The links between smoking and education

Business/research questions

Smoking is one of the leading reasons for the development of many diseases and death around the world. Cigarette smoking is responsible for more than 480,000 deaths per year in the United States [1]. Smoking has a huge impact on the US economy. The total economic cost of smoking in the US is estimated at more than $300 billion a year. This includes nearly $170 billion in direct medical care and more than $156 billion in lost productivity due to premature death and exposure to secondhand smoke [2]. In addition, thousands of young people start smoking cigarettes every day [3].

A new Yale study shows that the links between smoking and education in adulthood are explained by characteristics and choices made in adolescence [4]. Smokers often begin tobacco use in early adulthood and initiate quitting after middle age. Therefore, it is thought that education has a profound effect on smoking initiation in the younger population, while occupation can affect smoking cessation in the middle-aged and older population [5]. In this research, we aim to predict a correlation between smoking, the level of education, and income. It is shown that routine checkups can help in preventing diseases from getting chronic. Medical professionals advise an annual health check-up; we want to find if education or income affects how often people go for checkups.

We want to find answers to these three research questions:

Research question one: Does the higher the level of education have any impact on smoking?
Research question two: Is there any link between smoking and the level of income?
Research question three: Does the level of education or how much a person earns affect how often they go for a checkup?

Processing the data

The Behavioral Risk Factor Surveillance Survey data [6] includes 414 attributes constituting and monitors changeable risk behaviors and other factors that contribute to and cause morbidity and mortality in the population. The data set represents the non-institutionalized adult household population aged 18 years and older [6].

We studied data in Weka and Colab; we looked at different attributes present in the dataset, what values they contained, how many values a particular column can have, and applied preset algorithms such as decision tree J48, Logistic, and mulitiLayerPerceptron on selected attributes, closely monitored correctly and incorrectly classified instances. Furthermore, for more evaluation, we looked at true and false-positive rates, recalls, and F-Measures. We find out that many of the attributes were redundant in nature, like the attribute state, since it only had one value: ‘New York.’ Attribute DIABAGE2 had 90% of its value as Not asked or Missed, Refused, so attributes like this can introduce imbalance if we use it with other attributes. Some of the attributes were repeated with no change in data, so we think that 414 attributes can be brought down without affecting the dataset’s quality. Because of this, we suggest anyone working with this dataset spend a significant amount of time cleaning up data. We also found that cross-validation works better than splitting datasets into

training and testing when performing classification tasks. To find a correlation, we would suggest using the whole dataset or splitting it in training and testing rather than cross-validation because this introduces unnecessary complexities.

Data cleaning

We imported the given dataset in weka and dropped all the columns except income, education, smoker, and checkup to avoid unnecessary complications and speed up the process of training and testing as fewer data takes less time to train. Still, we had to make sure we didn't remove anything important. After that, we applied filters in weka to clean unwanted observations. Next, we dropped rows with missing values as we wouldn't be able to use the dataset without that. Then we exported that dataset to Colab for further cleaning. We removed rows containing values such as *"Don't Know, Missing, Not Sure, and Refused*" since these values don't provide any new information and can affect the accuracy of the model on which we train it. We also removed duplicates when we imported them back in weka as duplicates won't give us any new information, and if there are many exact duplicates, it may add bias if not removed, irrelevant, structural errors deleted. To answer the research questions, we imported the cleaned data set to Weka. After importing to Weka, we needed to convert attributes that were categorical in nature to binary values by creating dummy columns because machine learning models cannot directly work with categorical data, so we converted them to numerical data using dummy columns[7] for this, we used nominal to binary filter present in weka this converted education, checkup, and income into binary. To answer each question, we created individual arff files.

Data analysis

To answer the questions we had to find if there is a correlation between our attributes, we have used correlation because Correlation analysis is an extensively used technique that identifies interesting relationships in data. These relationships help us realize the relevance of attributes with respect to the target class to be predicted [8].

We tried to find a correlation value between education and smoking on our cleaned dataset using weka for the first question. Then we used different classification models present in weka like J48, Naive Bayes to find out if education can classify if a person is likely to smoke or not.

For the second question, we followed a similar approach. We started by finding the correlation between income and smoking and then used different classification models like J48, Naive Bayes to find out if income can classify whether a person is a smoker or not.

To find out if the level of education and income can have an effect on how often a person goes for a checkup, we started with correlation evaluation; after that, we used different classification models such as Naive Bayes, Logistic, Multilayer perceptron to find out if income and education can classify if the person goes for checkup often or not.

Research questions number one and two had similar results when we classified them using decision-tree, Naive Bayes, J48.

The result for research question number two

Correctly Classified Instances	8616	86.1772 %
Incorrectly Classified Instances	1382	13.8228 %

Reviewing the other metrics such as True and False Positive rates, Precision, Recall, and F-Measure contrasts with the level of accuracy; therefore, we can not determine if there is a link between smoking and the level of income. Reviewing data, we can confirm we are working with an Imbalanced data set to classify these problems because the class distribution is not uniform.

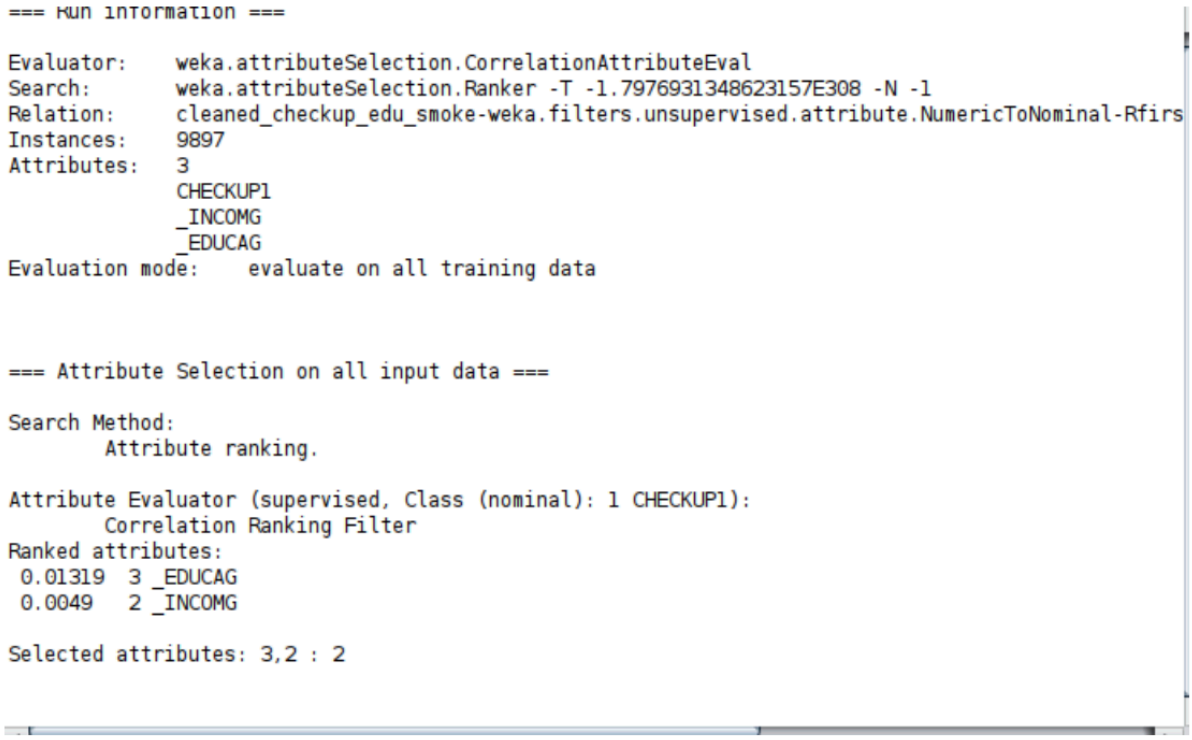
Result of correlation between education, income, and smoking

```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     cleaned_checkup_edu_smoke-weka.filters.unsupervised.attribute.NumericToNominal-Rfirs
Instances:    9897
Attributes:   3
              _INCOMG
              _EDUCAG
              _SMOKER3
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 3 _SMOKER3):
        Correlation Ranking Filter
Ranked attributes:
 0.123  2 _EDUCAG
 0.112  1 _INCOMG

Selected attributes: 2,1 : 2
```

As it is shown in the analysis, we can see that the correlation value is very small so we can say that neither education nor income correlates with smoking, so there are other factors like peers, the environment in which a person grows, or some other factors that determine if someone will be a smoker or not more research needs to be done in this.

For Research question three, we had to find the relation between education, income, and checkup. We followed a similar procedure. Even though we got an accuracy of around 70%, if we take a closer look at the confusion matrix, we can easily see that it is due to an imbalanced dataset. We cannot conclude if there is any relationship between them.

Correlation between Education, Income, and Checkup:



Correlation is also very low in this case, so we can say that education and income don't determine how often a person will go for checkups.

Summary

The summary results can not confirm a positive direct correlation between education, income, and smoking habits; there seems to be a weak correlation. A possible solution to these questions requires more in-depth analysis, and a new set of data must be collected to achieve more accuracy. Since the dataset was imbalanced 85% non-smokers so even if the classifier classifies everyone as a non-smoker, it will get 85 % accuracy because of this, some may think there is some relationship, but a closer look at other metrics like precision, recall F-measure will reveal that is not the case.

The Education attribute was better balanced in terms of the normal distribution of values. There is no correlation between education and income with checkups, health, and income on how often a person goes for a checkup. The takeaway from this research is there are no clear answers. Data analysis supports that smokers are found in all sections of society highly educated or illiterate, people earning above 50K, or below 20K, We can say that neither education nor income correlates with smoking, so there are other factors like peers, the environment in which a person grows or some other factors that determine if someone will be a smoker or not more research needs to be done in this.

The same is the case with health checkups; highly educated people and not so highly educated people also tend to visit the same number of times for checkups. It is the same case for income, and these questions can be analyzed from a different angle to determine a higher level of accuracy. One of the solutions could be to add more statistical attributes. We can also experiment with oversampling and generating synthetic datasets.

Research question one: Negative Relationship
Research question two: NegativeRelationship
Research question three: Negative Relationship

It is very hard to find an inevitable result from analyzing data. There are always many doubts. We believe in ethical and honest analysis. Therefore, dealing with an imbalanced data set raises the question of how much of an imbalanced data set gap can be attributed to the false prediction. In addition, the attributes which chose to make predictions have no scientific value compared to analysis based on biological tests such as blood pressure, blood sugar, an x-ray of lunges, to name a few. Maturation could be another threat to our analysis, considering the data set is not up-to-date and might not represent reality. There is always an internal validity that questions if our analysis is flawless. From an external validity point, our analysis might result from some artifact of the data we used. from the construct validity, we can question our methods and the choice of the attributes we used for analysis. These characteristics are very important to our analysis. Internal validity makes the conclusions of a causal relationship credible and trustworthy. Without high internal validity, an experiment cannot demonstrate a causal link between two variables[9]. From the internal validity, we can't confirm if our treatment and response variables changed simultaneously or if treatment precedes changes in response to variables.

“Threats to validity are not something that you can escape. What you have to do is reduce them as far as is practical, verify whether the residual threat is tolerable, and be appropriately honest about that residual threat when reporting your results[10].”

Dealing with large data sets

A relational database is a type of database that stores and provides access to data points related to one another. Relational databases are based on the relational model, an intuitive, straightforward way of representing data in tables. In a relational database, each row in the table is a record with a unique ID called the key[11]. If we want to store it as a relational database first, we should convert categorical columns into one-hot encoding using Weka Nominal to Binary filter; for our case, we have education, income, and checkup as categorical attributes so we convert it to one hot encoding. We will have these columns as a boolean since they can either be 0 or 1 and smoking as they can only contain yes or no. Next, we will have income which we will keep as Integer.

We will have 3 Relations(Tables), one to answer each question.

Relation 1(Education Vs. Smoking)
'_SMOKER3', **-Boolean**
'_EDUCAG=Attended college or technical school' **-Boolean**
'_EDUCAG=Graduated from college or technical school **-Boolean**

'_EDUCAG=Graduated High School' **-Boolean**
'_EDUCAG=Did not graduate High School' **-Boolean**

Relation 2 (Income Vs. Smoking)
It will have a similar structure to the above one, just that instead of all the Education columns, we will have columns relating to Income.

Relation 3 (Education and Income Vs. Checkup)

We will have 10 columns, 1 column related to checkup, 4 related to Education, and 5 related to Income, all of which are boolean in nature.
If we have a dataset of gigabytes, storing it in CSV or a similar format will not be efficient. It would be better to have a database to perform queries efficiently, and arriving data can be inserted easily into the database and processed subsequently in high traffic applications using thousands over even millions of users each day. It is crucial to distribute workloads across multiple servers. Load balancers send a request to servers to maximize speed and performance. Also, it prevents downtime. For many years load balancers used to be hardware in data centers. The modern load balancers evolved to application delivery controls that provide additional security, acceleration, and authentication. [12]. For load balancing, we can make use of a cloud service like AWS or Azure. If We use AWS, we can have a load balancer along with EC2 instances for horizontal scaling, so even if data size keeps increasing, our system is still scalable. We can also use Kubernetes and Dask, and if we use Tensorflow to make our machine learning model to be used it can be put on the cloud using tf.distributed of kubeflow. The parallelism can be trivially used during grid optimization since different models can be run on each worker node [13]. We can also use a hub that works with both Pytorch and Tensorflow; we know that Databases, data lakes, and data warehouses are best suited for tabular data and are not optimized for deep-learning applications. We can use hub to alleviate this problem, hub significantly increases data transfer speeds between network-connected machines and it can directly work with AWS. This eliminates the need to download entire datasets before running code because computations and data streaming can co-occur without increasing the total runtime. Since using a hub, we are streaming data instead of downloading it all, so even if the database size is enormous, it won't affect performance, and new data that comes we can stream it directly to our model in runtime this will significantly increase speed [14].

Privacy

Data Retention Limits: The Directive limits explicitly the period of time a controller may retain identifiable data to a period "no longer than is necessary" for the purposes for which they were collected or processed. Consequently, this implies that data must be erased or de-identified as soon as they are no longer needed [15].

Data could be harvested through hospital sensors and by extracting insights from medical records. Right to be informed: Individuals have the right to be informed about collecting and using their personal data. This data can be sold and reveal knowledge through participants' behaviors.

Since data contains sensitive or confidential information about individuals, improper disclosure of such data can have adverse consequences for a data subject's private information or even lead to civil liability or bodily harm. The development of formal privacy models like Differential Privacy will help in solving this problem. Differential privacy makes it possible for tech companies to collect and share aggregate information about user habits while maintaining the privacy of individual users[16]. It does that by adding randomness by randomly introducing yes or no instead of ground truth; now each person is protected with "plausible deniability"[17] e.g for a particular question answer can be either "yes" or "no" now through Differential Privacy what we can do is flip a coin if it is head return ground truth if it is tail then we flip again, now if it is head then we return "yes" otherwise "no" this adds privacy to participant answers. For some types of data, we can only apply aggregate queries like count, max, min, average e.g, now we can know how many people in a group smoke, but we can't tell anything about a particular participant since it's an aggregate query.

References

[1] Centers for Disease Control and Prevention, June 2, 2021, [Online]. Available: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm. [Accessed 21 June 2021].

[2] US National Library of Medicine National Institutes of Health "Annual Healthcare Spending Attributable to Cigarette Smoking: 2014," March 2015, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4603661/ [Accessed 21 June 2021].

[3] Samash.gov "NSDUH Detailed Tables: 2017," September 14, 2018, [Online]. Available: https://www.samhsa.gov/data/report/2017-nsduh-detailed-tables [Accessed 19 June 2021].

[4] Yale News "Why don't the highly educated smoke?: 2014," May 20, 2014, [Online]. Available: https://news.yale.edu/2014/05/20/why-don-t-highly-educated-smoke [Accessed 23 June 2021].

[5] US National Library of Medicine National Institutes of Health "Income, occupation and education: Are they related to smoking behaviors in China? : 2018," February 8, 2018, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805321/ [Accessed 22 June 2021].

[6] HealthData.gov, "Behavioral Risk Factor Surveillance Survey: 2015," March 12, 2021, [Online]. Available:https://healthdata.gov/State/Behavioral-Risk-Factor-Surveillance-Survey-2015/r4vf-8w6z [Accessed 20 June 2021].

[7] pandas.get_dummies [Online] Available: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html [Accessed 10 September 2021]

[8] International Journal of Environmental Research and Public Health, "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States Sunil Kumar and Ilyoung Chong": December 19, 2018, [Online] Available https://www.mdpi.com/1660-4601/15/12/2907/pdf [Accessed 8 September 2021]

[9] Scribber.com, “Understanding internal validity,”: May 1, 2020, [Online] Available https://www.scribbr.com/methodology/internal-validity/#:~:text=What%20are%20threats%20to%20internal,mean%2C%20social%20interaction%20and%20attrition. [Accessed 8 September 2021]

[10] Big Data Analytics, the University of York, “Lesson 4: Identify the remaining threats to validity for a given analysis” June, 2021, [Online], [Accessed 24 June 2021].

[11] Oracle “What is a Relational Database (RDBMS)?”, [Online]. Available:https://www.oracle.com/database/what-is-a-relational-database/ [Accessed 25 June 2021].

[12] NGINX, “What is Load Balancing,”, [Online]. Available:https://www.nginx.com/resources/glossary/load-balancing/ [Accessed 25 June 2021].

[13] Section II ‘Design Principles’ in Rubinstein, Ira & Good, Nathaniel: 2012,” August 12, 2012, [Online]. Available:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2128146 [Accessed 24 June 2021].

[14] Towards Data Science, “Understanding Differential Privacy [Online] Available”: June 30, 2019, https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a [Accessed 1 September 2021]

[15] Towards Data Science, “Handling Big Datasets for Machine Learning”: March 11, 2019, [Online] Available: https://towardsdatascience.com/machine-learning-with-big-data-86bcb39f2f0b [Accessed 1 September 2021]

[16] Harvard University, “Differential Privacy: A Primer for a Non-technical Audience”: February 14, 2018, [Online] Available: https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf [Accessed 5 September 2021]

[17] Activeloop/hub [Online] Available: https://github.com/activeloopai/Hub [Accessed 6 September 2021]