Data Mining for the World Health Organization: Risk Factors for Cervical Cancer

DSCI 64210-001 Data Science
Term Project Deliverable 3: Final Report
Li Maltba & Anne Sawyer
28 November 2017

Data Mining for the World Health Organization: Risk Factors for Cervical Cancer

ABSTRACT

As of 2014, the World Health Organization (WHO) estimated that approximately 1 million women worldwide have cervical cancer. The underlying cause of nearly all cases of cervical cancer is HPV, which affects roughly half of sexually active adults at some point in their lives. To change cancer outcomes, WHO established a Cancer Control Programme which is dedicated to preventing, treating, and educating the public about cancer by developing policies, plans, and programs that promote public health. To help WHO's Cancer Control Programme achieve its goals for meeting the reproductive health needs of women in particular, our data science team performed a series of predictive modeling and analysis techniques. Using a dataset of risk factors for cervical cancer provided by the Hospital Universitario de Caracas in Caracas, Venezuela, our team performed both supervised and unsupervised learning techniques such as clustering, decision trees, and logistic regression. These models allowed us to first identify the primary risk factors associated with a diagnosis of cervical cancer, and then to predict which of these is more likely to affect cancer outcomes. The results of our analysis will help WHO better leverage data-driven decisions for developing more effective cervical cancer prevention programs.

INTRODUCTION

Cervical cancer is one of the most preventable cancer types affecting women, yet 266,000 women died from the disease in 2012 (WHO, 2014). HPV has been identified as the leading cause of precancerous cervical cell changes and invasive cervical cancer in women. HPV today affects 50% of sexually active adults at some point in their lives (U.S. FDA, 2017). Early screening and diagnosis is critical for preventing cancer later in life. Though HPV infections often clear up on their own, for

the minority of women with chronic HPV infections, early treatment can prevent precancerous cells from metastasizing into invasive cancer.

WHO Cancer Control Programme Mission Statement:

The key mission of WHO's work in cancer control is to promote national cancer control policies, plans and programs that are harmonized with strategies for non-communicable diseases and other related health concerns. Our core functions are to set norms and standards for cancer control including the development of evidence-based prevention, early diagnosis, screening, treatment and palliative care programs as well as to promote monitoring and evaluation through registries and research that are tailored to the local disease burden and available resources.

WHO's Cancer Control Programme is dedicated to controlling cancer by developing policies, plans, and programs that promote public health. In this case, the organization wants to better leverage data-driven decisions for its Cancer Control Programme to help prevent cervical cancer and better meet the reproductive health needs of women. Specifically, WHO wants to use data science to more precisely target at-risk groups who could benefit from their education, screening, and treatment programs. Gathering and analyzing patient data from hospitals in developing countries (where cervical cancer is more prevalent) will help WHO extract insights about which risk factors and behaviors are more likely to lead to cervical cancer. Consistent, long-term collection of patient data can also help the organization assess which programs are more effective.

To help WHO's cancer program achieve its goals, our team used a dataset of risk factors for cervical cancer to:

- identify the specific risk factors that may lead to HPV infection.
- predict which risk factors will be more likely to lead to cervical cancer when HPV is present.
- classify which of the two most common forms of birth control, oral contraceptives (the Pill) or intrauterine devices (IUDs), is more likely to have a positive correlation with cervical cancer.
- determine to what extent the presence of other sexually-transmitted infections (STIs) lead to cervical cancer.

Data-driven decision-making can add value for mission-based organizations like WHO and make a difference in the quality of life for people across the globe. The results of our predictive modeling can be used to provide hospitals, clinics, governments, and the public with better information and resources for preventing cervical cancer in vulnerable populations. Armed with a deeper understanding of the data, health professionals will be able to implement best practices for monitoring and evaluating the course of diseases. Education and prevention efforts will be supported by empirical evidence from hard data, rather than well-intended assumptions. For example, most people would agree that tobacco use is unhealthy, and may cause lung and other oral cancer types. But to what extent does smoking lead to cervical cancer? Data mining can tell us the answer to this question, as well as many others.

If the Cancer Control Program can predict the likelihood of girls and women with certain risk factors to develop cervical cancer, they can use that information to improve cancer screening standards. The results of clustering analysis can help health professionals fine-tune health assessments and follow-up procedures for at-risk groups. Early detection and treatment will improve survival rates for cancer patients, and help localities better manage the costs associated with ongoing cancer treatment. Results of classification models can further help with cancer research, and open up opportunities for exploration into new algorithms and other data-based health solutions. WHO can share data models and methods with other agencies to improve public health education about the risks and benefits of specific birth control methods with regard to cervical cancer.

DATA UNDERSTANDING

Our data science team used a dataset provided by the Hospital Universitario de Caracas in Caracas, Venezuela. The set contains a total of 36 attributes and 858 instances. All data are numeric. Several types of behavioral and disease-based factors are represented, including:

- age of initial sexual activity
- number of sexual partners

- age of first pregnancy
- number of pregnancies
- tobacco use
- birth control method
- STD status
- other cancer-related diagnoses

As with all collections of raw data, the team's first challenge in working with this particular dataset was its cleanliness. Generally, instances of missing or corrupt values must be cleaned by imputing them, marking them missing, or removing them altogether (Brownlee, 2016). Another early challenge for preprocessing our data involved defining its features. Important attributes were not labeled clearly, and the team was not provided with any details about how, when, and for over what length of time the data was collected. For future data mining initiatives, our team would recommend that WHO's Cancer Control Programme collect more detailed patient data about high-risk vs low-risk HPV infections and HPV vaccination history, as well as screening and treatment information such as age of first detection or diagnosis of cancer, HPV, and other STIs, length of time between initial screening and follow-up. Finally, additional attributes for socioeconomic factors, such as income, geographic location, marital status, ethnicity, et al. would increase the accuracy of data predictions.

The data mining objectives for this project were threefold. Specifically, our team wanted to:

- identify trends and correlations between specific risk factors and the presence of HPV leading to diagnosis of cervical cancer.
- make probability predictions for women in target groups to develop cervical cancer.
- classify which risk factors are more likely to lead to HPV infection.

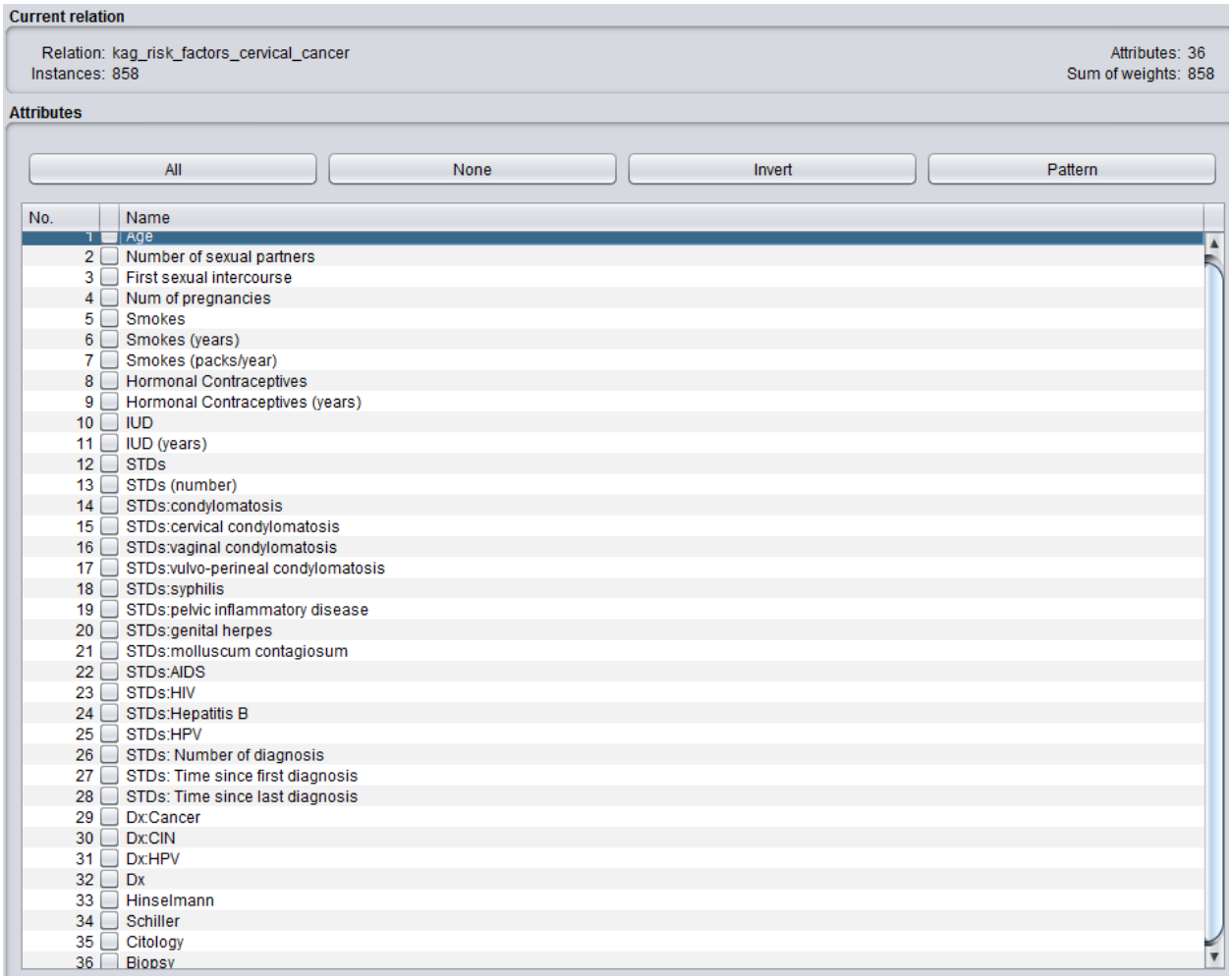

Figure 1. Attributes for cervical cancer risk.

DATA PREPARATION

Our data set contains values that are a mix of categorical and numeric, therefore in order to analyze the data using classification methods, we first converted selected attributes from numeric to nominal values. The attribute for number of sexual partners had 26 (3%) missing values. We cleaned this data by imputing the missing values as the mean value of the attribute (2.528).

During the initial process of mining the data, the first challenge we faced was interpreting the meaning of attributes without clear labels. Attributes with labels such as Dx: Cancer and Dx lack any explanation as to what they represent. Therefore, we decided to assume that column “Dx” represents the Cervical Cancer diagnosis, and

thus set it as Class. Then, by using all 36 attributes, we decided to use Logistic Regression analysis to determine if there is any relationship between all the other attributes and "Dx," i.e., cervical cancer diagnosis.

Results illustrated that attributes Dx: CIN (abnormal cells, dysplasia), Smoking, and Dx: Cancer are the three most significant factors for Dx: (cervical cancer) with 99% accuracy. This result makes sense since Dx: CIN represents diagnosis of positive cervical intraepithelial neoplasia, which indicates the presence of pre-cancerous cellular changes, and both smoking and the presence of other cancer types will lower a woman's immune system and make her more prone to illness. Thus, we confirmed that our assumption of Dx is correct for final diagnosis of cervical cancer, identifying Dx: Cancer as another type of cancer with which the patient was diagnosed.

After examining the instances at a granular level with our data mining goals in mind, we decided to not include the four attributes Hinselmann, Schiller, Citology, and Biopsy in our data mining as they did not appear relevant to the data mining problem. Moving forward, we divided all available remaining attributes into four different groups to examine the risk factors that may relate to cervical cancer diagnosis.

DATA MODELING: GROUP 1

With the first group, we wanted to see if any behavior factors might lead to cervical cancer diagnosis:

Data Attributes: Group 1

Number of sexual partners	*Number of sexual partners*
Num of pregnancies	*Number of pregnancies*
Smokes	*Currently smokes*
Smokes (years)	*Number of years of smoking*
Smokes (packs/year)	*Number of packs/year of smoking*
Dx	*Cervical cancer diagnosis*
Dx:Cancer	*Diagnosis of other type of cancer*
Dx:CIN	*Diagnosis of positive cervical intraepithelial neoplasia*
Dx:HPV	*Diagnosis of HPV infection*

Mining the data using logistic regression models reveals a 99.3% rate of accuracy with only 6 total errors. Though the model's accuracy was high, our team needed to better visualize which attribute actually was the leading factor for Dx, so we decided to use the J48 classification tree method. To improve the usefulness of this model and reduce the risk of overfitting our data, we first used a training set to produce a model with known output values. The remainder of the data was used for predictive mining, which produced a model with 98% accuracy. Having automatically pruned irrelevant data, the decision trees revealed that smoking is the most important factor for cancer diagnosis, especially when the patient also presents another type of cancer. However, when we ran additional models with different combinations of Group 1 attributes, the results showed that attributes for number of packs per year and how many years smoking have no significant effect on cancer outcomes. The fact of smoking makes the biggest difference.

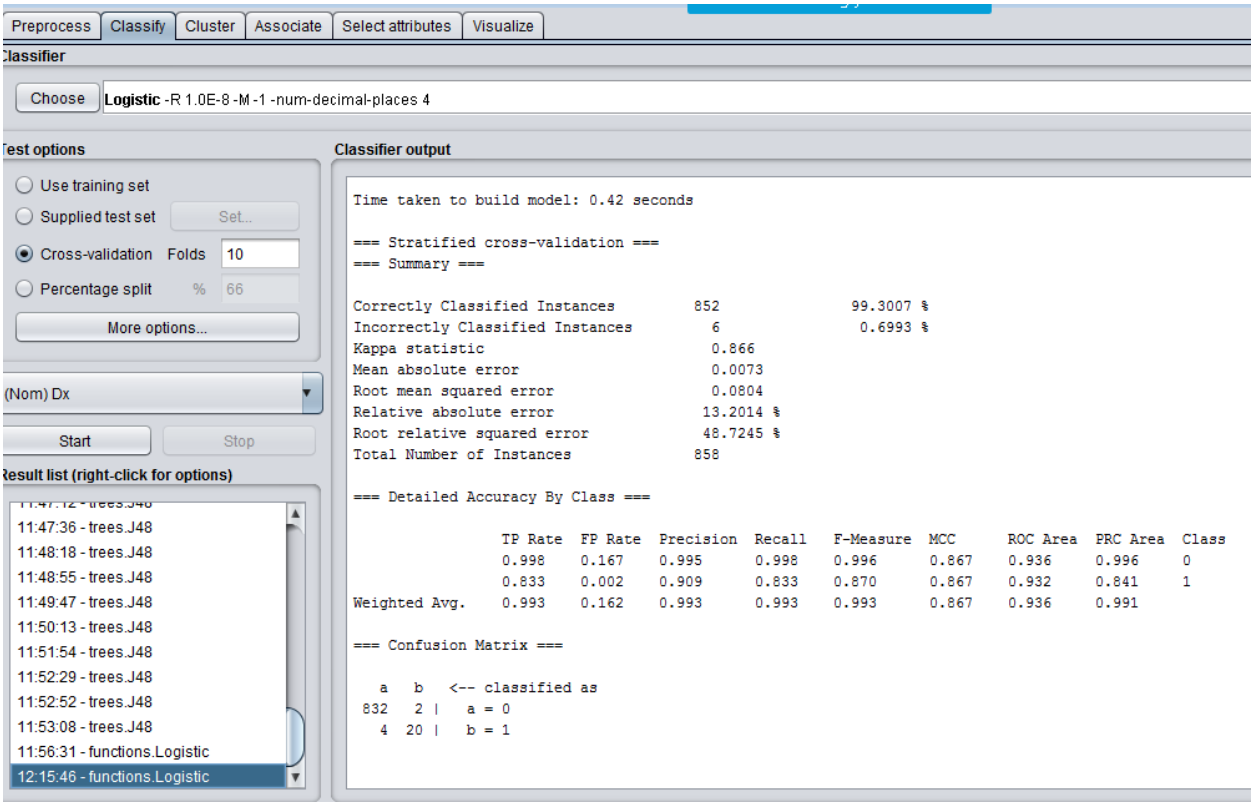

Figure 2. Data Model results using Logistic Regression method.

Alternative methods of supervised learning such as regression and causal modeling were initially considered, but our team determined that classification methods would be the most appropriate approach to mining the data since the

question we wanted to answer was categorical in nature (presence or absence of class: cervical cancer). The advantage of using decision trees for mining our data is that it quickly and cheaply produced an elegant yet robust model that allowed us to more clearly visualize the effects of various factors, or "decisions," that may lead to cervical cancer. Regression methods, which require a numerical target, were rejected based on their irrelevance to our data mining problem. Causal modeling, which attempts to predict which factors directly impact outcomes, has the potential to provide predictive value for WHO's fight against cancer, and should explored in future data mining initiatives. For this project, our team rejected this method because of the significant investment in data and A/B testing required for proper implementation (Provost & Fawcett, 2013).

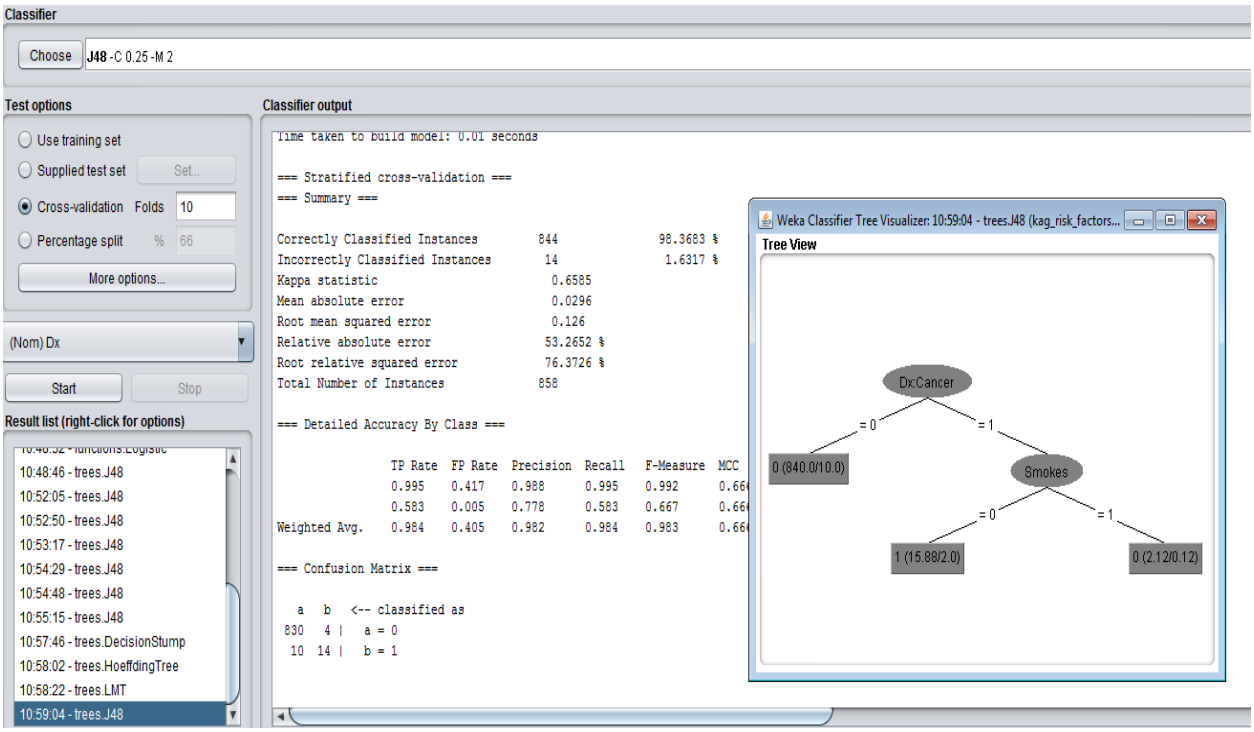

Figure 3. J48 tree model results for behavioral attributes.

DATA MODELING: GROUP 2

In the second group, we wanted to examine if choice of birth control method (hormonal or IUD) would make any difference in cervical cancer outcomes.

Data Attributes: Group 2

Hormonal Contraceptives	*Hormonal Contraceptives*
Hormonal Contraceptives (years)	*# of years hormonal contraceptives used*
IUD	*Intrauterine Device*
IUD (years)	*# of years intrauterine device used*
Dx	*Cervical cancer diagnosis*
Dx:Cancer	*Diagnosis of other type of cancer*
Dx:CIN	*Diagnosis of positive cervical intraepithelial neoplasia*
Dx:HPV	*Diagnosis of HPV infection*

Initial J48 tree results indicated that the type of birth control method is not a significant factor for cancer diagnosis. However, as with behavior attributes, analysis of birth control methods was stymied by missing values, which had to be removed from our models. Attributes for hormonal contraceptives had 108 missing values (13%), while that of IUDs had 117 (14%). This had the effect of further reducing an already small sample size, thereby rendering our results less useful for determining the impact of specific risk factors for our target.

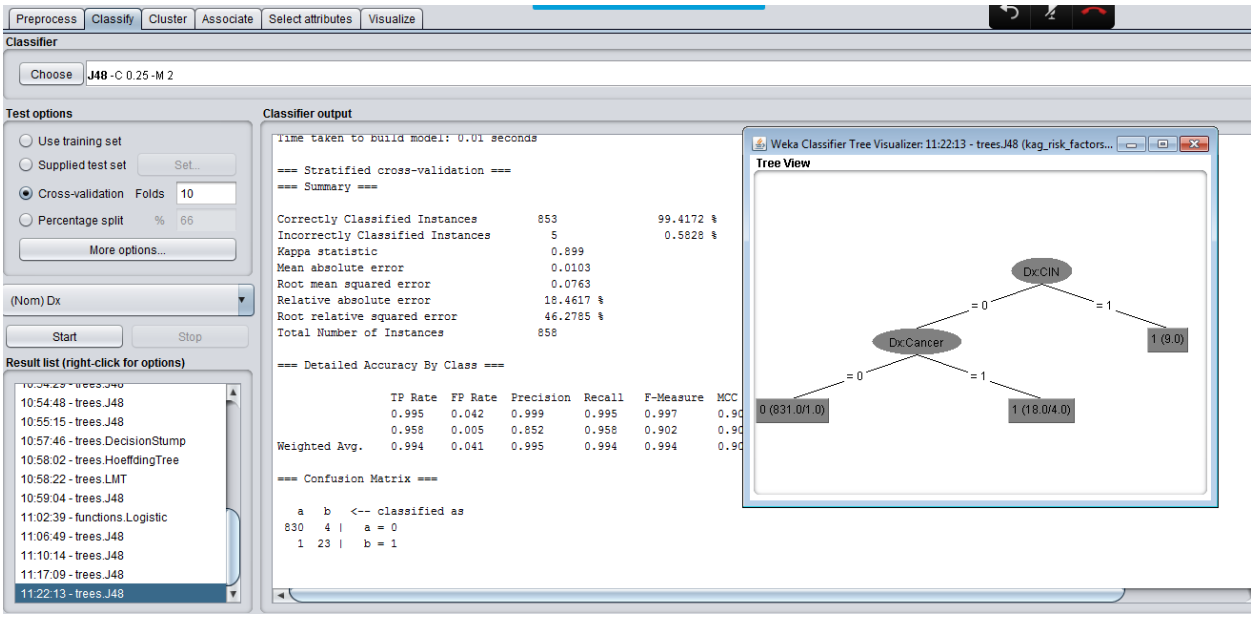

Figure 4. Initial J48 tree model results for birth control method.

After removing missing values, converting numeric values to nominal for categorical attributes, and discretizing the values for years of birth control use, we ran more decision tree models. As we expected, the results indicated that CIN and HPV

are the primary predictors of cervical cancer. However, the model also predicted that number of years hormonal contraceptives are used may influence cervical cancer outcomes. Interestingly, the bin with the longest duration of hormonal contraceptive use was less likely to develop cervical cancer than those using this birth control method for 0-7 or 8-14 years. Most instances were binned in the 0-7 years category (672 out of 738 total), with 269 of these instances assigned a value of 0. Since sexually active women who do not use birth control are more at risk not only for unplanned pregnancies but also for STIs, they may also increase their chances of developing cervical cancer.

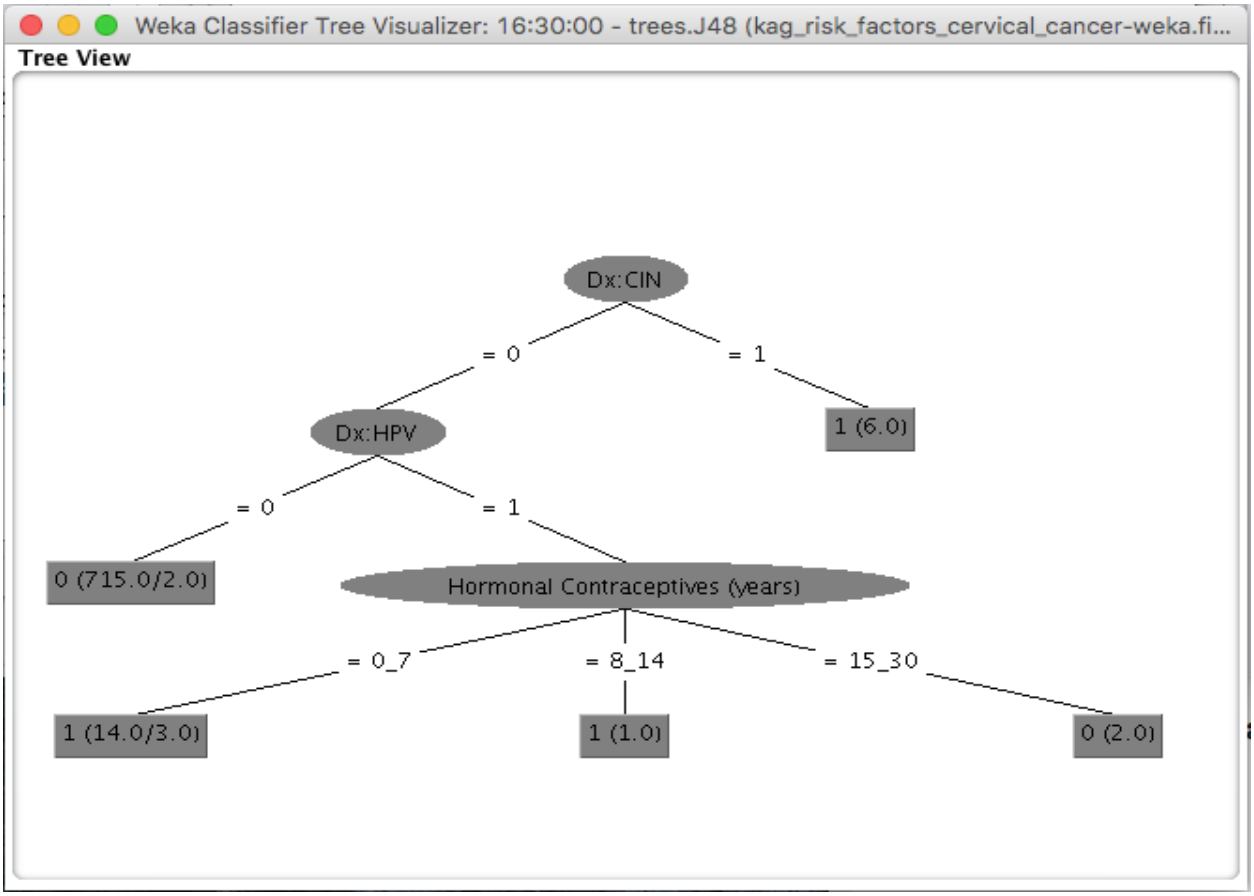

Figure 5. J48 tree results using birth control attributes.

More information about birth control habits would enable better predictions for the Cancer Control Programme's research and education efforts. For example, condom use as a birth control method is widely known to also reduce the risk of STD contraction, though perhaps less so for HPV, which can be transferred via skin-to-skin contact without having intercourse (U.S. FDA, 2017). Access to data about condom use

along with other methods would allow our team to more effectively evaluate how choice of birth control may contribute to HPV contraction. Women who only use oral contraceptives, for example, may be more at risk for abnormal cervical changes because they may overlook their risk for STIs. This and similar hypotheses could be validated or debunked with better data. For example, class probability estimation methods could be used to predict a percentage of likelihood that women who use condoms, IUDs, the Pill, or some combination of these will contract high-risk HPV.

DATA MODELING: GROUP 3

The third group contains different STD attributes:

Data Attributes: Group 3

STDs	*sexually transmitted diseases*
STDs (number)	*total # of sexually transmitted diseases present*
STDs:condylomatosis	*STDs:genital warts*
STDs:cervical condylomatosis	*STDs:cervical warts*
STDs:vaginal condylomatosis	*STDs:vaginal warts*
STDs:vulvo-perineal condylomatosis	*STDs:warts caused by certain types of HPV*
STDs:syphilis	*STDs:syphilis*
STDs:pelvic inflammatory disease	*STDs:pelvic inflammatory disease*
STDs:genital herpes	*STDs:genital herpes*
STDs:molluscum contagiosum	*STDs:pox virus*
STDs:AIDS	*STDs:AIDS*
STDs:HIV	*STDs:HIV*
STDs:Hepatitis B	*STDs:Hepatitis B*
STDs:HPV	*STDs:HPV*
STDs: Number of diagnosis	*STDs: Number of diagnosis*
STDs: Time since first diagnosis	*STDs: Time since first diagnosis*
STDs: Time since last diagnosis	*STDs: Time since last diagnosis*
Dx	*Cervical cancer diagnosis*
Dx:Cancer	*Diagnosis of other type of cancer*
Dx:CIN	*Diagnosis of positive cervical intraepithelial neoplasia*
Dx:HPV	*Diagnosis of HPV infection*

With this set of data, we wanted to see if the presence of STDs had any positive relationship with the target diagnosis, and if so, which type of STD is the leading factor. However, tree results revealed that the presence of other STDs does not appear to be significant for cancer outcomes. If this is true, then WHO could use this information to make better decisions about how to allocate funding for cancer research, treatment, and public health education efforts.

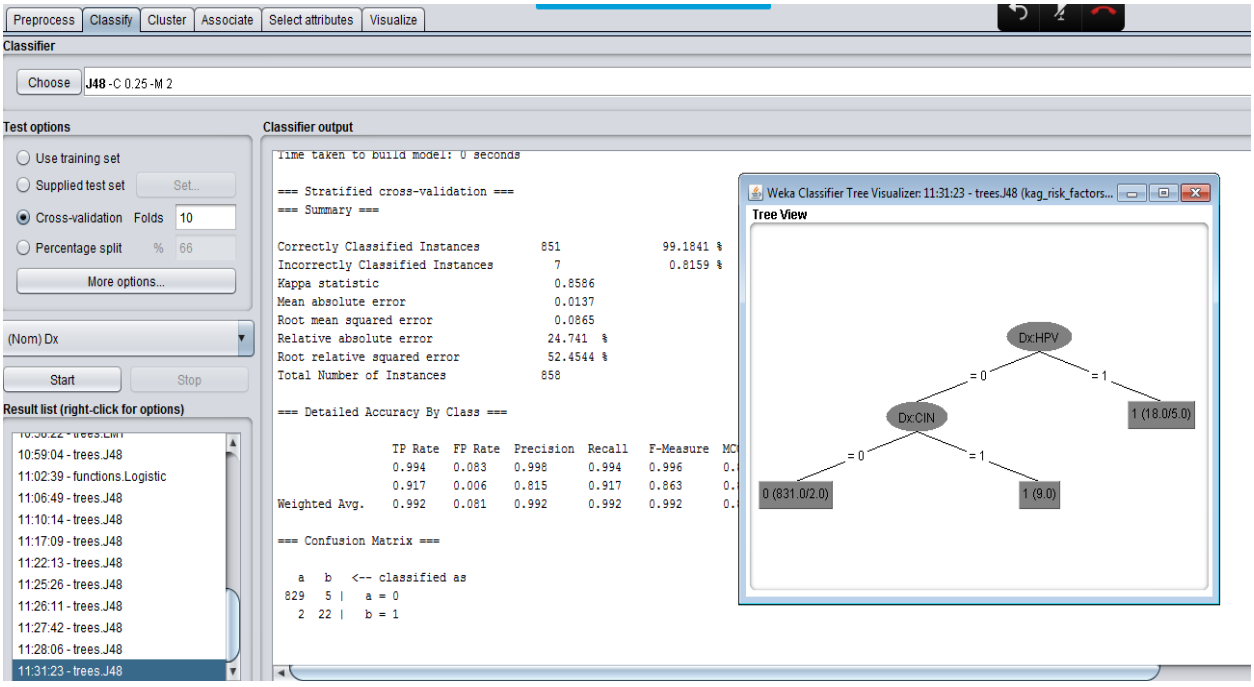

Figure 6. J48 tree results using STD attributes.

Additional patient data about these attributes would produce more meaningful models for predicting health outcomes for women, particularly those in developing countries with higher rates of infection and less stable sources of funding for disease prevention and treatment. The dataset used in our modeling efforts indicated only 2 instances of positive HPV diagnosis, which does not provide an accurate snapshot of general rates of HPV infection in world populations. A larger sample of HPV positive patients with more granular attributes for high-risk and low-risk types of the virus, as well as data for other STDs would allow our team to put classification methods to better use for cancer prediction. Naive Bayes as a simple, efficient, and low-cost form of analysis could be used to predict cancer risk for patients with high-risk HPV types and other factors. We would also be able to conduct exploratory modeling such as

cluster analysis or co-occurrence grouping to tease out patterns for cervical abnormalities among women with different combinations of STI and HPV infections.

DATA MODELING: GROUP 4

With Group 4, we wanted to see if any interesting patterns could be culled from unsupervised k-means clustering:

Data Attributes: Group 4

Age	*Patient's age*
Number of sexual partners	*Number of sexual partners*
First sexual intercourse	*Age of the first sexual intercourse*
Num of pregnancies	*Number of pregnancies*
Smokes	*Currently smokes*
Smokes (years)	*Number of years of smoking*
Smokes (packs/year)	*Number of packs/year of smoking*
Hormonal Contraceptives	*Hormonal Contraceptives*
Hormonal Contraceptives (years)	*# of years hormonal contraceptives used*
IUD	*Intrauterine Device*
IUD (years)	*# of years Intrauterine device used*
STDs:genital herpes	*STDs:genital herpes*
STDs:molluscum contagiosum	*STDs:pox virus*
STDs:AIDS	*STDs:AIDS*
STDs:HIV	*STDs:HIV*
STDs:Hepatitis B	*STDs:Hepatitis B*
STDs:HPV	*STDs:HPV*
STDs: Number of diagnosis	*STDs: Number of diagnosis*
STDs: Time since first diagnosis	*STDs: Time since first diagnosis*
STDs: Time since last diagnosis	*STDs: Time since last diagnosis*
Dx	*Cervical cancer diagnosis*
Dx:Cancer	*Diagnosis of other type of cancer*
Dx:CIN	*Diagnosis of positive cervical intraepithelial neoplasia*
Dx:HPV	*Diagnosis of HPV infection*
Biopsy	*Biopsy*

Initially, all attributes (with the exception Hinselmann, Schiller, and Citology, which were eliminated as irrelevant) were used as a baseline for exploration, and then further refined based on the results. Defining the number k involved trial and error: models were run using 3, 4, and 5 clusters. Initial models using all 33 attributes in 3 clusters revealed that several attributes could be eliminated based on a lack of distinct values. For example, there are so few positive values for most STDs that these attributes provide no value for meaningful analysis. Similar to what we discovered using J48 trees, the small number of positive values for risk factors tends to skew cluster results. In most models, the majority (50% or more) of instances were grouped together, while the remainder comprised the other 2, 3, or 4 clusters.

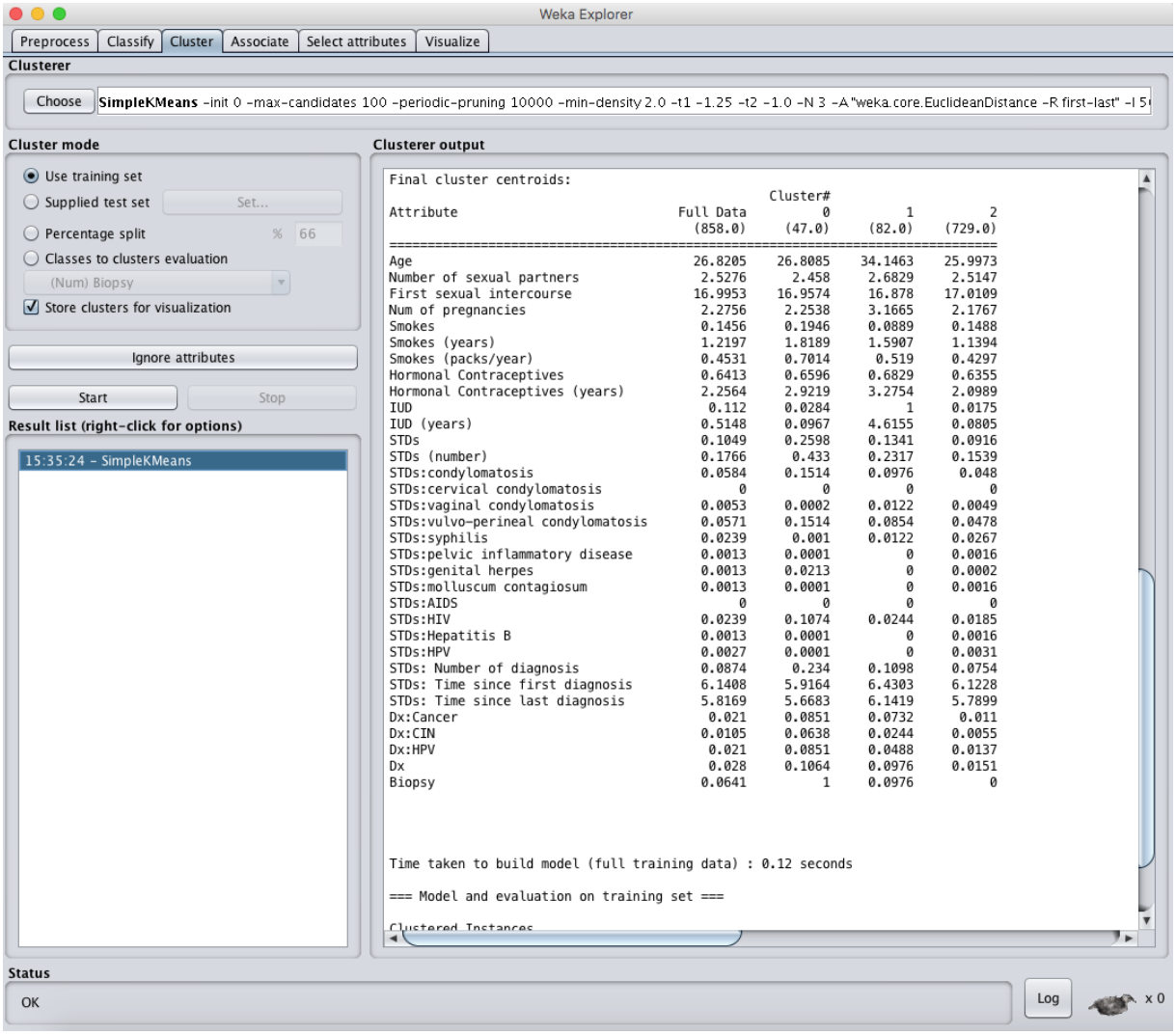

Figure 7. Cluster results using all 36 attributes for 3 clusters.

The results pictured in Figure 8 using fewer attributes proved more revealing than initial modeling efforts. Four clusters were generated using 14 attributes, with Cluster 0 forming the highest-risk group for cervical cancer. Women in this group tend to be smokers who are less likely to use birth control and more likely to have STDs (and HPV in particular), CIN (precancerous cervical cell changes), and cervical cancer. Interestingly, Cluster 0 did not differ significantly from the lower risk clusters in terms of age of first sexual intercourse, number of pregnancies, or number of sexual partners. This indicates that these factors are less indicative of high risk for cervical cancer than smoking and HPV.

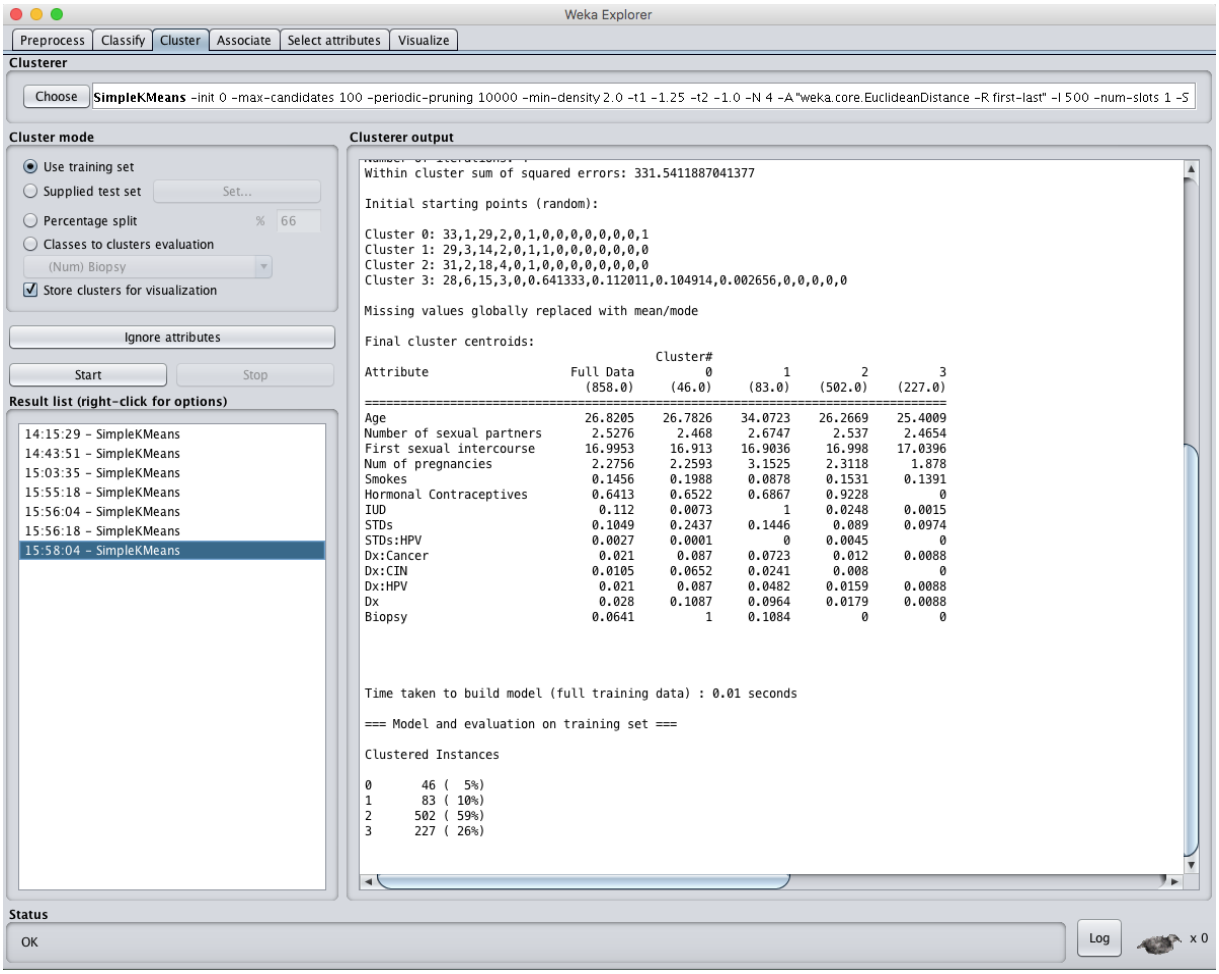

Figure 8. Cluster results using 4 groups with 14 attributes.

Our team chose to use clustering techniques over other methods of unsupervised learning to uncover patterns among various behavioral and health risk factors for cervical cancer. Nearest neighbor analysis, a technique which uses a

combination of clustering and classification, was considered but discarded due to its higher cost for accurate prediction, intelligibility of modeling, and the need for deep domain knowledge, which was unavailable to our team for this project (Provost & Fawcett, 3429-3489).

Both methods may provide WHO with information about which behaviors tend to naturally correlate or occur together to affect health outcomes, but clustering is cheaper, faster to construct, and can be iterated many times for better results. WHO's Cancer Control Programme can use results of clustering analysis of hospital patient data to refine the direction of their strategic plan for HPV prevention and cancer treatment programs. Hospitals with higher rates of reproductive health problems can be targeted for deploying HPV vaccines, biopsy screenings, and use of LEEP procedures for cancer prevention.

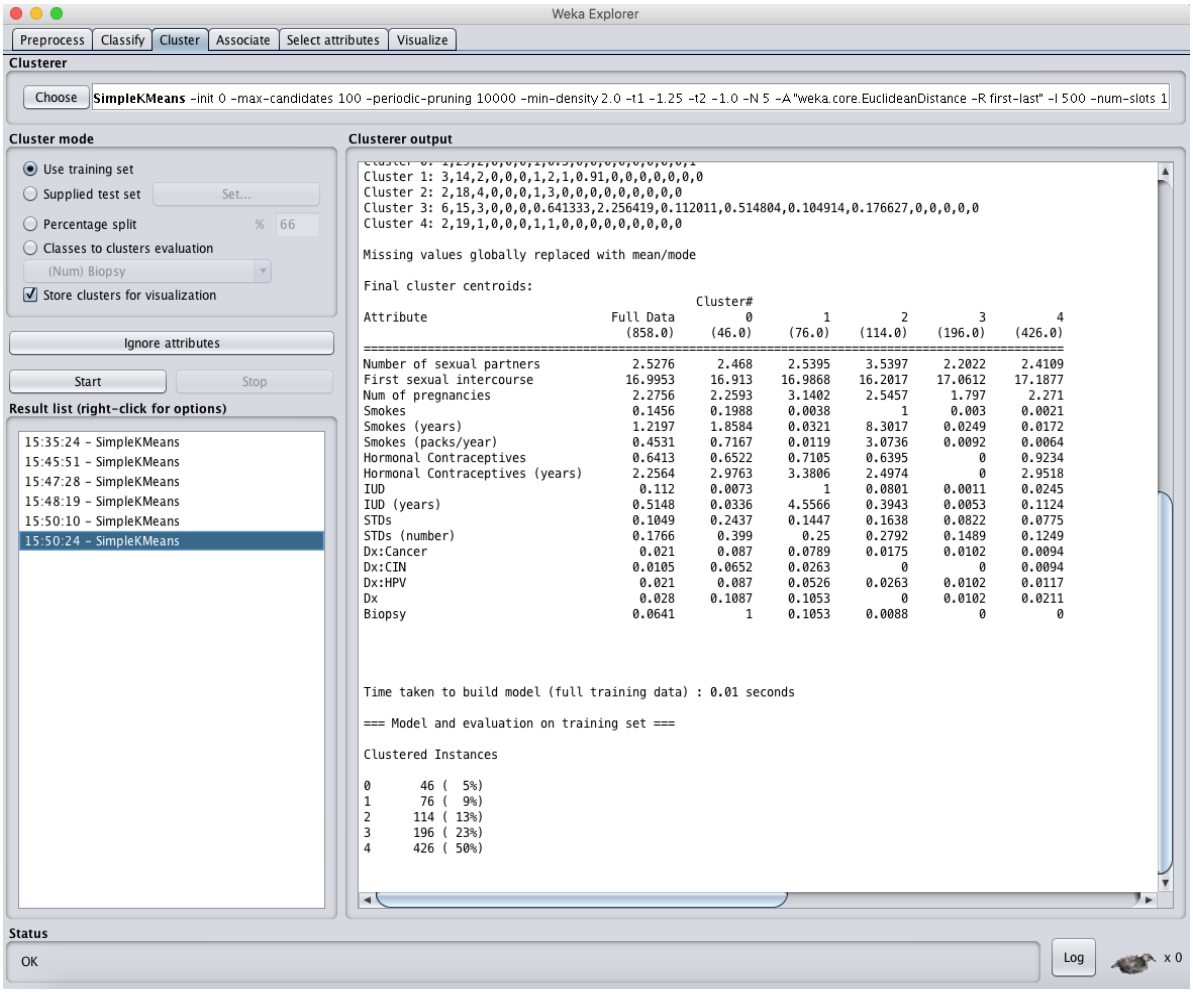

Figure 9. Additional K-means model using 5 clusters.

DEPLOYMENT

Results of our team's data mining and predictive modeling may be deployed to advance WHO's Cancer Control Programme initiatives for reducing cervical cancer rates and improving the reproductive health of women. To address immediate health needs, WHO can use our clustering models to identify the most at-risk populations of women. The organization can then apply our predictive models using new data instances to predict cervical cancer in these groups. For long-term strategic planning, WHO can use the results of our study to develop recommendations and programs for cancer prevention and public health education. Our model may also be used a foundation to guide further data collection and analysis for researching other risk factors for cervical dysplasia and precancer, such as the presence of other cancers and compromised immune system functioning.

Working with big data to solve the world's health outcomes is fraught with complex and sometimes contradictory findings. While the models themselves produce accurate results, WHO should be aware that important socioeconomic data such as income, geographic location, marital status, ethnicity, etc. was not available for our study. These could be important contributors for identifying where, how, and for whom to deploy data mining models. Low-income populations and certain ethnic groups may be more at risk for low immunity, higher rates of disease, and more rapid onset of cancer. Our model should not be viewed myopically or used to predict cancer divorced from the socioeconomic context of the group to which it is applied. The problem of "dirty data" is always an issue with any data mining initiative, and our study is no different. The dataset on which our models are based was missing values for several attributes. In addition, our dataset lacked key details about patient medical history such as HPV vaccination, the presence of other cancers, and past treatments.

Before deploying this model, our team recommends that WHO aggregate larger and more robust data for further study. This will not only improve the accuracy and predictive power of our models, but also open the door to new behavioral insights, risk associations, and more meaningful health solutions. When gathering data, ethical considerations such as HIPAA compliance, patient privacy, and informed consent should be fully understood by both patients and health professionals. Personal data

about patients, including names and contact information is not necessary for data mining, and therefore should not be included in data collection.

EVALUATION

In conducting data analysis, our team found that J48 decision trees are very useful for visualizing the relationships between attributes and interpreting the results. However, there are many valid concerns and limitations regarding the data:

- Attribute labels for Dx, Dx: CIN, Dx: Cancer are not clearly defined, which makes analysis more difficult.
- Deeper knowledge of other analysis methods would help our understanding of the data and knowing which methods would be more useful.
- Missing information and a small sample size are both challenges for modeling. For example, Dx: Cancer does not specify type, location, stage and treatment history of cancer. We know that HPV causes cervical cancer, but our data shows that Dx: CIN is more important than the HPV attribute. How do we ascertain whether this is true, or only a function of an incomplete picture of the data?
- Even though the presence of other STDs does not appear to be significant for cancer outcomes, the small size of our dataset as well as the small number of STD diagnoses and lack of treatment history may affect the results of data analysis.
- There is no detailed information about high-risk vs low-risk HPV infections and HPV vaccination history to rule out some possibilities that may have affected the data mining result.

The general conclusion we have determined for this analysis is that HPV and CIN (dysplasia, abnormal cell changes) are both contributors to cervical cancer diagnosis, but CIN is most significant contributor and smoking is more likely to result in cell abnormalities than HPV alone. When presenting another form of cancer, a smoker will be at higher risk of developing cervical cancer. This result indicates that educating and enforcing smoking cessation may help prevent CIN from developing into full-blown

cervical cancer. WHO can use this information to promote smoke cessation programs and provide assistance for women who have difficulty maintaining smoke-free status.

The results of our models also illustrate that WHO may need to study other causes of CIN. Additional data could be collected that is more focused on CIN risk factors, such as genetic, behavioral, and other contributors to disease. Thus, related immunizations and education programs can be provided to target populations. However, WHO should bear in mind that this level of research may require a significant investment of time and resources to prepare, validate, and deploy before significant results may be seen. Our preliminary studies also point to diminished immunity as a contributor to the development of cervical cancer. A short-term solution to this problem that could produce more immediate results is strengthening existing programs that focus on general health of women, particularly nutrition and stress-reduction.

Based on our data mining results, we believe that cervical cancer prevention can be improved by helping people change their lifestyles. We recommend that WHO focus on designing and implementing programs that promote women's health and general well-being.

APPENDIX: TEAM MEMBER CONTRIBUTION

Team members Li Maltba and Anne Sawyer collaborated on this project using a variety of communication methods, including email, text, and voice calls. We used Google drive for editing our work and sharing resources, and conducted 2 Webex conference sessions, each with an approximate duration of 2-3 hours. All data models were developed using Weka software. Li Maltba led the team in constructing data models during our Webex sessions and Anne Sawyer supplemented these with additional models for Deliverables 2 and 3. Our team shared the roles of Writer (responsible for composing initial drafts) and Editor (overseeing changes and preparing work for final submission).

DELIVERABLE 1: Li Maltba, Editor | Anne Sawyer, Writer

DELIVERABLE 2: Li Maltba, Writer & Editor | Anne Sawyer, Writer & Editor

DELIVERABLE 3: Li Maltba, Writer | Anne Sawyer, Editor

REFERENCES

Brownlee, J. (2016). How to handle missing values in machine learning data with Weka. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/how-to-handle-missing-values-in-machine-learning-data-with-weka/

Cancer Control. (2015). Mobile technology in cancer control for emerging health systems: digital divide or digital provide? [JPEG image]. Retrieved from http://www.cancercontrol.info/cc2015/

Nations Encyclopedia. (n.d.) The World Health Organization (WHO) – Activities. [JPEG image]. Retrieved from http://www.nationsencyclopedia.com/United-Nations-Related-Agencies/The-World-Health-Organization-WHO-ACTIVITIES.html

O'Neal, M. (2016). Cervical cancer prevention in Cape Coast. [JPEG image]. Retrieved from http://blog.globalmamas.org/?page=2

Provost, F., & Fawcett, T. (2013). *Data science for business.* [Kindle version]. Retrieved from Amazon.com

UCI Repository. Cervical Cancer Risk Factors for Biopsy. Retrieved from https://www.kaggle.com/loveall/cervical-cancer-risk-classification.Data source:https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

U.S. Food & Drug Administration. (2017). HPV (human papillomavirus). Retrieved from https://www.fda.gov/forconsumers/byaudience/forwomen/ucm118530.htm

WHO/N. Lkhagvasuren. (n.d.) A cervical cancer screening in Mongolia. [JPEG image]. Retrieved from http://wunrn.com/2014/11/cervical-cancer-world-health-organization-issues-new-guidelines-on-treating-preventing-cervical-cancer/

World Health Organization. (2014). Comprehensive cervical cancer control: a guide to essential practice. [2nd Ed.]. PDF document downloaded from http://apps.who.int/iris/bitstream/10665/144785/1/9789241548953_eng.pdf

World Health Organization. (2017). Cancer Control Programme. Retrieved from http://www.who.int/cancer/en/

World Health Organization. (2017). World Cancer Day 2017. [JPEG image]. Retrieved from http://www.who.int/cancer/world-cancer-day/2017/en/