining states per pack of cigarettes does not affect them at all.

The first reason why this analysis has been done is to reduce the cigarette consumption rate; at least to a certain amount, by applying taxes for every single pack of cigarette [12] otherwise the benefit of the economic growth will be slowly erased. For instance, the tax should constitute approximately 70%-80% of the total price, or the tax should be increased regularly to keep price with inflation in order to control the tobacco consumption.

At the same time, this study helps the smokers or cigarette consumers to save income and their personal finances. It is also contributing to the economic stability because less income to spend on tobacco especially the poor people and 16 years old unemployed smoker [13]. Since this kind of people are more responsive to price, they may actually save money from buying cigarettes.

The annual number of deaths attributed to cigarettes, allowing this study to prove that use of tobacco has great effect on the smokers and also the public health [14]. There are hundreds of thousands of people die each year from smoking-related diseases thus more stop-smoking efforts and anti-tobacco education needed for population with higher rates of smoking in order to cut down the number of deaths every year.

5.0 Data Descriptions

Our dataset is about the cigarette consumption in the United State. [15] It consists of 1380 number of observations from the year 1963 to 1992 from over 50 states in the United States. There is a total of ten columns as part of the attributes. They include column 0, state, year, price per pack of cigarettes, overall population, population above the age of 16, consumer price index, per capita disposable income, cigarette sales in packs per capita and minimum price in adjoining per pack of cigarettes. The type of the dataset is supervised learning because the data of the class is known and provided in the training phase as well as the desired output.

6.0 Experimental Setup

6.1 Data Preprocessing

Based on the dataset being used, it is usual to have incomplete data. There may also have redundant data that could affect the quality of the mining results. Therefore, it is important to have the data preprocessed in order to have a quality dataset. We use Microsoft Azure for our data preprocessing. Figure 1 shows the steps taken for the data preprocessing.

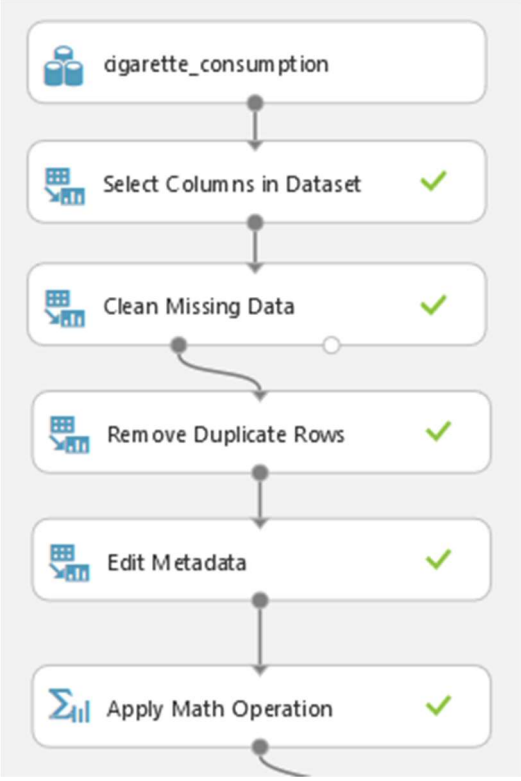

Figure 1: Steps for Data Preprocessing

For our dataset, which is about cigarette consumption, there are originally 1380 rows with 10 columns in total. Unused columns are removed, and in this case, only one column is removed which is the 'Column 0' as it does not have any importance to the dataset. After removing, there are nine columns left. Cleaning missing data is done next. Incomplete data

are removed during the process. However, there are no changes to the number of rows which means that there are no missing data in the dataset. Duplicate rows are also removed. Same as previous process, there are no changes to the number of rows as there are no redundancies of data in the dataset. Next, the column names are renamed for easier observation. Finally, a calculated column is created using (Ln) function to create a balanced logarithmic distribution values to a predictive model. Table 1 and 2 shows the preprocessed data.

state	year	price	pop	pop16	consumerpriceindex	percapitadispincome	sales

Table 1: Preprocessed Data

minpricestate	Ln(minpricestate)

Table 2: Continuation of Preprocessed Data

6.2 Data Transformation

After finishing with data preprocessing, the data will undergo transformation, in which the data is split into training set and testing set. For this project, the training set is assigned to be 70% of the data set whereas the testing set is 30%. It is important to split the data into training and testing sets in order to have an accurate prediction of data. The larger the percentage of training set, the higher the accuracy of the predictions. A training model is selected to train the training set and that for this project; Two-Class Decision Forest model is used. Decision forests are known to be fast and supervised ensemble model [16]. With this, the training set is trained and learned by using decision forest with backpropagation algorithm. After finishing with training data, the testing set is tested in Score Model where the data is visualized after it finished learning. The data is then visualized using Evaluate Model to check its accuracy, the true positives and negatives, and also the false positives and negatives. Figure 2 below shows the steps taken for data transformation:

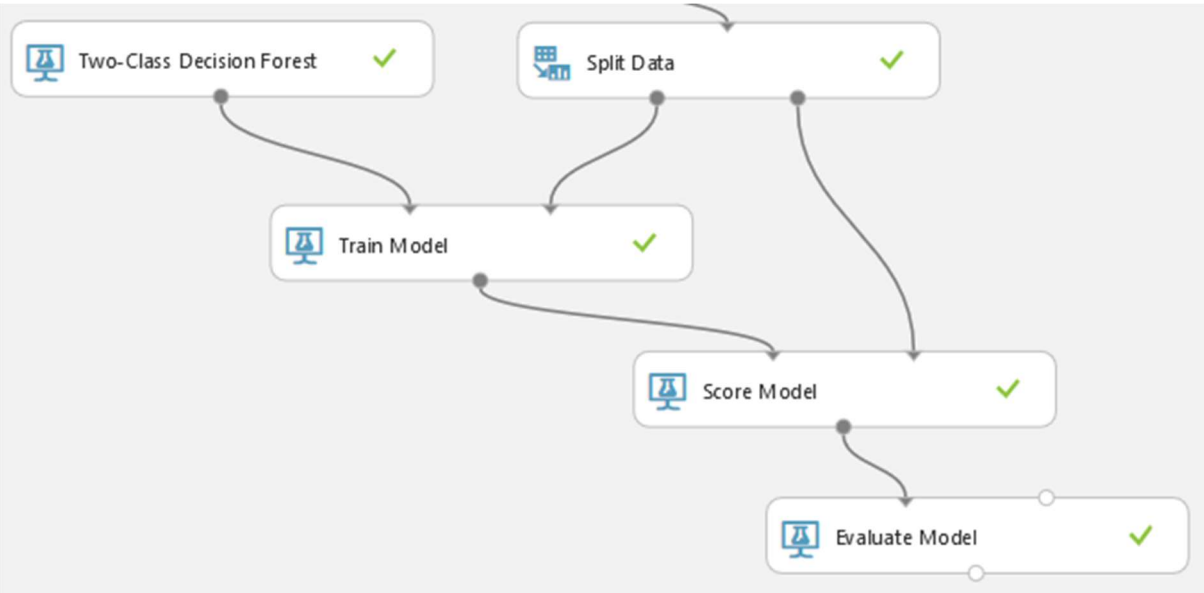

Figure 2: Steps for Data Transformation

7.0 Results

Evaluate Model is used to evaluate the result of the experiment. Moreover, this model outputs a confusion matrix that shows the numbers of true positive (TP), false positive (FP), true negative (TN) and false negative (FN), as well as positive label, and negative label. The binary classification model has Accuracy, Precision, Recall, F1 score and also Area under Curve (AUC) for their evaluation metrics [17]. Table 3 below shows the binary classification confusion matrix of the dataset:

	Predicted	
	Positive	**Negative**
Actual True	412	0
Actual False	2	0

Table 3: Binary Classification Confusion Matrix

Accuracy is simply the proportion of correctly classified instances and it is the first value we will look after evaluating the model of this experiment. It deals only with ones and zeros. **Threshold** represents the trade-off between false positive and false negative. The most natural threshold is 0.5. In this experiment, we get accuracy near to 1 which is 0.995. In addition, the class labels can only take the maximum up to 2 values that we will consider as positive and negative, which in this experiment, the value for positive label is 24.0 and negative label is 23.9.

If anyone asks about "How many were classified correctly (TP), based on the minimum price in states per pack of cigarette?" which the answer would be best when we look at **Precision** that defines better as the proportion of positives that is classified correctly. Another metrics are **Recall** which is the true positive rate and also **F1 Score** which takes recall and precision into consideration to summarize the evaluation in a single number.

Moreover, the curve of Receiver Operating Characteristic (ROC) that contains the true positive rate vs. the false positive rate and the corresponding **AUC** can be inspected. If the curve is nearly to the upper left corner that means the classifier's performance is getting better. Table 4 below shows the result after the dataset is tested.

Positive Label	Negative Label	Accuracy	Precision	Recall	F1 Score
24.0	23.9	0.995	0.995	1.000	0.998

Table 4: Result of the Data

Figure 3 below shows the graph of true positive rate with false positive rate:

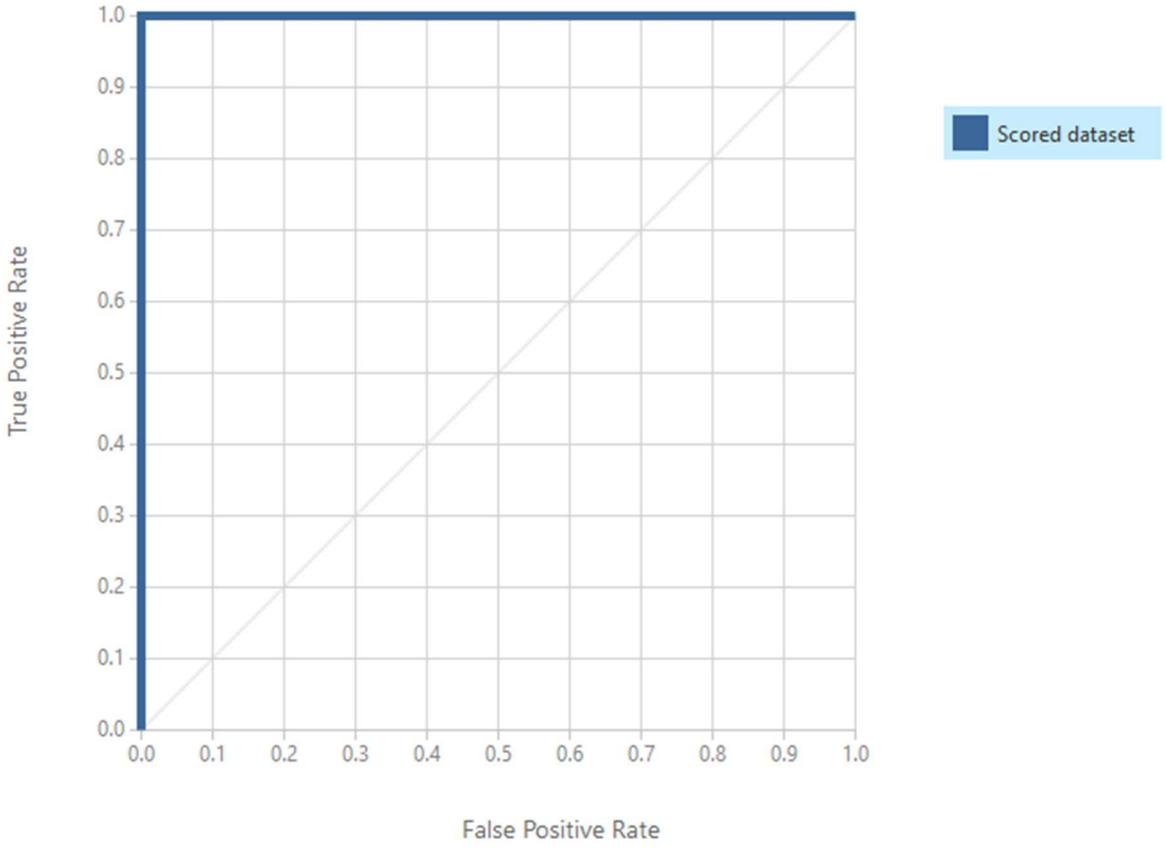

Figure 3: Graph of True Positive Rate with False Positive Rate

Figure 4 below shows the actual result from the evaluate model based on the cigarette consumption dataset:

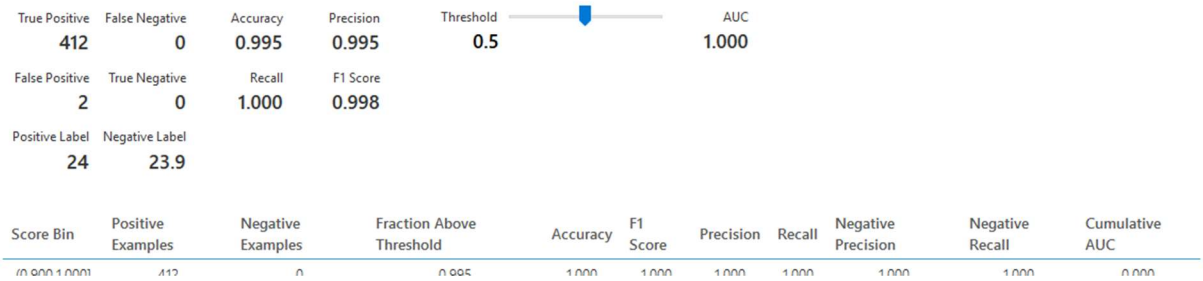

Figure 4: Visualization of Evaluate Model

8.0 References

[1] Eriksen, M. (2015). *The tobacco atlas* (5th ed.) [5]. Retrieved April 17, 2017, from http://www.tobaccoatlas.org/

[2] U. S. Department of Health and Human Services (USDHHS). A Report of the Surgeon General: How Tobacco Smoke Causes Disease: What It Means to You (Consumer Booklet). Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2010.

[3] HHS, Preventing Tobacco Use Among Young People: A Report of the Surgeon General, 1994.

[4] Pechmann, C., & Knight, S. J. (2002, June 01). An Experimental Investigation of the Joint Effects of Advertising and Peers on Adolescents' Beliefs and Intentions about Cigarette Consumption. Retrieved April 16, 2017, from https://academic.oup.com/jcr/article-abstract/29/1/5/1796214/An-Experimental-Investigation-of-the-Joint-Effects

[5] MS, R. F. (1999, November 03). Adverse Childhood Experiences and Smoking During Adolescence and Adulthood. Retrieved April 16, 2017, from https://jamanetwork.com/journals/jama/fullarticle/192056

[6] Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., . . . Zieba, K. (n.d.). Self-driving cars. *AccessScience*. doi:10.1036/1097-8542.br0326141

[7] Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks, 6*(4), 525-533. doi:10.1016/s0893-6080(05)80056-5

[8] Bellegarda, J. R. (2010). Unsupervised document clustering using multi-resolution latent semantic density analysis. *2010 IEEE International Workshop on Machine Learning for Signal Processing*. doi:10.1109/mlsp.2010.5587982

[9] Stoyanov, I. (1996). An improved backpropagation neural network learning. *Proceedings of 13th International Conference on Pattern Recognition*. doi:10.1109/icpr.1996.547632

[10] Nandini, K. (2014, September 22). Top 10 Dangerous Negative Effects of Smoking on Health. Retrieved April 19, 2017, from http://listovative.com/dangerous-negative-effects-of-smoking-on-health/

[11] Bahaya Merokok 5 Kesan Yang Ramai Tak Tahu (n.d.). Retrieved April 19, 2017, from http://www.healthmassa.com/bahaya-merokok-5-kesan-yang-perokok-tak-tahu/

[12] Callison, K., & Kaestner, R. (2013). Do Higher Tobacco Taxes Reduce Adult Smoking? New Evidence of The Effect of Recent Cigarette Tax Increases on Adult Smoking. *Economic Inquiry, 52*(1), 42-46. doi:10.1111/ecin.12027

[13] Michael, P. (n.d.). Are Cigarette Excise Taxes Effective in Reducing the Habit? The Impact of Income versus Price on the Percentage of Adults Who Smoke (n.d.). Retrieved April 17, 2017, from https://www.american.edu/spa/publicpurpose/upload/2011-Public-Purpose-Cigarette-Taxes.pdf

[14] European Commission. (2017, April 07). Retrieved April 18, 2017, from http://ec.europa.eu/health/sites/health/files/tobacco/docs

[15] Cigarette Consumption. (n.d.). Retrieved April 12, 2017, from https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Cigar.html

[16] Two-Class Decision Forest. (2016, June 9). Retrieved May 14, 2017, from https://msdn.microsoft.com/en-us/library/azure/dn906008.aspx

[17] Gary Ericson, Seokjin Han, Larry Franks, Neeraj Khanchandani and Brad Severtson. 2017. How to evaluate model performance in Azure Machine Learning. *Evaluating a Binary Classification Model: Inspecting the Evaluation Results*. Retrieved from website: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-evaluate-model-performance