We certify that this is all our own original work. If we took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission. We will show we agree to this honor code by typing "Yes": ***Yes.***

Practical Data Science Research on PDS Cervical Cancer Dataset

Authors: Vijit Kumar (s3799493)

Chenyu Xiao (s3829221)

Affiliation: RMIT

Contact Details: s3799493@student.rmit.edu.au

s3829221@student.rmit.edu.au

Date of Report: 20 May 2021

Table of Contents

Abstract

Nowadays, cervical cancer is a very severe disease suffered by females around the globe. Despite its incurability, cervical cancer can still be prevented if proper precautionary measures are taken. In this report, two classifiers, K Nearest Neighbours (KNN) and Decision Tree (DT), are used as the classifiers to detect whether a female has ca cervix based on a series of her personal health behaviour and recognise which behaviours have significant influence on ca cervix. The result of classification shows that our KNN classifier is able to predict ca cervix with the test accuracy of 91% and our DT classifier has the test accuracy of 86%.

1. Introduction

Nowadays, cervical cancer is a very severe disease suffered by females around the globe. With approximately 11.7% of females worldwide suffering from cervical cancer [1], it is currently the second most prevalent cancer among females [2]. One of the reasons why cervical cancer is so prevalent is that people, especially those from the middle or low class, start caring about their health issues only after they notice some symptoms developing on them. And since cervical cancer cannot be completely cured, it is already too late when those people start seeking medical treatment.

Despite its incurability, cervical cancer can still be prevented if proper precautionary measures are taken. Therefore, it is vital for females to pay attention to their personal health behaviour, which can greatly reduce the risk of getting cervical cancer.

The research goal of this report is to find an accurate classifier to identify whether the respondent is going to have cervical cancer or not given a set of her characteristics and personal health behaviours. Moreover, by finding out which behaviours and characteristics have significant influence on cervical cancer, we hope to generate a guide that informs females which health behaviours they should pay enough attention to prevent cervical cancer.

2. Methodology

2.1 Feature Information

In the field of health science and psychology there are many theories that are related to behaviour. According to some psychological behaviour of a person is affected by the intention which in turn is determined by three factors namely: attitudes, subjective norms, and perceived behavioural control [3]. Furthermore, according to the Social Cognitive Theory, the attitude and subjective norm when interacted with perceived control independently are strong predictors of intentions [4]. All together the seven major predictors of a behaviour namely: perception, intention, motivation, subjective norm, attitude, social support, and empowerment [5].

The dataset we used for this task is called “Cervical Cancer Behaviour Risk Data Set” which was obtained from the dataset archives of University of California, Irvine. This dataset consists of 72 instances with each instance representing the questionnaire filled by the respondents. This dataset consists of behavioural related questions in a likert scale manner thus each feature in this dataset is an ordinal categorical variable. In total there are 19 features which are sub-categories of the seven variables that are major predictors of a behaviour as mentioned earlier. The target variable is ca_cervix that contains class 0 and class 1. Class 0 represents the cancer negative and class 1 represents the cancer positive. The features in the dataset are shown in Table 1.

Table 1: Feature Information

Feature	Type	Feature	Type	Feature	Type
behavior_sexualRisk	int	behavior_personalHygine	int	behavior_eating	int
intention_aggregation	int	intention_commitment	int	attitude_consistency	int
attitude_spontaneity	int	norm_significantPerson	int	norm_fulfillment	int
empowerment_desires	int	empowerment_knowledge	int	empowerment_abilities	int
socialSupport_appreciation	int	socialSupport_emotionality	int	socialSupport_instrumental	int
motivation_willingness	int	motivation_strength	int	perception_vulnerability	int
perception_severity	int	**ca_cervix (class)**	int		

2.2 Feature Selection

Before splitting the data, we implemented the hill climbing method to extract the important features from the dataset that could help to eliminate the features from the dataset that are not useful. Since the hill climbing method requires the use of a heuristic function, we used the baseline model of each classifier to extract the important features. After the feature selection we split the data into three parts: training (50%), validation(20%) and testing (30%) set.

The 15 selected features for KNN classifier are: 'intention_aggregation', 'attitude_spontaneity', 'norm_significantPerson', 'empowerment_knowledge', 'socialSupport_instrumental', 'behavior_personalHygine', 'perception_vulnerability', 'perception_severity', 'empowerment_abilities', 'motivation_strength', 'socialSupport_appreciation', 'behavior_sexualRisk', 'socialSupport_emotionality', 'norm_fulfillment', 'behavior_eating'.

The 9 selected features for DT classifier are: 'intention_aggregation', 'behavior_personalHygine', 'perception_vulnerability', 'perception_severity', 'empowerment_abilities', 'motivation_strength', 'behavior_sexualRisk', 'motivation_willingness', 'behavior_eating'.

2.3 Classification for early detection of Cervical cancer

In this study we use K Nearest Neighbours (KNN) and Decision tree (DT) classifiers for the early detection of cervical cancer based on the behaviour determinants of the respondents. KNN is a simple and yet an efficient algorithm which is often used in quick calculation time. It stores the training dataset and learns from it only at the time of making real time predictions. Therefore, this makes it much faster than other ML algorithms such as SVM, Logistic Regression etc [6].

DT is also a common algorithm which is often used in classification tasks. In fact, it is among the top 10 algorithms for data mining [7]. DT does not require normalization and scaling of the data which makes it an appropriate choice for categorical features. Furthermore, during the pre-processing task it is not necessary to handle the missing values in the dataset.

In order to train, test and tune the model we have divided the dataset into three parts. The first part is the training set on which the model is trained. The second part is the validation set which is used to fine tune the parameters of the classifier algorithm which will give us higher accuracy. The third part is the testing set which is used to evaluate the performance of the model trained. The ratio we have used is 50-20-30 for the training, validation and testing set respectively. In order to show the

importance of the validation set, we first run the algorithm directly on the test set, and then tune the parameters using the validation set before re-assessing the testing set and comparing the results.

After the training of the model from the given dataset, the next step is to test the performance of the model on the testing dataset. In order to test the performance, there are many different types of performance metrics available for classification and regression tasks. This study uses accuracy and confusion matrix to measure our model's performance. In a confusion matrix the results are stored in the form of following values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP means when the actual value is 1 and the model also predicted 1 in case of cancer positive. Similarly, when the actual value is 0 and the model correctly predicts 0 for non-cancer cases then it is TN. FP is when the actual value is 0 but the model labelled it as 1. And similarly, FN is then actual value is 1 but the model labelled it as 0. Another metric that we are using is Accuracy which is calculated using the four values of the confusion matrix. The formula for accuracy is as follows:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN)$$

3. Results

3.1 Data Exploration

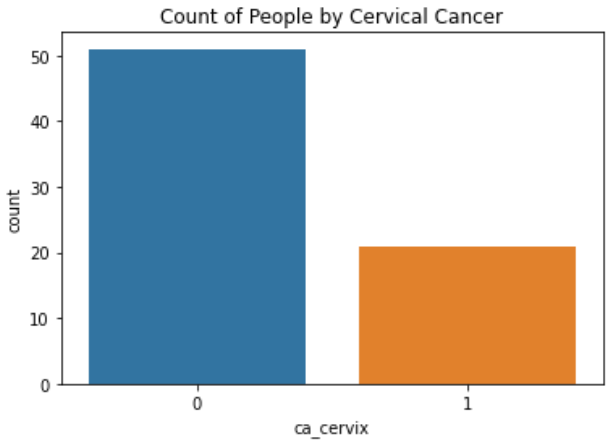
3.1.1 Individual Attributes

To explore each attribute individually, we firstly gather the descriptive statistics for all the attributes in a table then plot histogram or count plot for 10 selected attributes to display their data distribution. According to the descriptive statistics, the means of most attributes are very close to the maximum values and far away from the minimum values. Therefore, most attributes are highly skewed.

We divide the 10 graphs into 4 groups according to the characteristics of data distributions. Here we show one graph for each group.

(1). Cervical Cancer (class attribute)

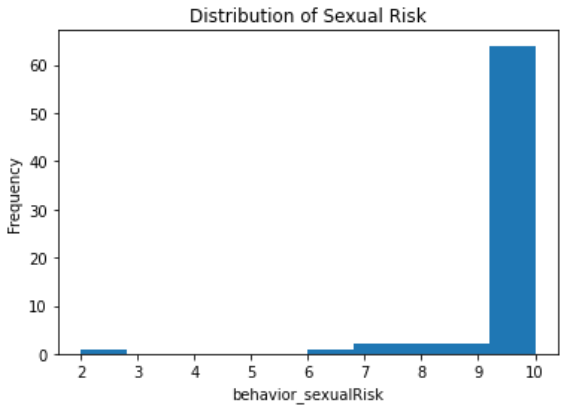
Figure 1: Count Plot for ca_cervix



The count of people tested positive for cervical cancer is more than twice as those tested negative. It indicates there is data imbalance issue within the class attribute.

(2). Sexual Risk, Personal Hygiene, and Eating

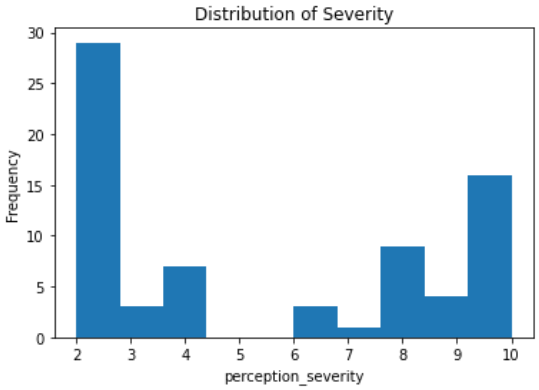
Figure 2: Histogram for behavior_sexualRisk



Over 80% of instances have sexual risk higher than 9 which is close to the maximum value of this attribute. It can make sexual risk insignificant in explaining the cause of cervical cancer. The other attributes having the similar negatively skewed distribution are personal hygiene and eating, with the characteristic that the mode and mean are very close to the maximum value of the corresponding attribute.

(3). Severity, Vulnerability, Knowledge, and Desire

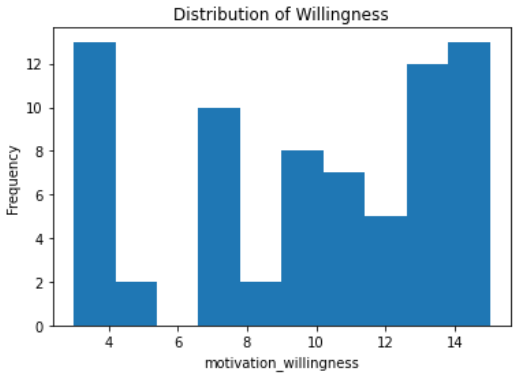
Figure 3: Histogram for perception_serivity



The majority of instances in these four attributes are either close to the minimum value or the maximum value, but there is only a small number of instances in the middle.

(4). Willingness and Ability

Figure 4: Histogram for motivation_willingness

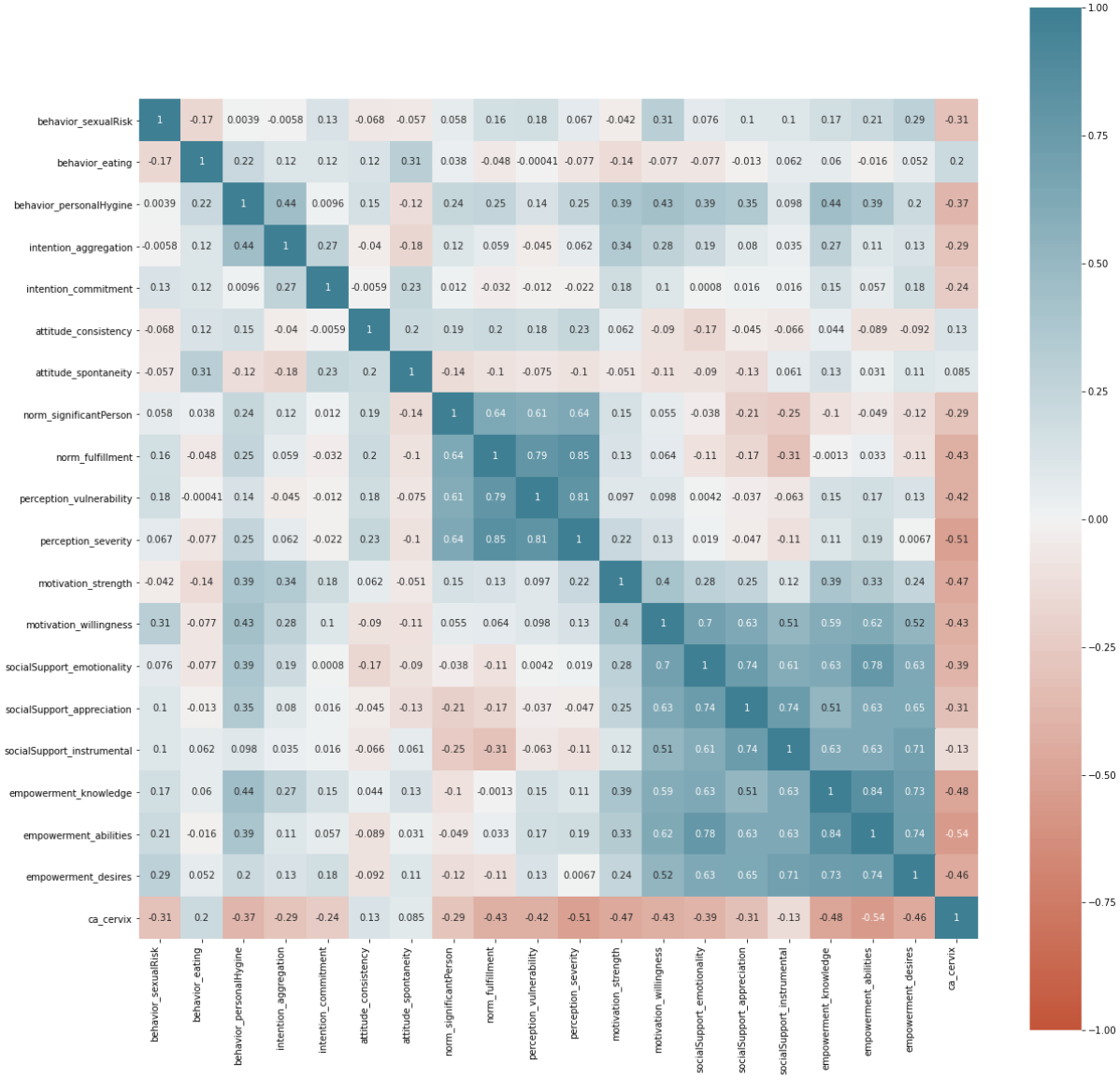


There is no central tendency observed from the distribution of data in these two attributes. Therefore, we conclude that they follow random distribution.

3.1.2 Attribute Pairs

We firstly use a heat map to provide an overview of correlations between all attribute pairs. The blue cells indicate positive correlations between features in the corresponding rows and columns whereas red cells indicate negative correlations. Additionally, darker colour represents stronger correlation whereas lighter colour indicates weaker correlation.

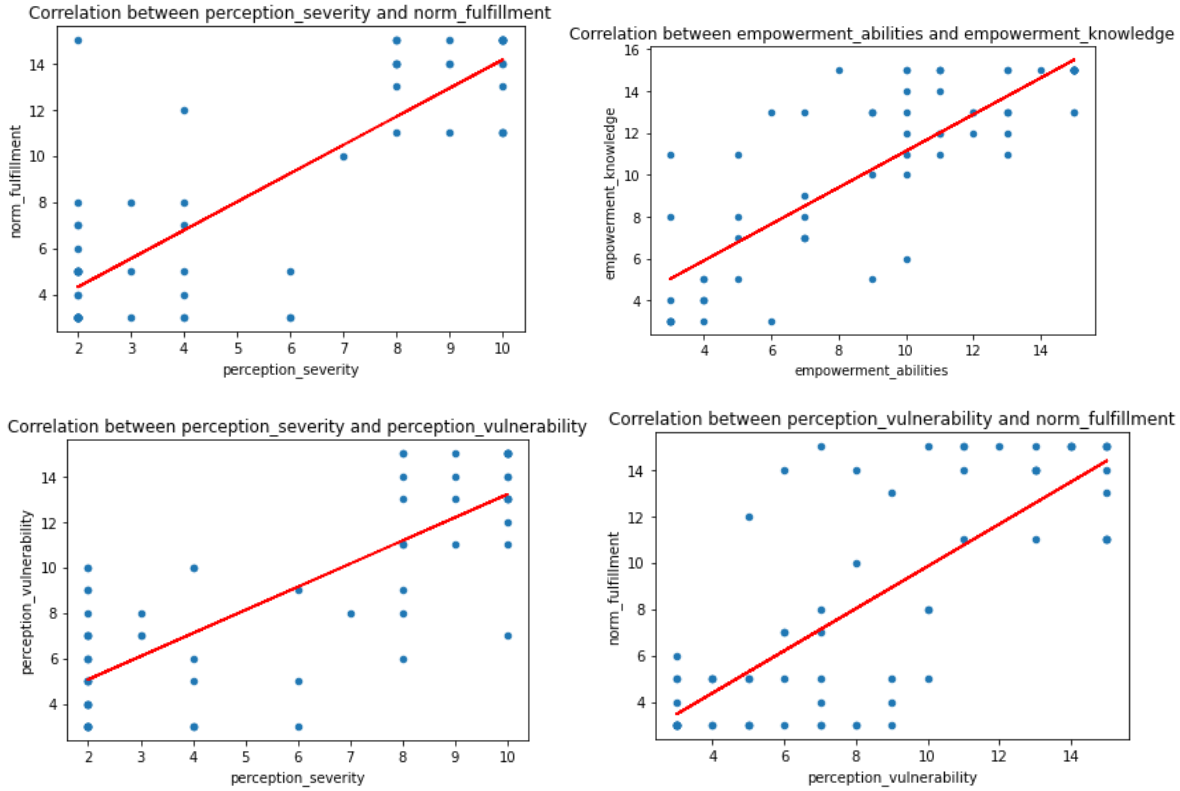
Figure 5: Heat Map of All Attribute Pairs



According to the distribution of dark blue cells in the heat map, most of the mild-to-strong positive correlations (greater than 0.6) appear between the norm and perception variables and also between the social support and empowerment variables. All of the cells in the target row/column are light red (i.e., insignificant correlation), indicating that no single feature can explain the target attribute by itself. However, since none of the cells has correlation below -0.55, there is no strong negative correlation observed between any pair of attributes.

We select 10 pairs of attributes and draw a scatter plot to visualise the correlation for each attribute pair. We observe that, due to the ordinal data type and insufficient data samples, the scatters do not gather tightly around the trendline. The following scatter plots are those representative ones.

Figure 6: Scatter Plot and Linear Fit for Representative Attribute Pairs



The hypothesis of the relationship between each pair of attributes as well as the actual observation are also listed below in Table 2.

Table 2: Hypothesis and Observation on Feature Pairs

Attribute Pair	Hypothesis	Observation
behavior_sexualRisk ca_cervix	People with higher sexual risk are supposed to have higher risk of getting ca cervix (positive relationship)	The hypothesis is false, and the correlation is insignificant (-0.31).
behavior_personalHygine ca_cervix	People who pay more attention to personal hygiene are less risky in terms of ca cervix (positive relationship).	The hypothesis is false, and the correlation is insignificant (-0.37).
perception_severity norm_fulfillment	People that take ca cervix more seriously take more actions to prevent ca cervix (positive relationship)	The hypothesis is true, and the correlation is strongly positive (0.85).
empowerment_abilities empowerment_knowledge	People with stronger financial ability do not necessarily have more knowledge on ca cervix and vice versa (no relationship).	The hypothesis is false, and the correlation is strongly positive (0.84).
perception_vulnerability perception_severity	People who are more susceptible to ca cervix tend to take it more seriously (positive relationship).	The hypothesis is true, and the correlation is strongly positive (0.81).

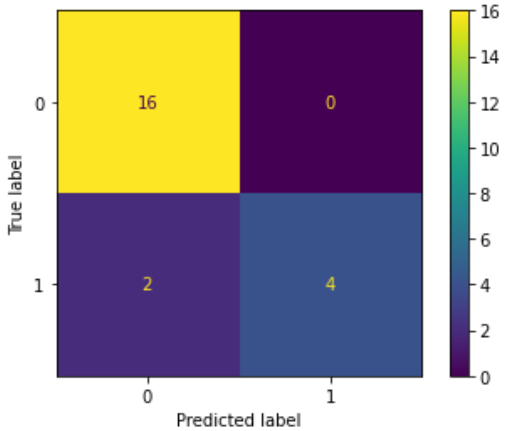
perception_vulnerability norm_fulfillment	People who are more susceptible to ca cervix tend to take more actions to prevent ca cervix (positive relationship).	The hypothesis is true, and the correlation is mildly positive (0.79).
empowerment_abilities socialSupport_emotionality	People with stronger financial ability do not necessarily feel more emotionally supported and vice versa (no relationship).	The hypothesis is false, and the correlation is mildly positive (0.78).
socialSupport_appreciation socialSupport_emotionality	People getting more appreciation naturally feel more emotionally supported (positive relationship).	The hypothesis is true, and the correlation is mildly positive (0.74).
empowerment_desires empowerment_abilities	People with more desire of preventing ca cervix do not necessarily have stronger financial ability and vice versa (no relationship).	The hypothesis is false, and the correlation is mildly positive (0.74).
empowerment_desires empowerment_knowledge	People with more desire to prevent ca cervix tend to seek more information and acquire more knowledge about ca cervix (positive relationship).	The hypothesis is true, and the correlation is mildly positive (0.73).

3.2 Data Modelling

In this study we conducted two experiments to evaluate the behaviour determinant as attributes in order to classify whether the respondent has cervical cancer or not with KNN and DT classifiers. The training and validation set were used to train and hyper-tune the parameters of the model, and finally the performance of the model was evaluated on the testing set.

The results of the first experiment on the testing set using the KNN model are shown in Figure 7. From the confusion matrix we can see that the accuracy of the model on the testing set is approximately 91%.

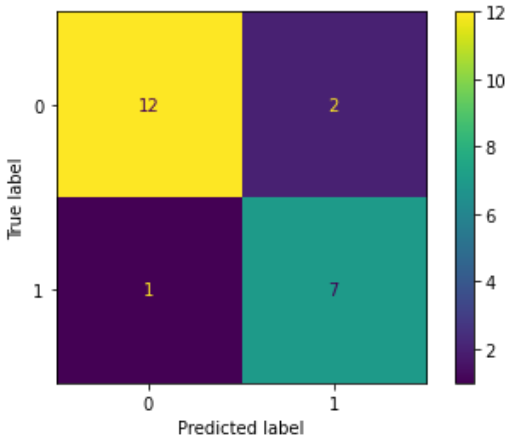
Figure 7: Confusion matrix on KNN



Accuracy =(TN+TP)/(TN+FP+FN+TP) = (16+4)/(16+0+2+4) = 0.909

The results of the second experiment on the testing set using the DT model are shown in Figure 8. From the confusion matrix we can see that the accuracy on the testing set is 86%. However, due to the randomness of tree splitting, the result can be different each time the classifier is executed.

Figure 8: Confusion matrix on DT



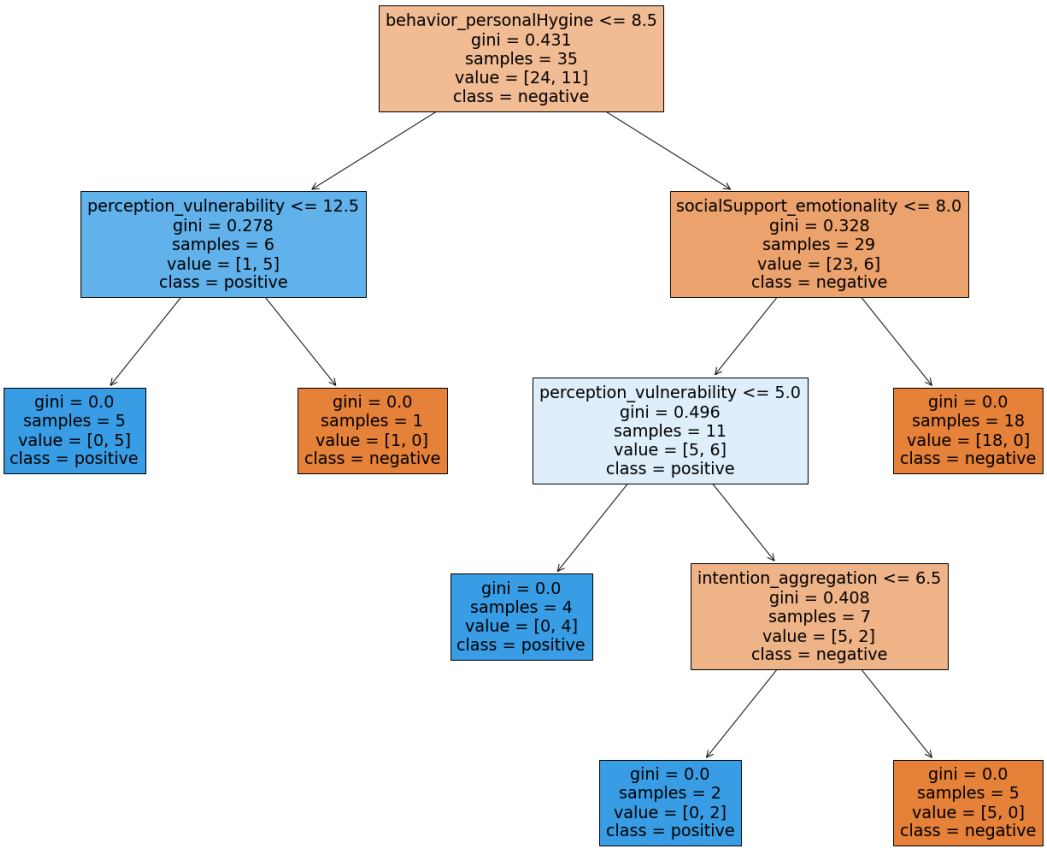
Accuracy =(TN+TP)/(TN+FP+FN+TP) = (12+7)/(12+2+1+7) = 0.864

4. Discussion

The dataset used in this study is quite small with only 72 instances. In the case of KNN we saw an accuracy of 91% which is quite remarkable considering the number of instances we have. Furthermore, it should also be noted that the data has high variance as with each random sampling of data showed us a variation in the accuracy.

In the case of DT, the accuracy on the testing set was 86% which is almost close to KNN. The primary difference between the results of the two experiments is that DT has one FN value whereas in the case of KNN there is none. This means that the KNN is performing better in predicting if the respondent does not have Cervical cancer. In the case of predicting whether the respondent is Cancer positive (class 1), both the models had similar results. It should also be noted that with each run of DT classifier there will be a difference in the results. As for the parameter tuning, we are using the max_features as ‘sqrt’ which will always choose a set of features to build the tree. Below we have attached a tree from the DT model on the testing set to show the split of relevant features.

Figure 9: Visualisation of Decision Tree



From the above tree we can see that some of the relevant features in determining the Cancer positive class are behavior_personalHygine, perception_vulnerability, socialSupport_emotionality, and intention_aggregation. However, the decision tree can be different each time the classifier is executed.

It should also be noted that since the testing dataset is small with only 16 instances of class 0 and 6 instances of class 1, there can be a huge variation in the accuracy of the models with even 1 or 2 misclassifications. Therefore, it is imperative to look for more dataset as the more generalized number of instances can prove to be better for classification tasks in machine learning.

5. Conclusion

Ca Cervix is a concerning public health problem among women throughout the world but can be prevented if detected at the right time. In this study we attempted to predict the Cervical cancer cases based on the behavioural patterns of the respondent. This could potentially be beneficial research as this might help in reducing the mortality rates and even increase the effectiveness of the treatment. Predicting the cancer based on the behaviour determinant is a cheaper approach than any other on site testing approaches.

Based on the experiments we conducted, the KNN classifier outperformed the DT classifier and showed promising results in the early classification of the cancer with an accuracy of 91%. However, due to the small dataset we had a variation in the testing results. We believe that with more relevant dataset we can certainly bring up the accuracy on the unseen data.

References

[1] A. F. Rositch, A. Gatuguta, R. Y. Choi, B. L.Guthrie, R. D. Mackelprang, R. Bosire, L.Manyara, J. N. Kiarie, J. S. Smith, and C. Farquhar,"Knowledge and acceptability of Pap smears, self-sampling and HPV vaccination among adult women in Kenya," *PLoS One*, vol. 7, no. 7, 2012.

[2] F. H. Zhao, S. M. Tiggelaar, S. Y. Hu, L. N. Xu, Y.Hong, M. Niyazi, X. H. Gao, L. R. Ju, L. Q.Zhang, X. X. Feng, X. Z. Duan, X. L. Song, J.Wang, Y. Yang, C. Q. Li, J. H. Liu, J. H. Liu, Y. B.Lu, L. Li, Q. Zhou, J. F. Liu, N. Zhao, J. E.Schmidt, and Y. L. Qiao, "A multi-center survey of age of sexual debut and sexual behavior in Chinese women: Suggestions for optimal age of human papillomavirus vaccination in China," *Cancer Epidemiol*., vol. 36, pp. 384–390, 2012.

[3] T. L. Webb and P. Sheeran, "Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence.," *Psychol. Bull*., vol. 132, no. 2, pp. 249–268, 2006

[4] J. P. Dillard, "An Application of the Integrative Model to Women's Intention to Be Vaccinated Against HPV: Implications for Message Design.," *Health Commun*., vol. 26, no. January, pp. 479–486, 2015

[5] Sobar, Machmud, R., & Wijaya, A. (2016). "Behavior Determinant Based Cervical Cancer Early Detection with Machine Learning Algorithm". *Advanced Science Letters*, 22(10), 3120â€"3123.

[6] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, "Top 10 algorithms in data mining," *Knowl. Inf. Syst*., vol. 14, no. 1, pp. 1–37, 2008.

[7] H. Xiao, F. Sun and Y. Liang, "A Fast Incremental Learning Algorithm for SVM Based on K Nearest Neighbors," *2010 International Conference on Artificial Intelligence and Computational Intelligence*, 2010, pp. 413-416, doi: 10.1109/AICI.2010.207.