Abstract:

Cervical Cancer is cancer that arises due to the abnormal growth of cells originating from the cervix. The carcinoma generally presents little no symptoms early on, with later symptoms continually insidiously discreet thus, those vulnerable are usually at risk of being misdiagnosed. Although 90% of all causes are due to hpv, other risk factors include lifestyle decisions, including smoking, birth control pills, starting to become sexually active at a young age and having many partners. Although these factors contribute to an individually small amount of cases, there is a lot of crossover with these behaviours and the aforemented sexually transmitted disease, hpv, which increases the risk factor of contracting both the disease and in turn cervical cancer (although no all cases of hpv progress to cervical cancer). Considering this, we theorize it should be hypothetically possible that a machine learning classification model trained on patient clinical records could be used to identify high risk individuals. In this study we have tested Knn and decision trees for the identification of individuals at risk of cervical cancer.

Introduction:

Cervical cancer is a relatively rare form of cancer compared to other cancer diagnoses amongst women, accounting for 1.4% of all women’s cancer diagnosis in 2020(https://www.canceraustralia.gov.au/affected-cancer/cancer-types/cervical-cancer/cervical-cancer-australia-statistics) . The survival rate of cervical cancer for five years is 74% from 2012-2016 in Australia, as evidenced by recorded statics, since the introduction of organizing screening beginning in 1991, diagnosis of cervical cancer has dropped on average 4.5% each year. 80% of cervical cancers are originate from developing countries, this is relevant as the dataset was provided by and from the school of public health in west Sumatera, Indonesia(reference article here), which can be regarded as a developing area. In this report, we have explored whether based on the provided questionnaires results, using Machine learning we are able to accurately predict the occurrence of cervical cancer in an individual based on their responses to the questionnaire. From a practical perspective, due to it’s subtle symptoms, it would be beneficial to be able to accurately identify patients for therapeutic treatment.

Methodology

Data  
the Sobar’s diagnosis dataset (Sobar, Machmud, Wijaya 2016) was obtained from the UCI machine learning repository. The dataset contains responses from a study conducted in 2016 between 3 universities (school of public health: university of Andalas, school of information Technology: STIMIK Eresha, School of health science, STIKIM Jarkarta) for presentation at the international conference on internet services technology and information engineering 2016. The questionnaire asks questions based on social science theory modules:   
1. The health belief model  
2. Protection Motivation Theory  
3. Theory of Planned behavior  
4. Social Cognitive Theory

From these theories, eight determinants of behavior were extracted:

-behaviour  
-Perception  
-intention  
-Motivation  
-Subjective Norm   
-Attitude  
-Social Support  
-Empowerment

These determinants were broken down further to individual variables which are used as features to create a classification model to detect cervical cancer risk:

Behaviour:  
-sexual risk  
-eating  
-personal hygiene

Perception:  
-vulnerability  
-severity

Intention:  
-aggregation  
-commitment

Motivation:  
-strength  
-willingness  
  
subjective norms:  
-significant person  
-fulfimment

Attitude:  
-consistency  
-spontaneity

Social Support:  
-emotionality  
-appreciation  
-instrumental

Empowerment:  
-knowledge  
-abilities  
-desires

Data analysis  
Data manipulation, data cleaning and analysis were performed using the python packages pandas and numpy. The data was checked for null values, statistical parameters, and inconsistent data values, from this tertiary view, it became clear that all the values were based on ordinal values, thus outliers removal wasn’t necessary.  
  
Data modelling was done using matplotlib, and of the data visualizations we were prompted to use, the 11 visualizations that were generated for the results are:  
-behavior\_sexualRisk  
-behavior\_eating  
-behavior\_personalHygine  
-intention\_commitment  
-attitude\_consistency  
-perception\_vulnerability  
-norm\_fulfillment  
-empowerment\_knowledge  
-empowerment\_abilities  
-empowerment\_desires  
  
As all the data sets consisted of the same rating system, they were all processed as stacked bar-graphs that drew distinction amongst participants who had cervical cancer and those who don’t.  
  
Scatterplots were made to show potential correlation and connections of the following variables.  
- behavior\_personalHygine/ behavior\_eating

- attitude\_consistency/ behavior\_eating

- attitude\_consistency/behavior\_personalHygine

- attitude\_consistency/ intention\_commitment

- empowerment\_knowledge/ empowerment\_abilities

- empowerment\_abilities/empowerment\_desires

- empowerment\_knowledge/perception\_vulnerability

- empowerment\_knowledge/behavior\_personalHygine

- empowerment\_desires/norm\_fulfillment

- empowerment\_desires/empowerment\_knowledge

Sklearn library was used for the data modelling section, of the classification tasks, we chose two methods, which are the k neighbour classifer and decision Tree Classifier to predict whether an individual has cervical cancer based on their attributes, since all values were of similar weighting, there was minimal manipulation. The classification were divided into 3 tasks

Task 1: 0.5 of the data for training, the other 0.5 for testing  
task 2: 0.6 of the data for training, the other 0.4 for testing  
task 3: 0.8 of the data for training, the other 0.2 for testing  
  
in each report: precision, recall and f1-scores were provided for each label 0 = no cancer, 1 = has cancer, accuracy, macro accuracy and weighted accuracy were given [MORE ELABORATION ON KNEIGHBOUR/DECISION TREE]

DATA EXPLORATION ANALYSIS:

Data exploration elaboration had to be speculated, due to the language divide and incomplete explanation of the social skill variables, the exact nature of the graphs representation had to be speculated.

Chart, pie chart

Description automatically generated

This pie chart shows the distribution of those without and with cervical cancer for context (blue and orange respectively). And the colours of the graphs will consistently dictate blue as cancer free, and orange/red as those with cervical cancer.

A picture containing text

Description automatically generated

graph 1: behavior\_sexualRisk  
From this graph, very few were willing to engage in risky sexual intercourse with or without cervical cancer, although those who were willing, indicated that they already had cervical cancer.  
  
Chart, histogram

Description automatically generated

graph 2: Behaviour\_eating  
The individuals tested indicated that the majority indicated their eating behavior highly, those with cancer probably were more health conscious due to their diagnosis, whereas the few that rated their diet low likely had less health concerns.

Chart, bar chart

Description automatically generated

Table 3: behavior\_personalHygiene  
The ratings for personal hygiene show that most of those tested rating their hygiene as high, although of those who have cervical cancer, their distribution was spread evenly across the graph.Chart

Description automatically generated

Table 4: intention\_commitment

Generally it was shown the the majority of participants rated their intentions of committing very highly.

Chart

Description automatically generated

Table 5 Constitency\_attitude:  
participants attitudes were rated generally above average, although it is surprising that most of those who had cancer were optimistic based on interpretation of the graph

Chart

Description automatically generated

Table 6 perception\_vulnerability:  
It was generally shown, based on my interpretation of the graph, that generally most of those with cancer saw themselves as vulnerable, and although some of those without cancer shared this sentiment, of those who did not believe that their were vulnerable, were entirely cancer free.

Chart, bar chart

Description automatically generated

Table 7: norm\_fulfillment  
This tables interpretation came across as a rating about the fulfillment the individuals felt towards societal norms of their country. Indonesia is a conservative country, as such, it is a reasonable hypothesis that more liberal behaviours based on sexual freedom in this instance, run counter to the countries general conservative norms, leading to those unfulfilled by said norms to engage in more risky behaviour, This is reflected by those whose ratings are lower have the majority of cancer cases.

Chart

Description automatically generated

Table 8: empowerment\_knowledge  
This variable is interpreted as the rating of empowerment based on how much knowledge they have about the disease, those without cancer were found to generally feel empowered potentially due to having the knowledge to act towards prevention, whilst those with cancer likely did not feel empowered due to already contracting the illness or became optimistic due to gaining more information.

Chart, waterfall chart

Description automatically generated

Table 9: empowerment\_abilities  
This variable explored their empowerment based on the abilities of the individual, this graph indicates that many of those with cancer felt hopeless, whilst those without generally felt more empowered with their abilities.

Chart, bar chart

Description automatically generated

Table 10: empowerment\_desire  
The graph shows that a large proportion of those with cervical cancer, rated their desires did not empower them, potentially due to a belief of pessimistic outcome, although those who did not have cancer were empowered by their desires.

Chart, scatter chart

Description automatically generated

Table 11: hygiene ratings crossed with eating rating.

We hypothesized that those with good hygiene would engage in a healthier diet, there proved to be a minor positive correlation to hygiene to diet, however the correlation was weak.

Chart, scatter chart

Description automatically generated

Table 14: Consistency to commitment  
  
Consistency is often linked to either commitment or habit, we tested for a correlation between these 2 aspects, however, their showed to be almost zero correlation between the 2 variables which went against our initial assumptions.

Chart, scatter chart

Description automatically generated

Table 15: abilities and knowledge  
  
There was a strong correlation between the empowerment rating of those with abilities and knowledge, visually it was clear that those who did not feel like they had either rated low on both of these graphs, and there is a clear positive correlation. We hypothesized that those with higher rated abilities, would in turn be able to act on the knowledge the received, and thus felt empowered, conversely the opposite was shown to be true too.

Chart, scatter chart

Description automatically generated

Table 20:  
  
We hypothesized that those with high desire ratings would lead to higher knowledge ratings, as desire translates to action, and thus a pursuit of knowledge. The Strong positive correlation indicates that generally these individuals would rate highly across both variables. There is a clear visual divide as well amongst those with cancer and those without, leading to a potential indication that generally those with cancer did not feel or generally felt less empowered overall. There are several small outliers, but the clustering of those with cancer in table 20 and 15 are a clear indication, that generally those who are already suffering cervical cancer do not feel empowered on average.

TASK 3 ANALYSIS  
  
CONCLUSION  
  
REFERENCES(probably reformat a bit):

[UCI Machine Learning Repository: Cervical Cancer Behavior Risk Data Set](http://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk)

[sobar-rizanda-adi-asl-2016.pdf (wordpress.com)](https://adivb.files.wordpress.com/2017/11/sobar-rizanda-adi-asl-2016.pdf)

https://slideplayer.com/slide/11492208/