# COSC 2670
# Practical data science with Python
# Assignment 2

# Screening cervical cancer

Team members:

s3560484 Ailin Dou
s3560484@student.rmit.edu.au
RMIT University,
Melbourne Australia

s3749857 Ziqing Yan
s3749857@student.rmit.edu.au
RMIT University,
Melbourne Australia

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*.

Date of report: 18/05/2021

# Table of Contents



# Table of Figures

## 1. Executive summary:

The purpose of the assignment is to accurately predict whether a patient has cervical cancer through the related cervical cancer data set, which involves data processing, and data exploration to avoid data clutter. Two classifiers are trained to predict the labels, which are k-Nearest Neighbors and the Decision Tree. The result shows that both two classifiers perform well in this data set and the average accuracy of k-Nearest Neighbors is higher than that of the Decision Tree. The report concludes that the two classifiers trained by this data set may not be stable because the size of the data set is too small. It is recommended that people can use k-Nearest Neighbors to screen cervical cancer if they do not have much time to train models by different training sets. Otherwise, it is better to choose the Decision Tree that can produce tree visualization and make it easier for people to understand the results.

## 2. Introduction:

Cervical cancer is a common gynecological malignancy that is the increase of abnormal cells within the lining of the cervix in females. According to the surveys, Cancer Council [1][2] presented that cervical cancer is gradually becoming younger, teenagers also have the risk of cervical cancer so that how to screen and judge for cervical cancer has become the most important step in the medical field. In addition, Cancer Council [3] also described that the large loop excision of the transformation zone (LLETZ) is currently the most advanced technique for screening for cervical cancer, but the LLETZ method is still inadequate that leads to a low diagnosis rate so that missing the best time for treatment [4]. Therefore, we are considering analyzing the relevant data and developing suitable data modeling to improve the accuracy of screening for cervical cancer by the data science approach for providing appropriate services in the medical field.

## 3. Methodology:

According to the study, The Cervical Cancer Behavior Risk data set is collected from UCI Machine Learning Repository [5]. In addition, the data set 19 features of each patient that are divided into 8 major sections to analyze for prediction of cervical cancer, including the personal behaviors, the individual intentions, the personal attitudes, the personal norms, the individual perceptions, the individual motivations, the social supports, the empowerments to build machine learning models for screening of cervical cancer. We use Pandas and Numpy, which belong to the Python data analysis library, to complete the process of data retrieving and data preparation. We also use Matplotlib, the Python visualization library, to create a variety of graphs for data exploration. For data modelling, we use the Scikit-learn library in Python to build classifiers(k-Nearest Neighbors and Decision Tree).

### 3.1 Data retrieving

The task aims to improve the accuracy of cervical cancer screening by analysing the specific data and establishing proper data modelling to provide services for the medical field. Thence, the data retrieving and data cleaning become an essential part of the task, which has to be completed before the data modelling. Moreover, the obvious errors and missing values need to be retrieved and manipulated by the most suitable methods in data retrieving and data cleaning. If this step is omitted, the errors and missing values will remain in the dataset, leading to increasing the modelling deviations in the task. In addition, the size of the data set and the data types of the variables in the data set need to be viewed firstly, and then using the specific function to check which variables exist missing values and obvious errors, especially calling the is.null() function and the value_counts () function. Thus, there are not missing values and obvious errors in the cervical cancer data set, which are able to establish and analyse the data modelling by the original data set.

### 3.2 Data exploration

The data exploration has been divided into two parts in the data set, including demonstrating the data visualisation and analysing the specific descriptive statistics of each variable so that combining the bar chart and histogram to show the data visualisation is the best method to provide definite information for people. Also, discovering the relationship between the ca_cervix and other variables in the data set by data visualisations, thereby analysing initially whether there is a direct or indirect association between judging for cervical cancer and other variables in the data set. If a strong correlation appears between the variables and the ca_cervix column, which should focus on the roles of these variables in the data modelling in the data set.

### 3.3 Data modelling

We first implement feature selection with the method Hill Climbing for each classifier to improve classification accuracy of classifiers. Then, the two classification models are compared in terms of their classification accuracy

# 4. Results:

## 4.1 Data retrieving

The Pandas and Numpy libraries were called to support us to search for the missing values and obvious errors in the cervical cancer data set by the based functions. In addition, there are no missing values and potential issues with a complete data set, and then writing the checked data into a .csv file called 'cleaned_Cervical_Cancer.csv' file.

## 4.2 Data Exploration

The data exploration was used to discover the specific information and the relationships between all variables in the cervical cancer data set. Moreover, the descriptive statistics, the data visualisations of the major variables and their relationships can be displayed below.

### 4.2.1 The descriptive statistics in the data set

There are 72 patients' information recorded in the data set. The descriptive statistics were shown below to show the specific information of each variable in the cervical cancer data set in **Figure 1, Figure 2, and Figure 3**.

| | behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_spontaneity |
|---|---|---|---|---|---|---|---|
| count | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 |
| mean | 9.666667 | 12.791667 | 11.083333 | 7.902778 | 13.347222 | 7.180556 | 8.611111 |
| std | 1.186782 | 2.361293 | 3.033847 | 2.738148 | 2.374511 | 1.522844 | 1.515698 |
| min | 2.000000 | 3.000000 | 3.000000 | 2.000000 | 6.000000 | 2.000000 | 4.000000 |
| 25% | 10.000000 | 11.000000 | 9.000000 | 6.000000 | 11.000000 | 6.000000 | 8.000000 |
| 50% | 10.000000 | 13.000000 | 11.000000 | 10.000000 | 15.000000 | 7.000000 | 9.000000 |
| 75% | 10.000000 | 15.000000 | 14.000000 | 10.000000 | 15.000000 | 8.000000 | 10.000000 |
| max | 10.000000 | 15.000000 | 15.000000 | 10.000000 | 15.000000 | 10.000000 | 10.000000 |

*Figure 1: The descriptive statistics of the variables in the cervical cancer data set 1*

| | norm_significantPerson | norm_fulfillment | perception_vulnerability | perception_severity | motivation_strength | motivation_willingness | socialSupport_emotionality |
|---|---|---|---|---|---|---|---|
| count | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 |
| mean | 3.125000 | 8.486111 | 8.513889 | 5.388889 | 12.652778 | 9.694444 | 8.097222 |
| std | 1.845722 | 4.907577 | 4.275686 | 3.400727 | 3.207209 | 4.130406 | 4.243171 |
| min | 1.000000 | 3.000000 | 3.000000 | 2.000000 | 3.000000 | 3.000000 | 3.000000 |
| 25% | 1.000000 | 3.000000 | 5.000000 | 2.000000 | 11.000000 | 7.000000 | 3.000000 |
| 50% | 3.000000 | 7.000000 | 8.000000 | 4.000000 | 14.000000 | 11.000000 | 9.000000 |
| 75% | 5.000000 | 14.000000 | 13.000000 | 9.000000 | 15.000000 | 13.000000 | 11.250000 |
| max | 5.000000 | 15.000000 | 15.000000 | 10.000000 | 15.000000 | 15.000000 | 15.000000 |

*Figure 2: The descriptive statistics of the variables in the cervical cancer data set 2*

| | socialSupport_appreciation | socialSupport_instrumental | empowerment_knowledge | empowerment_abilities | empowerment_desires | ca_cervix |
|---|---|---|---|---|---|---|
| count | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 | 72.000000 |
| mean | 6.166667 | 10.375000 | 10.541667 | 9.319444 | 10.277778 | 0.291667 |
| std | 2.897303 | 4.316485 | 4.366768 | 4.181874 | 4.482273 | 0.457719 |
| min | 2.000000 | 3.000000 | 3.000000 | 3.000000 | 3.000000 | 0.000000 |
| 25% | 3.750000 | 6.750000 | 7.000000 | 5.000000 | 6.750000 | 0.000000 |
| 50% | 6.500000 | 12.000000 | 12.000000 | 10.000000 | 11.000000 | 0.000000 |
| 75% | 9.000000 | 14.250000 | 15.000000 | 13.000000 | 15.000000 | 1.000000 |
| max | 10.000000 | 15.000000 | 15.000000 | 15.000000 | 15.000000 | 1.000000 |

*Figure 3: The descriptive statistics of the variables in the cervical cancer data set 3*

### 4.2.2 Searching the main variables in the cervical cancer

The ca_cervix variable included 0, and 1, where 0 means not cervical cancer shown in pick one, 1 means cervical cancer revealed in the blue one. Around 50 people have cervical cancer and approximately 20 people who do not have cervical cancer were shown in the

data set. Besides, people without cervical cancer more than those having cervical cancer in the data set, which is shown in **Figure 4**.

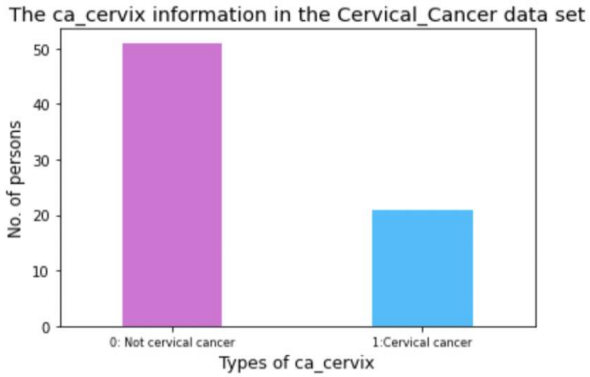


*Figure 4: The ca_cervix variable in the cervical cancer data set*

The behavior_personalHygine variable was unevenly distributed, it exposed a right-skewed distribution. Moreover, the highest number of people in the interval from 14 to 15 was up to 20 patients in this range, but none of the patients were in the interval between 6 and 7 in this variable of the cervical cancer data set, which is epitomised in **Figure 5**.

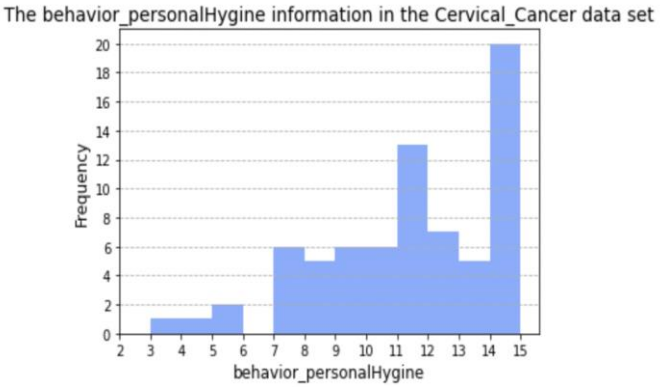


*Figure 5: The behavior_personalHygine variable in the cervical cancer data set*

The largest number of people in the empowerment_abilities variable was located in the interval of 14-15 as 14 patients. Also, there are more than 8 patients in the range of 3-4, but the interval of 8-9 is less than 2 patients in the empowerment_abilities variables of the data set, which is epitomised in **Figure 6**.

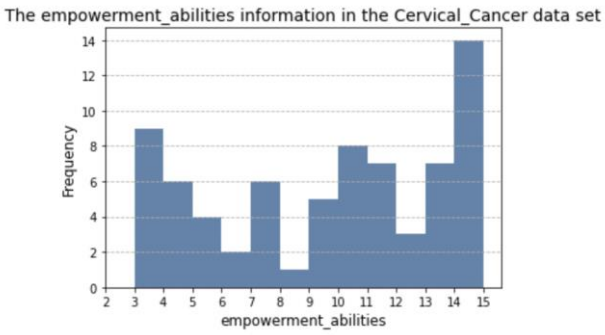


*Figure 6: The empowerment_abilities information of the cervical cancer data set*

The largest number of people in the empowerment_desires variable was located in the interval between 14 and 15, which had 25 patients. This is followed by more than 10 patients in the range of 3-4. Whereas, there are less than 5 patients in the interval of 6-7 and 8-9. The specific data shown in **Figure 7**.

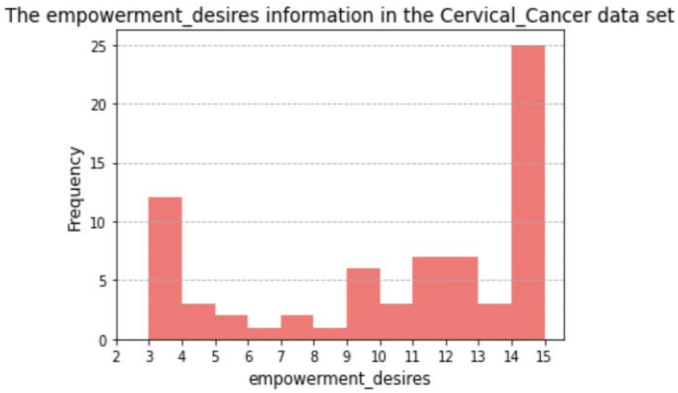


*Figure 7: The empowerment_desires variable in the cervical cancer data set*

### 4.2.3 Exploring the relationships between the ca_cervix and other features in the data set

Using box plots to analyse the primary relationships between the ca_cervix and other elements in the data set.

Based on **Figure 8**, the median, minimum, and maximum number of the behavior_personHygine variable could be found in the box plot, but the median is the most important parameter to figure the relationship in the box plot. Also, a minor interval existed between the “0: Not cervical cancer” and the “1: Cervical cancer”. Thus, once the personal hygine (behavior_personHygine) is larger, the lower the risk of getting cervical cancer. Oppositely, the lower the personal hygine (behavior_personHygine) lower, the higher risk of cervical cancer.

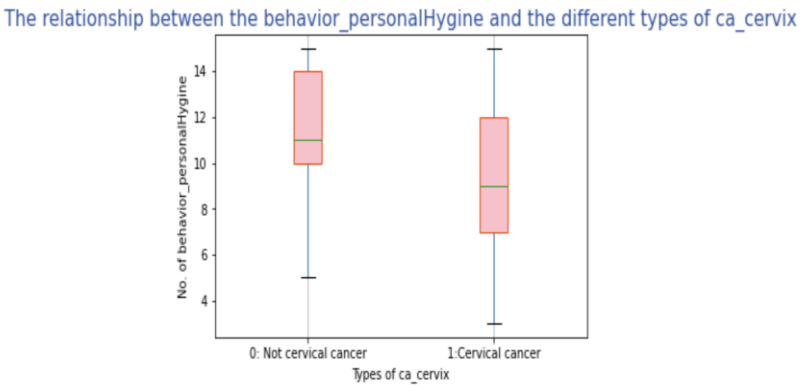


*Figure 8: The relationship between behavior_personalHygine and the ca_cervix*

Based on **Figure 9**, the median, minimum, and maximum number of the empowerment_abilities variable could be found in the box plot, but the median is the most important parameter to figure the relationship in the box plot. Also, a big interval existed between the “0: Not cervical cancer” and the “1: Cervical cancer”. Thus, the

empowerment_abilities feature is higher, the lower the risk of getting cervical cancer. Oppositely, the empowerment abilities lower, the higher risk to gain cervical cancer.

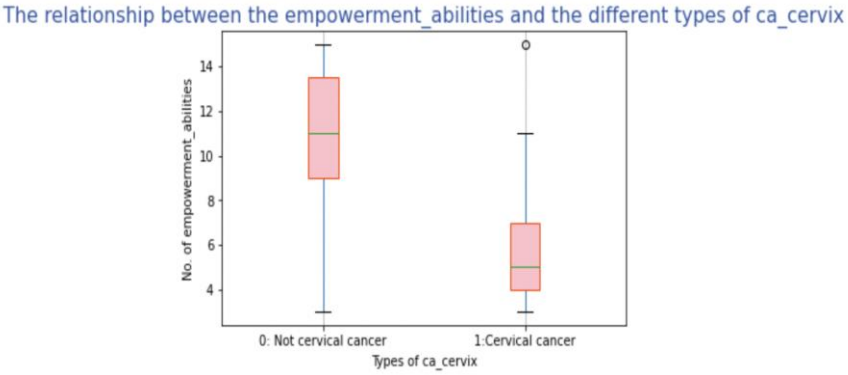


*Figure 9: The relationship between the empowement_abilities and the ca_cervix*

Based on **Figure 10**, the median, minimum, and maximum number of the empowerment_desires variable could be found in the box plot, especially the median is the most important parameter to figure the relationship in the box plot. Also, a large interval existed between the “0: Not cervical cancer” and the “1: Cervical cancer”. If the empowerment_desires larger, the risk of having cervical cancer will reduce for people. In contrast, the empowerment_desires smaller, people may contract cervical cancer.

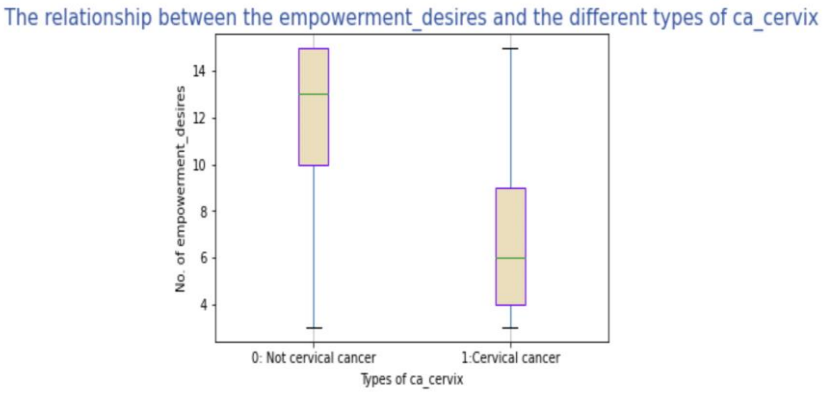


*Figure 10: The relationship between the empowerment_desires and the ca_cervix*

## 4.3 Data Modelling

In this part, we will show the performance of each model applied in the Cervical Cancer Behavior Risk data set and the comparison between their performance. Due to the small size of the data set, we randomly split the data set into training and testing sets for 10 times and apply them for training models respectively, and we calculate their average accuracy as the evaluation of the model.

### 4.3.1 k-Nearest Neighbors

In this classifier, we focus on two parameters which are k value and weight. Figure 11 shows the average classification accuracy of the k-Nearest Neighbors classifier with different k values and weights. The classifier has the highest accuracy (0.9444) when its k value is

equal to 3 and weighted by uniform, or its k value is equal to 5, 7 or 9 and weighted by distance, so we should take more evaluation in order to choose which values as the parameters of the k-Nearest Neighbors classifier.

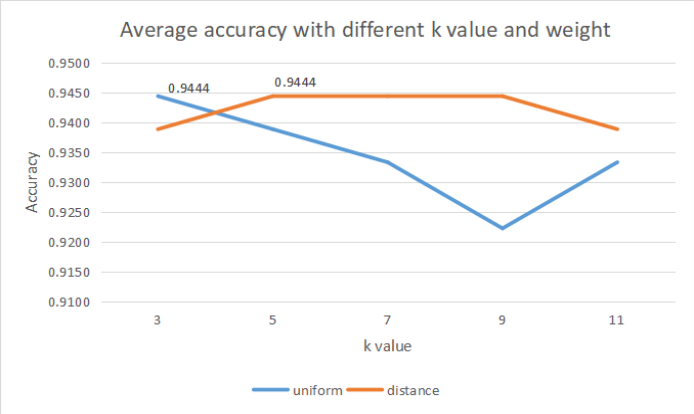


*Figure 11: Average accuracy with different k value and weights*

Figure 12,13 and 14 show the classification report of the k-Nearest Neighbors classifier with different k value and weight that make the classifier get high accuracy.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.945806 | 0.977857 | 0.959909 | 13.000000 |
| 1 | 0.940714 | 0.882381 | 0.901354 | 5.000000 |
| accuracy | 0.944444 | 0.944444 | 0.944444 | 0.944444 |
| macro avg | 0.943260 | 0.930119 | 0.930632 | 18.000000 |
| weighted avg | 0.951675 | 0.944444 | 0.943873 | 18.000000 |

*Figure 12: Classification report (k = 3, weights = 'uniform')*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.936598 | 0.981667 | 0.957896 | 13.000000 |
| 1 | 0.969048 | 0.868095 | 0.912179 | 5.000000 |
| accuracy | 0.944444 | 0.944444 | 0.944444 | 0.944444 |
| macro avg | 0.952823 | 0.924881 | 0.935037 | 18.000000 |
| weighted avg | 0.947565 | 0.944444 | 0.943175 | 18.000000 |

*Figure 13: Classification report (k = 5, weights = 'distance')*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.932088 | 0.993333 | 0.960527 | 13.000000 |
| 1 | 0.975000 | 0.843095 | 0.896898 | 5.000000 |
| accuracy | 0.944444 | 0.944444 | 0.944444 | 0.944444 |
| macro avg | 0.953544 | 0.918214 | 0.928712 | 18.000000 |
| weighted avg | 0.951387 | 0.944444 | 0.942903 | 18.000000 |

*Figure 14: Classification report (k = 7, weights = 'distance')*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.931447 | 0.991667 | 0.959642 | 13.000000 |
| 1 | 0.983333 | 0.843095 | 0.903608 | 5.000000 |
| accuracy | 0.944444 | 0.944444 | 0.944444 | 0.944444 |
| macro avg | 0.957390 | 0.917381 | 0.931625 | 18.000000 |
| weighted avg | 0.949571 | 0.944444 | 0.942744 | 18.000000 |

*Figure 15: Classification report (k = 9, weights = 'distance')*

Screening for cervical cancer is serious work. Both misdiagnosis and missed diagnosis can threaten people's lives. Therefore, we consider comprehensively and focus on f1-score. By comparing these classification reports, we find that the k-Nearest Neighbors classifier has

the highest f1-score which is about 0.9350 when its k value is equal to 5 and weighted by distance, so we choose them as the values of the classifier. The accuracy of the k-Nearest Neighbors classifier can reach 1 in this data set.

### 4.3.2 Decision Tree

Due to the small size of the data set, we can use the default value of the parameters in the Decision Tree. The accuracy of the Decision Tree may be different even if the same training data set is used to train the classifier, so we decide to calculate its accuracy of 10 times for each training and testing data set to avoid accidental results. Eventually, the average accuracy of the Decision Tree for this data set is about 0.9094, which is lower than the accuracy of k-Nearest Neighbors. The accuracy of the Decision Tree can also reach 1 in this data set. Figure 13 shows the visualization of the Decision Tree whose accuracy is equal to 1.

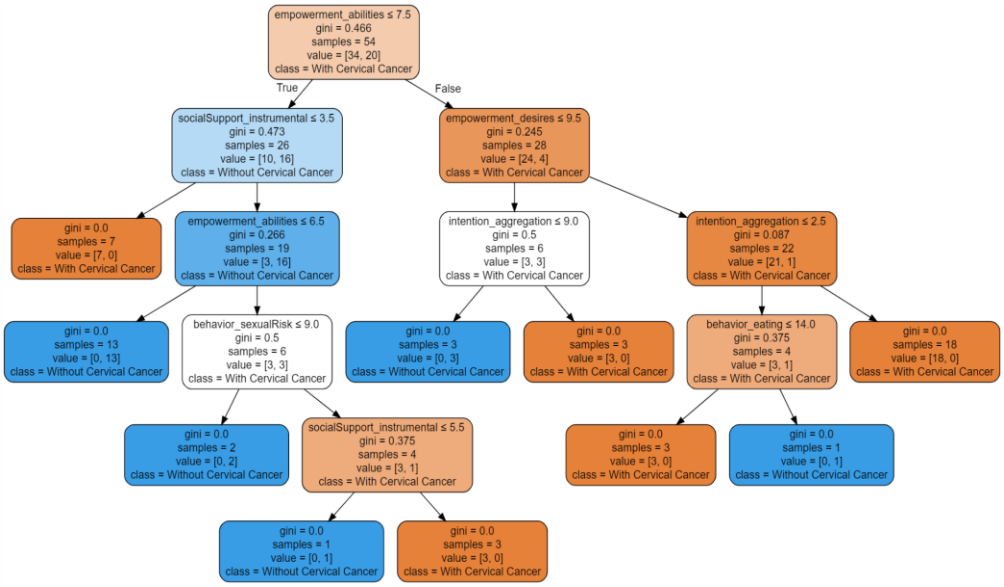


*Figure13: The visualization of the Decision Tree whose accuracy is equal to 1*

## 5. Discussion:

The result shows that the average accuracy of k-Nearest Neighbors is about 0.9444, which is higher than that of the Decision Tree (around 0.9094). However, it does not mean that the performance of k-Nearest Neighbors must be better than the Decision Tree, because both two classifiers can be trained to make their accuracy equal to 1 when using some random training and testing set. Both two classifiers can perform very well if the specific training data are splitted to train these models. This may be because the size of this dataset is only 72, which is not very stable for training models.
For the parameters of the k-Nearest Neighbors classifier, it is better that k value is equal to 5 and weighted by distance, because the classifier with these parameters has higher classification accuracy and f1-score.
The average accuracy of k-Nearest Neighbors is higher, so if you do not have much time to try different training sets to train the model, k-Nearest Neighbors could be more suitable. If you have some time to test the performance of models with different parameters, Decision Tree may be the best choice. Also, tree visualization can be used to make it easier for people to explore the results.

## 6. Conclusion:

Both the k-Nearest Neighbors and the Decision Tree classifier have the ability to perform well in this data set. If k-Nearest Neighbors classifier is chosen, it is better to make its k value equal to 5 and weighted by distance. However, the two classifiers trained by this data set may not be very stable to screen cervical cancer because the size of the data set is too small. If the size of the data set of the training model becomes larger, the two classifiers would be more powerful for screen cervical cancer.

## 7. References:

[1] Cancer council. “Cervical cancer.” Cancer council. https://www.cancer.org.au/cancer-information/types-of-cancer/cervical-cancer. (accessed May 21, 2021).

[2] Cancer council. “About cervical cancer”. Cancer council. https://www.cancerwa.asn.au/resources/specific-cancers/gynaecological-cancers/cervical-cancer/. (accessed May 21, 2021).

[3] Cancer council Victoria. “Cervical cancer”. Cancer council Victoria. https://www.cancervic.org.au/cancer-information/types-of-cancer/cervical_cancer/diagnosing_cervical_cancer.html. (accessed May 21, 2021).

[4] E.Reynolds, “What to expect during a LLETA procedure for abnormal cells.” Patient. https://patient.info/news-and-features/what-its-like-to-have-a-lletz-procedure. (accessed May 21, 2021).

[5] Sobar, R.Machmud, and A.Wijaya, “Cervical cancer behavior risk data set.” UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk. (accessed May 21, 2021).