Rémi Wong

Cervical Cancer Prediction

April 2020

Table of Contents

1 Data Analysis and Visualization

In this section, Tableau data visualization tools were used to see patterns in the cervical cancer dataset based on the research led by Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes from the Hospital University of Caracas, Venezuela (2017) after the different data filtering and pre-processing techniques have been carried out in the part A of this coursework. The data was loaded on the tableau through the data connection feature. The different variables were separated in dimensions and measures to be able to generate proper graphs.

1.1 Age distribution of Different Types of Cancers

A view has been pulled out of the dataset to see how the cases of cancer are spread over different age groups. First the different participants in the dataset have been aged grouped to have an overview of the demographics by age of the patients that will be analyzed. Considering the normal lifespan of human beings ranging up to 100 years old, it can be seen from the chart that the participants were mostly young between the ages of 10 to 45. Hence the model that would be devised in this study would be most appropriate for predicting the risks of contracting cancer for people up to the age of 50 as not enough data was recorded for older ages.

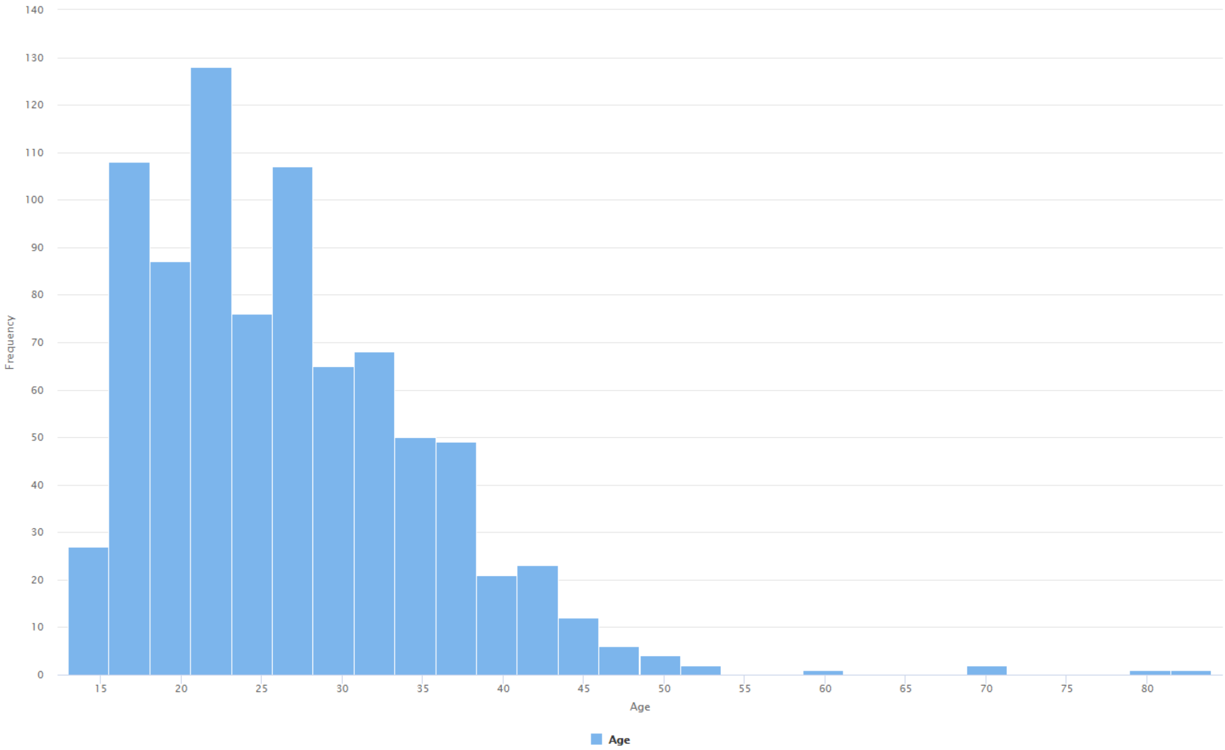


Figure 1. Age Distribution

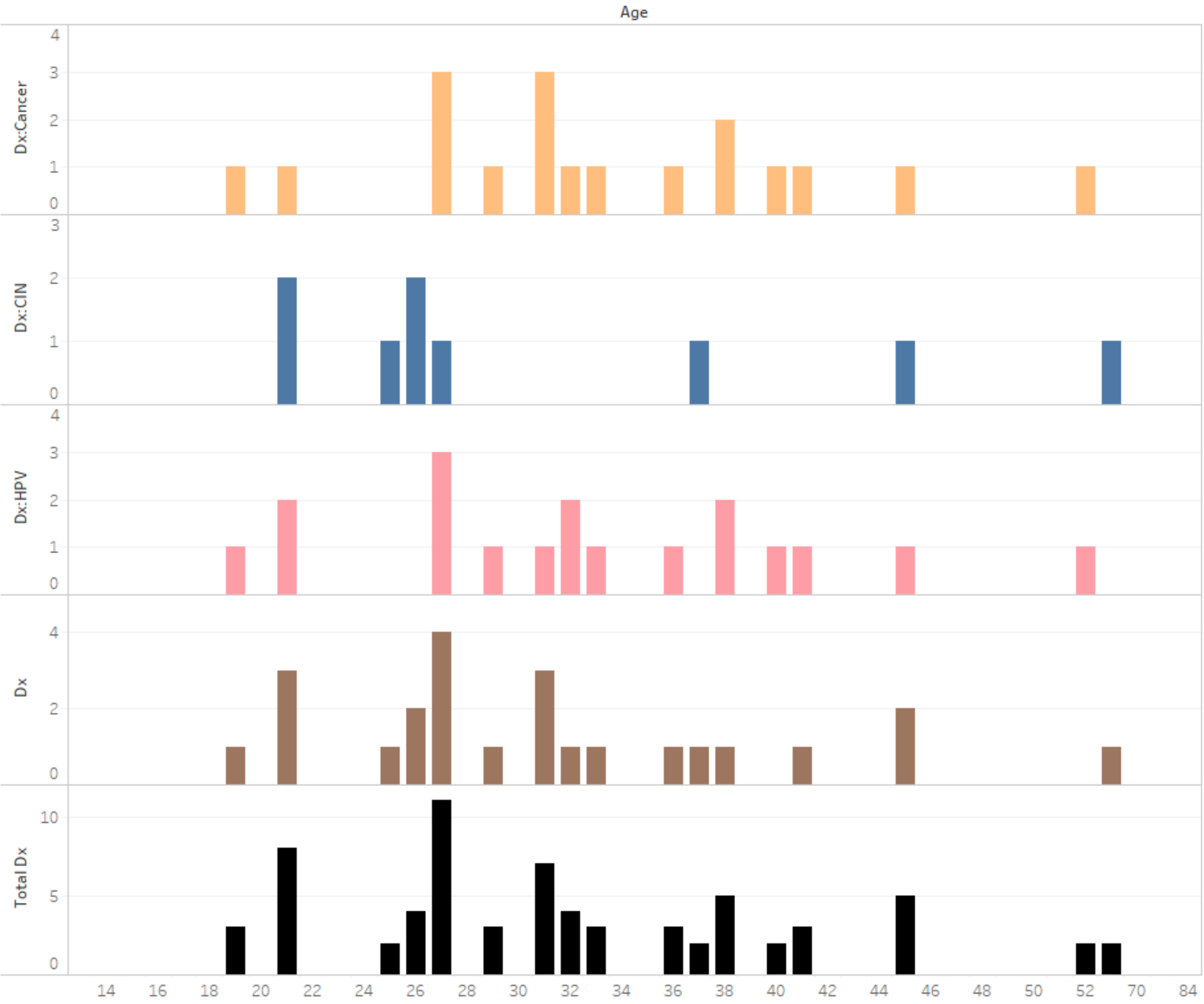


Figure 2. Cancer Age Distribution

Since most participants of this research that middle-aged women between 25 and 45, there were more cases in that age group, this might support the argument that these age groups might be more prone to suffer from cancer based on the total Dx chart or that cervical cancer might be related to age since not enough elderly people have participated in this research. This pattern tends to repeat for different types of cancer with the exception of CIN (Cervical Intraepithelial Neoplasia) which is more frequent in the younger age groups.

1.2 Relation between Sexually Transmitted Diseases and Cancers

When it comes to sexual organs, it is very important to determine the risks of cancer linked to contracting STDs. The table devised below shows the different cases where a cancer positive participant was also suffering from a STD infection. It should be noted that most of the cancerous cases were not positive with a STD with only 5 women who are HPV positive and 2 HIV positive. Hence it can be deduced that patients with HPV have a higher risk of developing some form of cancer compared to HIV. Moreover the study shows that a patient which tested positive with HPV infection does not always tend to suffer HPV cancer complications but if left untreated, there could be a higher risk though.

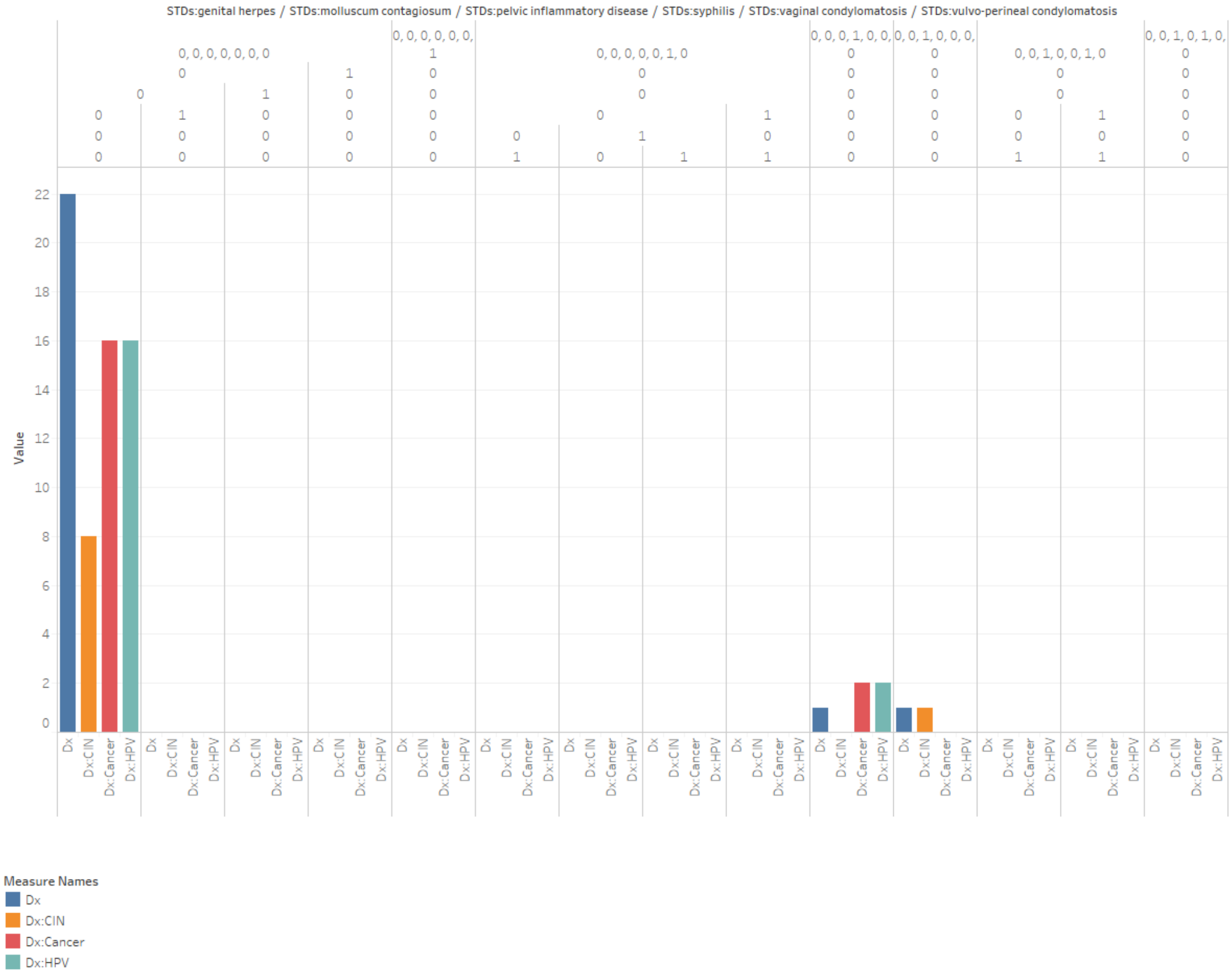


Figure 3. STD-Cervical Cancers Relation

1.3 Risks of Contracting Cancers due to Smoking

Two graphs were plotted to see the number of cases of cancer whereby the patient has been smoking and the number of years the patients have been actively smoking to see the impact of smoking on cancer risks. The first graph shows that there is a possible risk of contracting cervical cancer and HPV compared to CIN and other types of cancers Dx. However the fact that the cases spread over different amount of years and quantity of tobacco smoked per annum shows that these cancer cases might have occurred by other factors since the majority of cancer victims were non-smokers. Moreover there were not enough positive cancer cases in the dataset to reach a definite verdict but based on the data investigated, smoking did not pose a high risk of cancer.

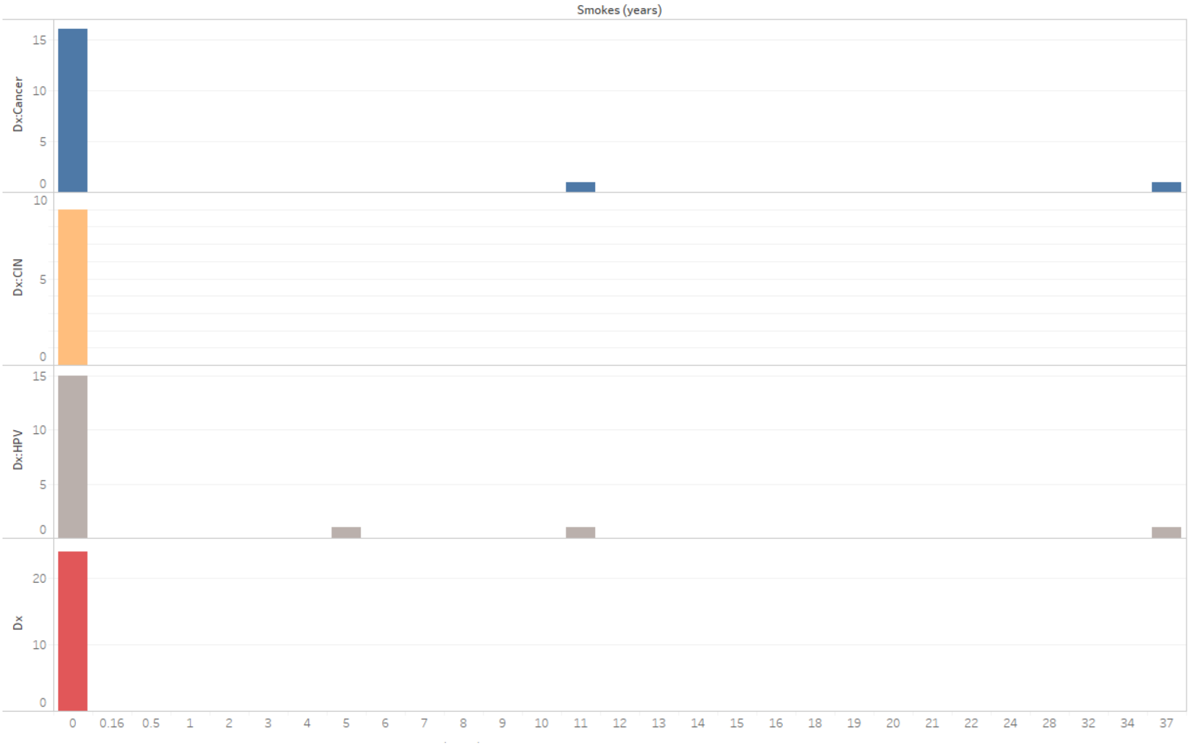


Figure 4. Cervical Cancer Due to Years Smoked

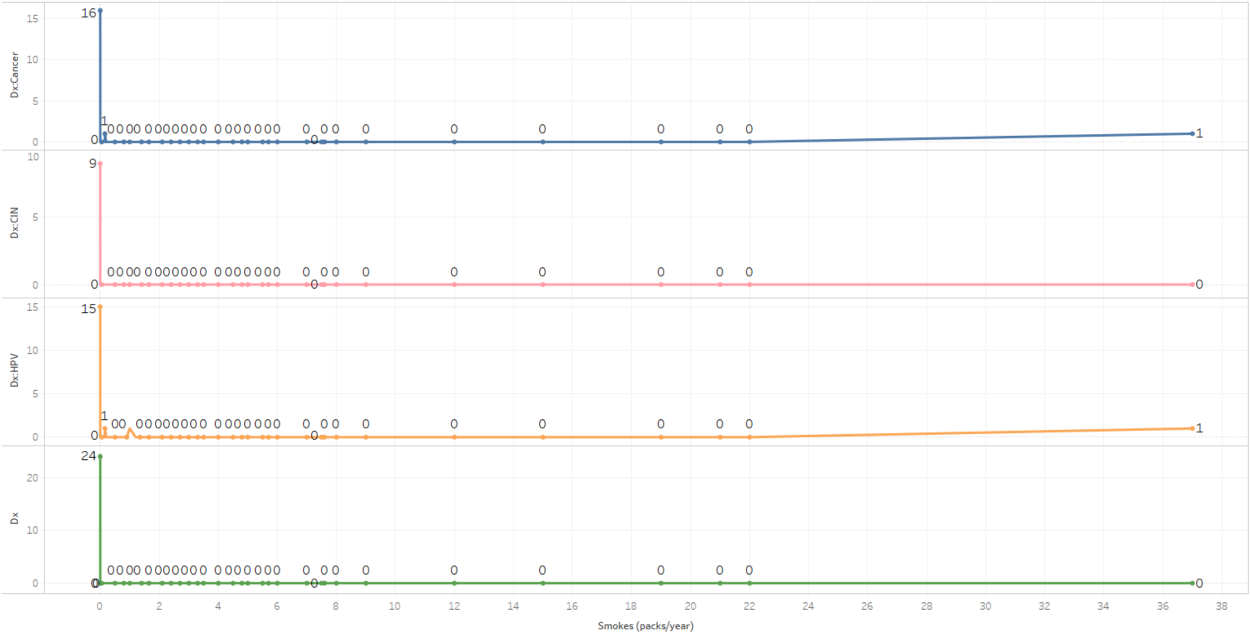


Figure 5. Cancer Impact in relation to number of packs smokes per annum

1.4 Impact of use of Hormonal Contraceptives

The graph obtained below shows substantial evidence that hormonal contraceptives usage could develop different forms of cancers. More concerning is that the data shows that different lengths of use of such products can be dangerous and more significantly for patients who have used this method of contraception for more than a year. Out of all the different dimensions seen in this research, hormonal contraceptives tends to have a greater influence of the outcome of cervical cancer tests. Hence based on the results obtained, the use of hormonal contraceptives needs to be re-evaluated and possibly banned if deemed unsafe for the health of consumers.

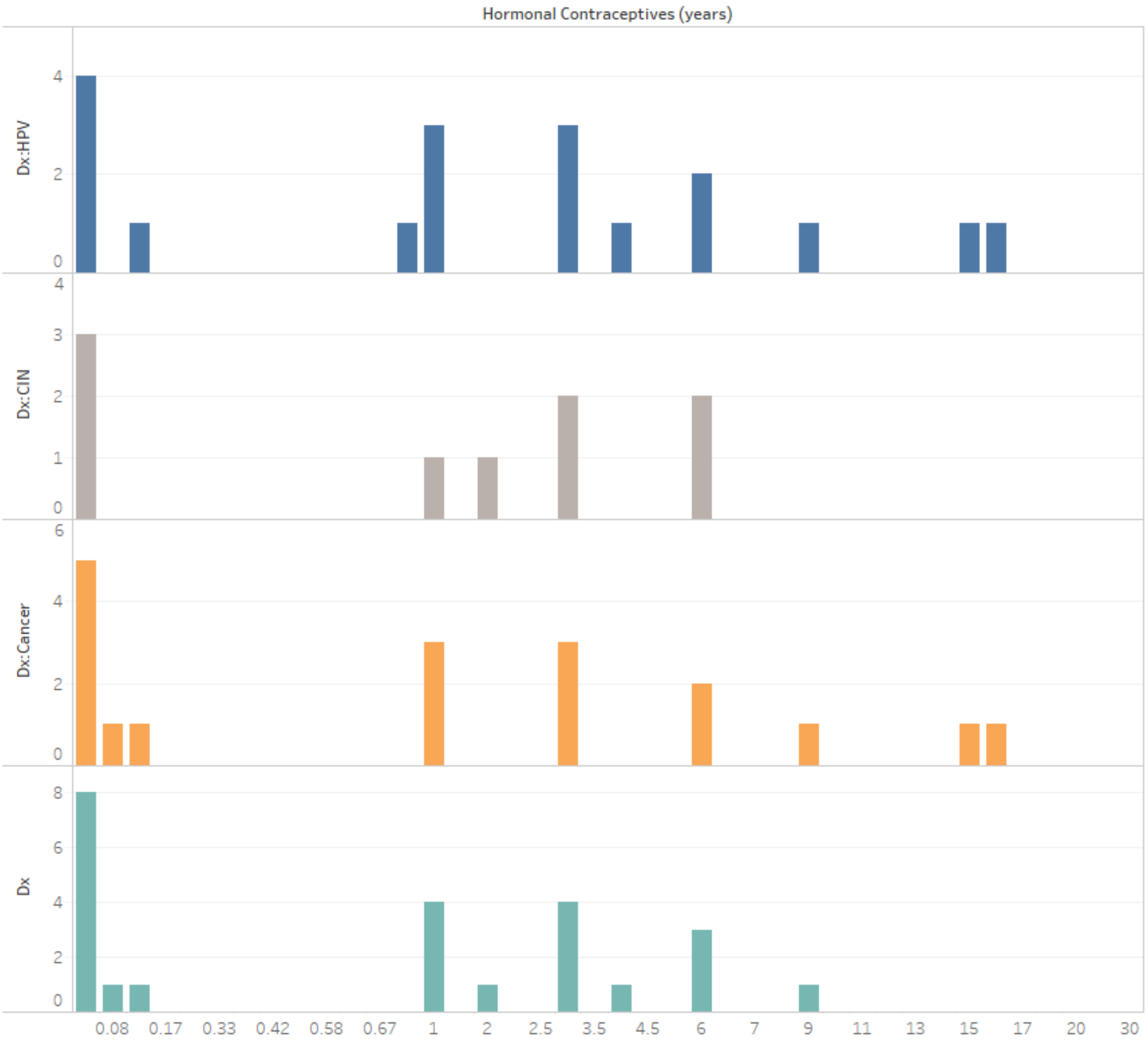


Figure 6. Impact of Hormonal Contraceptives

1.5 Impact of Intrauterine Device (IUD) on cancer

The graph of IUD against different cancers tend to suggest that there is a possible risk of developing different types of cervical cancers when used between 2 and 8 years. There were two cancer cases for 0.41 year which also shows that the conditions can appear earlier depending on the other input variables. However there were no cases for those having used the contraceptive technique for over 8 years up to 19 which shows that those who contracted cancers that the pattern to not conclusive.

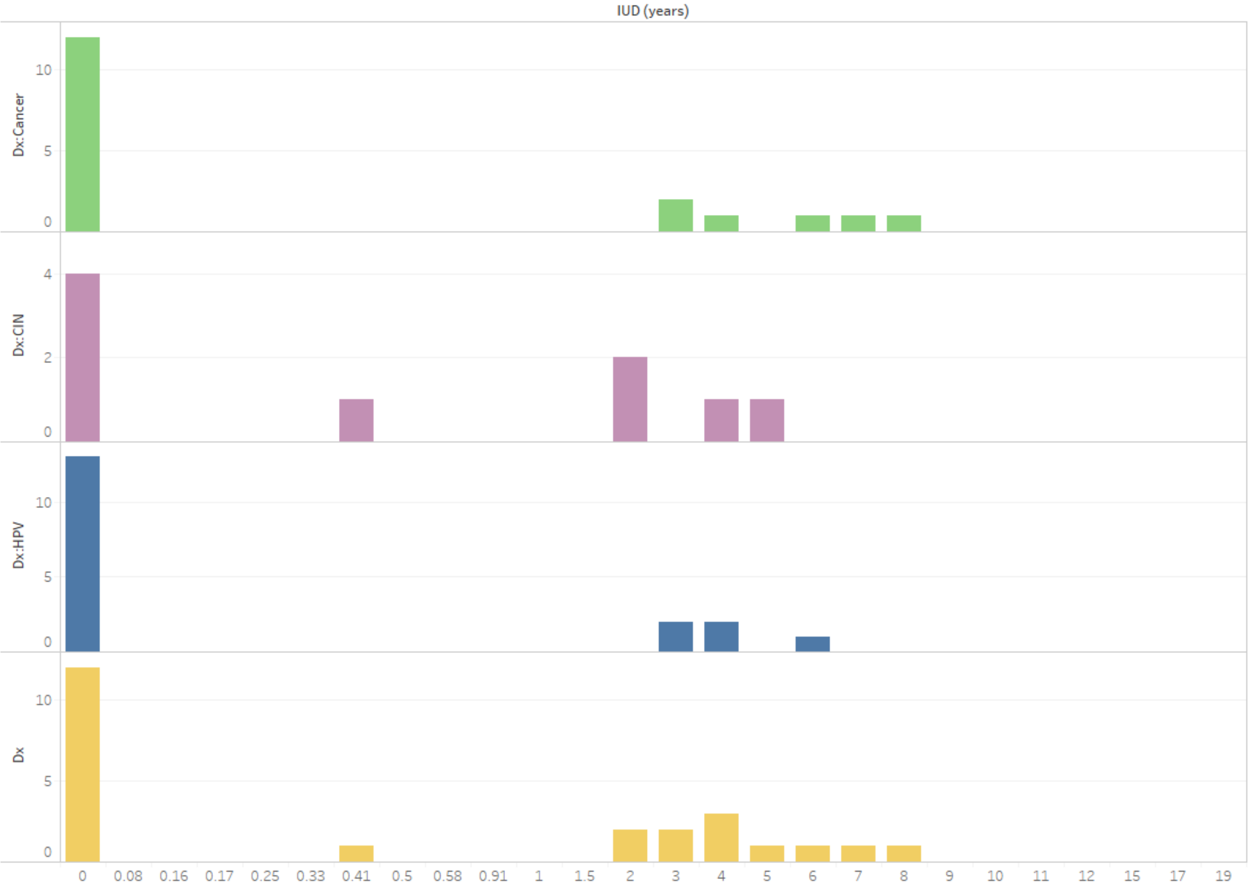


Figure 7. Impact of IUD (Intrauterine Devices)

1.6 Risk of Cancers due to Pregnancies

Out of all input variables considered to have an impact on cervical cancers, the pregnancy seemed to be the most important factor. The number of pregnancies ranges from 0 to 12 and the trend indicates that patients with a pregnancy range from 1 to 4 have a greater risk at developing cervical cancer based on the dataset as most cancer positive patients have had at least one pregnancy.

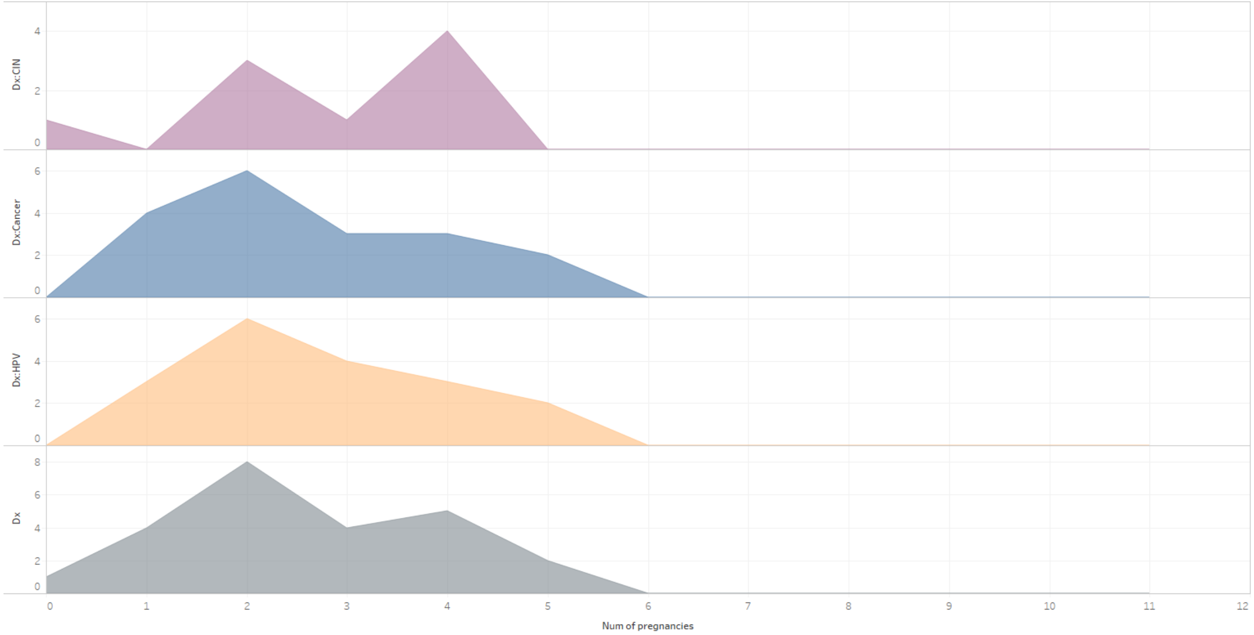


Figure 8. Pregnancies Linked to Cervical Cancers

1.7 Data Visualization Critical Analysis

From the different views obtained from the dataset using Tableau visualization tool, the findings have been crucial to determine the possible data mining model that could be used to predict the contraction of cancer and be able to investigate the possible causes of cervical cancers. Out of all positive cervical cancer cases, there were three input variables which were most significantly high to all patients as seen in the graphs generated, namely; Hormonal Contraceptives, IUD (Intrauterine Devices) and Pregnancy. However there were cases affiliated to the other inputs but these inputs had relatively a lower number of positive cases. Amongst the different inputs, it can be deduced that age is not a clear indicator when it comes to the probability of contracting cervical cancer since the majority of people who participated in the research were aged between 10 and 45, and furthermore most of the positive cases were recorded within that same age group.

The patterns obtained indicate that each and every input variable of the dataset might have a degree of influence on the value of the different cervical cancer class variables since they all have at least one case of cancer related to them. Some of the inputs have a higher influence whereas all others have a lower influence. This shows that a data mining model that can learn from each input variable might be best suited for the current scenario due to varying levels of influence.

2 Data Mining

Based on the requirements of the coursework, the use of a data mining algorithm was used to be able to use the current dataset to predict of possible outcomes of new data. Since the cervical cancer dataset was a historical dataset whereby the data was collected from real-life individuals, it did not come with a test dataset. In order to solve this issue, the chosen dataset has been divided into a training set and test set to be able to perform this study. As such the dataset which was composed of 858 rows was split into an 838-row training set and a 20-row test set to be able to first train a data mining model and then be able to test the model to determine the accuracy of the model for predicting cervical cancer cases.

From the patterns obtained from tableau it can be deduced that there are several factors which affect the chances of contracting cancers and because of a high amount of uncertainty, the neural network algorithm through the use of a multi-layer perceptron concept was seen as the most suitable data mining algorithm due to its ability to work with different types of data which do not show any relation to one another. For example the number of pregnancies and smoking are very different in nature when it comes to predictions but with the help neural network, some sort of relationship based on the learning of patterns could give insights on test data.

2.1 Neural Network Training Model

Before being able to predict a possible occurrence of different cervical cancers, one very important task was to build a training model for the study's specific use-case scenario. A neural network algorithm relies on the training of all the network weighs through a process known as the forward-propagation and the backward-propagation. Forward propagation deals with the calculation of the output of a neural network by feeding the inputs in the network and passing them in a forward direction to produce an output after going through different hidden layers and activation function. The hidden layers were used to increase the accuracy of the algorithm and to improve its ability to tackle problems at a higher degree of accuracy. A sigmoid activation function was deemed to be sufficient for this study's neural network since it corresponds to the values class variables used in the current dataset, which was either 0 or 1 since sigmoid function gives an output between 0 and 1.

The figure below shows the neural network devised for the data mining process of the dataset with 21 input variables, layer-1 having 8 nodes, and last the layer-2 having 10 nodes. In each hidden layer, there is an additional input known as the bias input which is added to stabilize the weights throughout the network. The output will act as the class variable, in this case a type of cancer. The weights in the network are initially randomized.

In the training process, the output obtained from the different inputs is compared against the expected value from the training set class variable to obtain an error margin. This error margin is then used to calibrate the different weights through the network through a process known as backward propagation till the proper desired output is obtained.

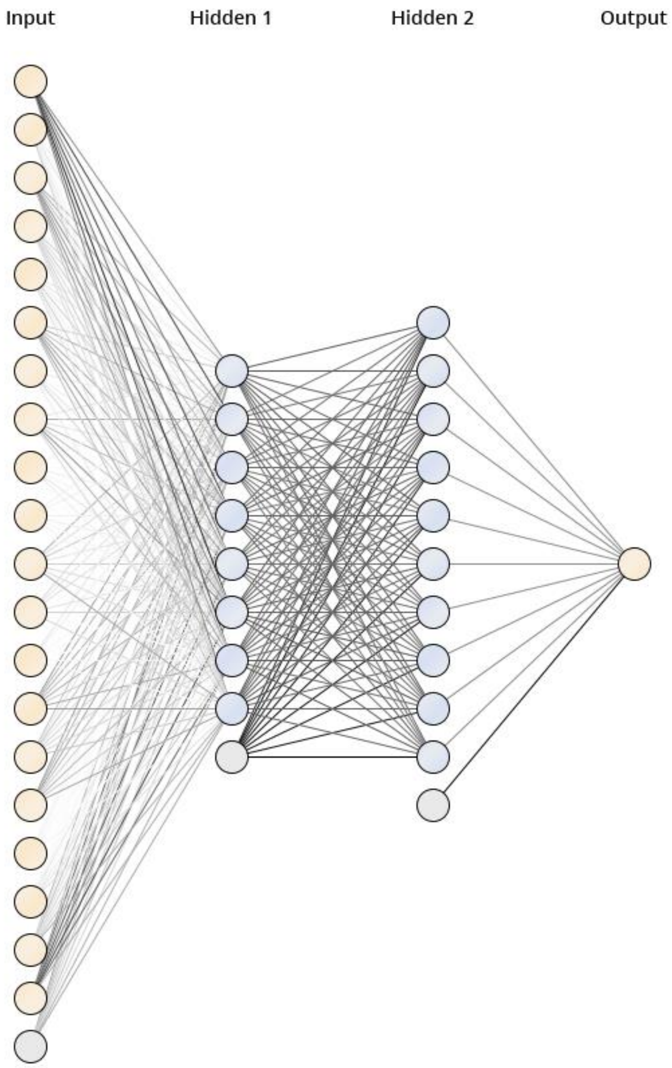


Figure 9. Neural Network Topology

This process has been repeated for each row of the training set (up to 838 times) and looped for 500 times (also known as the epoch). The different weights were adjusted with a learning rate of 0.005 and a momentum of 0.9. The whole process has been applied in the Rapidminer studio data mining software through the use of the operators below:

- **Retriever**: To load the dataset
- **Select Attribute**: To select the required data inputs and desired class variable
- **Neural Net**: The neural network algorithm function

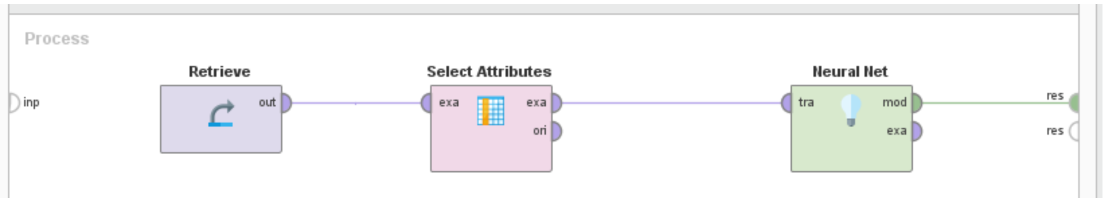


Figure 10. Initial Rapidminer Neural Network Design Process

After several trials, the topology devised above produced the best optimum results because using less or more hidden layers and nodes tended to cause under-fitting and over-fitting issues whereby the neural network would become unstable. After training the algorithm, it was converted into a training model through the use of a **cross-validation operator**. The cross validator model tool incorporates a training side and a testing side for applying the trained algorithm to a test dataset with the help of a **regression performance** gauging operator. The regression performance indicator was used since it is suitable for prediction scenarios which is what this study is dealing with. The updated process flow of Rapidminer is shown in the figures below with the root mean square error obtained after the training of the model through the **Apply Model operator**.

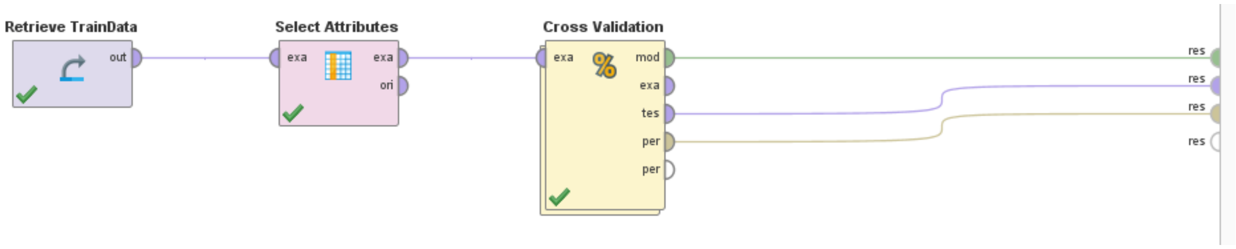


Figure 11. Training Model Cross Validation Design

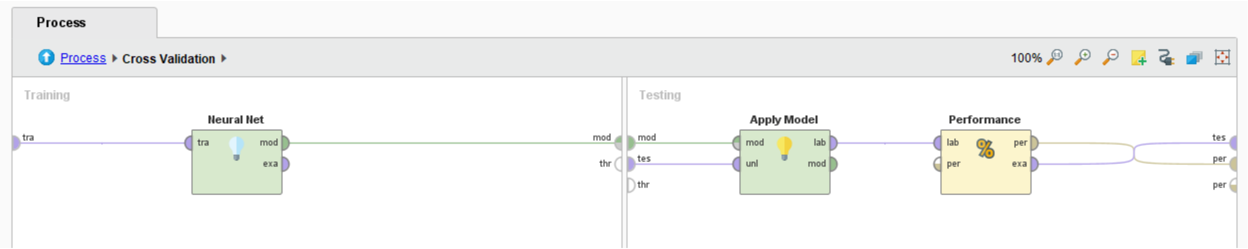
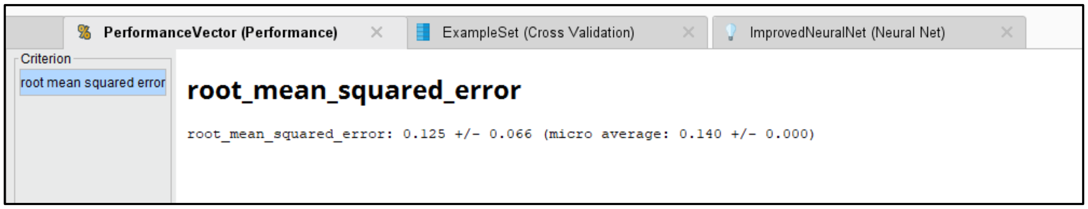


Figure 12. Training Model Performance Calibration


PerformanceVector (Performance)
ExampleSet (Cross Validation)
ImprovedNeuralNet (Neural Net)
Criterion
root mean squared error

root_mean_squared_error

```
root_mean_squared_error: 0.125 +/- 0.066 (micro average: 0.140 +/- 0.000)
```

Figure 13. Training Model Performance Metric

2.2 Test Model

The model designed in section 2.1 was used to predict the classification of different types of cervical cancers by applying the trained neural network to the unclassified test dataset. The Rapidminer process was appended to retrieve the test data inputs and class variable and apply the neural network algorithm model to the test dataset to predict expected class variable output value.

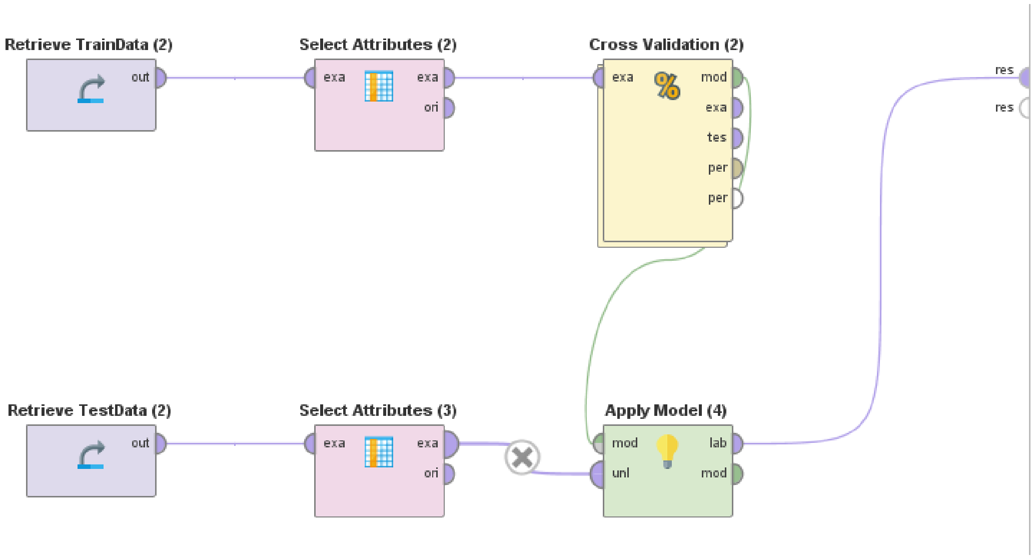
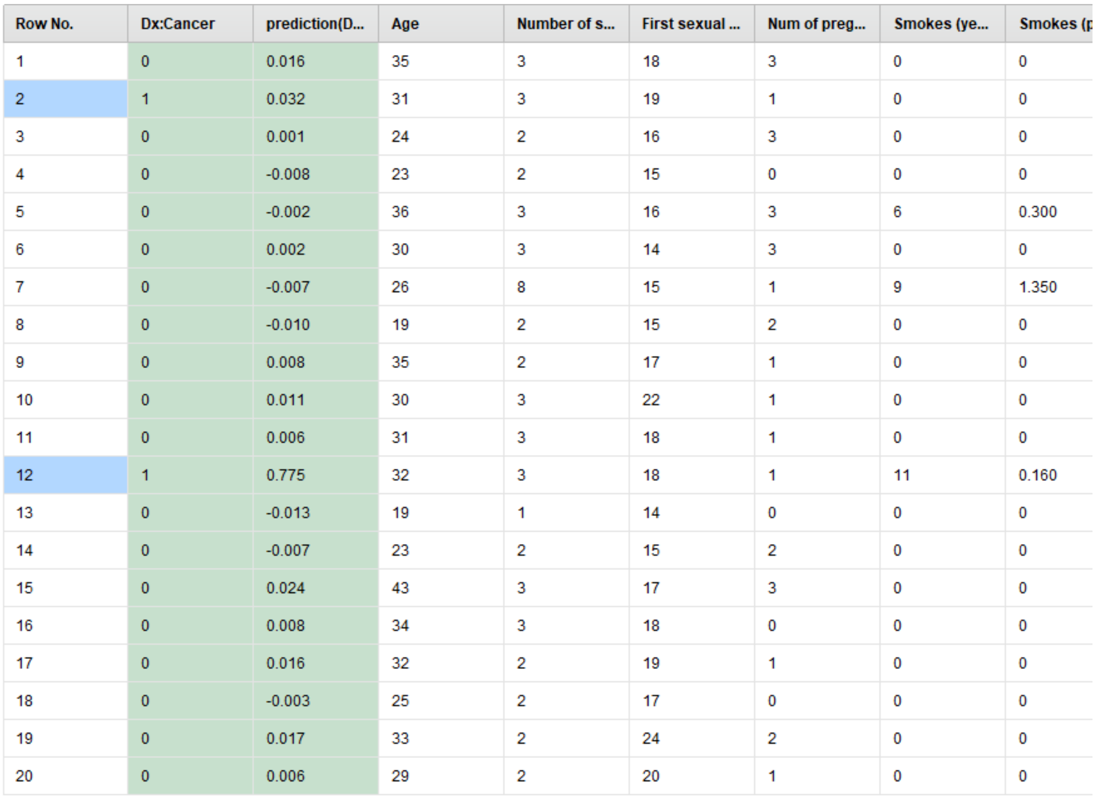


Figure 14. Rapidminer Test Model Design Process

2.2.1 Dx: Cancer Prediction

The figure below shows the model’s Dx Cancer results and comparing the predicted value against the expected value, we can see that the algorithm is working very well. Only 2 positive cancer cases were expected and assuming that for the cancer result of the neural network to be positive, the prediction value needs to be greater than 0.5 (prediction > 0.5).



Row No.	Dx:Cancer	prediction(D...	Age	Number of s...	First sexual ...	Num of preg...	Smokes (ye...	Smokes (p
1	0	0.016	35	3	18	3	0	0
2	1	0.032	31	3	19	1	0	0
3	0	0.001	24	2	16	3	0	0
4	0	-0.008	23	2	15	0	0	0
5	0	-0.002	36	3	16	3	6	0.300
6	0	0.002	30	3	14	3	0	0
7	0	-0.007	26	8	15	1	9	1.350
8	0	-0.010	19	2	15	2	0	0
9	0	0.008	35	2	17	1	0	0
10	0	0.011	30	3	22	1	0	0
11	0	0.006	31	3	18	1	0	0
12	1	0.775	32	3	18	1	11	0.160
13	0	-0.013	19	1	14	0	0	0
14	0	-0.007	23	2	15	2	0	0
15	0	0.024	43	3	17	3	0	0
16	0	0.008	34	3	18	0	0	0
17	0	0.016	32	2	19	1	0	0
18	0	-0.003	25	2	17	0	0	0
19	0	0.017	33	2	24	2	0	0
20	0	0.006	29	2	20	1	0	0

Figure 15. Dx Cancer Testset Perdiction

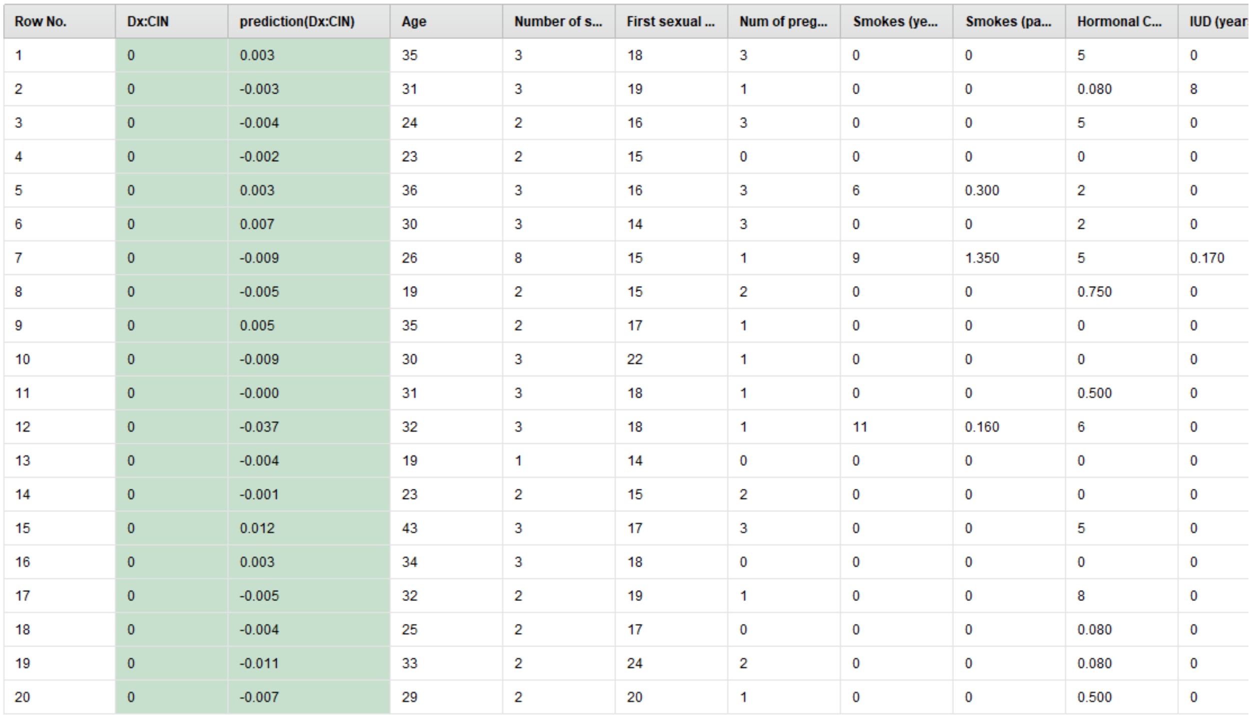
From the results we can deduce that the algorithm has predicted that there would be 1 positive case which was the row-12 with value 0.775.

Correct Prediction Score: 19

Accuracy: 19/20 = 95 %

2.2.2 Dx: CIN (Cervical Intraepithelial Neoplasia) Prediction

The model was again applied for the Dx CIN class variable and this time the prediction was correct 20 out of 20, hence with an accuracy of 100 %



Row No.	Dx:CIN	prediction(Dx:CIN)	Age	Number of s...	First sexual ...	Num of preg...	Smokes (ye...	Smokes (pa...	Hormonal C...	IUD (year
1	0	0.003	35	3	18	3	0	0	5	0
2	0	-0.003	31	3	19	1	0	0	0.080	8
3	0	-0.004	24	2	16	3	0	0	5	0
4	0	-0.002	23	2	15	0	0	0	0	0
5	0	0.003	36	3	16	3	6	0.300	2	0
6	0	0.007	30	3	14	3	0	0	2	0
7	0	-0.009	26	8	15	1	9	1.350	5	0.170
8	0	-0.005	19	2	15	2	0	0	0.750	0
9	0	0.005	35	2	17	1	0	0	0	0
10	0	-0.009	30	3	22	1	0	0	0	0
11	0	-0.000	31	3	18	1	0	0	0.500	0
12	0	-0.037	32	3	18	1	11	0.160	6	0
13	0	-0.004	19	1	14	0	0	0	0	0
14	0	-0.001	23	2	15	2	0	0	0	0
15	0	0.012	43	3	17	3	0	0	5	0
16	0	0.003	34	3	18	0	0	0	0	0
17	0	-0.005	32	2	19	1	0	0	8	0
18	0	-0.004	25	2	17	0	0	0	0.080	0
19	0	-0.011	33	2	24	2	0	0	0.080	0
20	0	-0.007	29	2	20	1	0	0	0.500	0

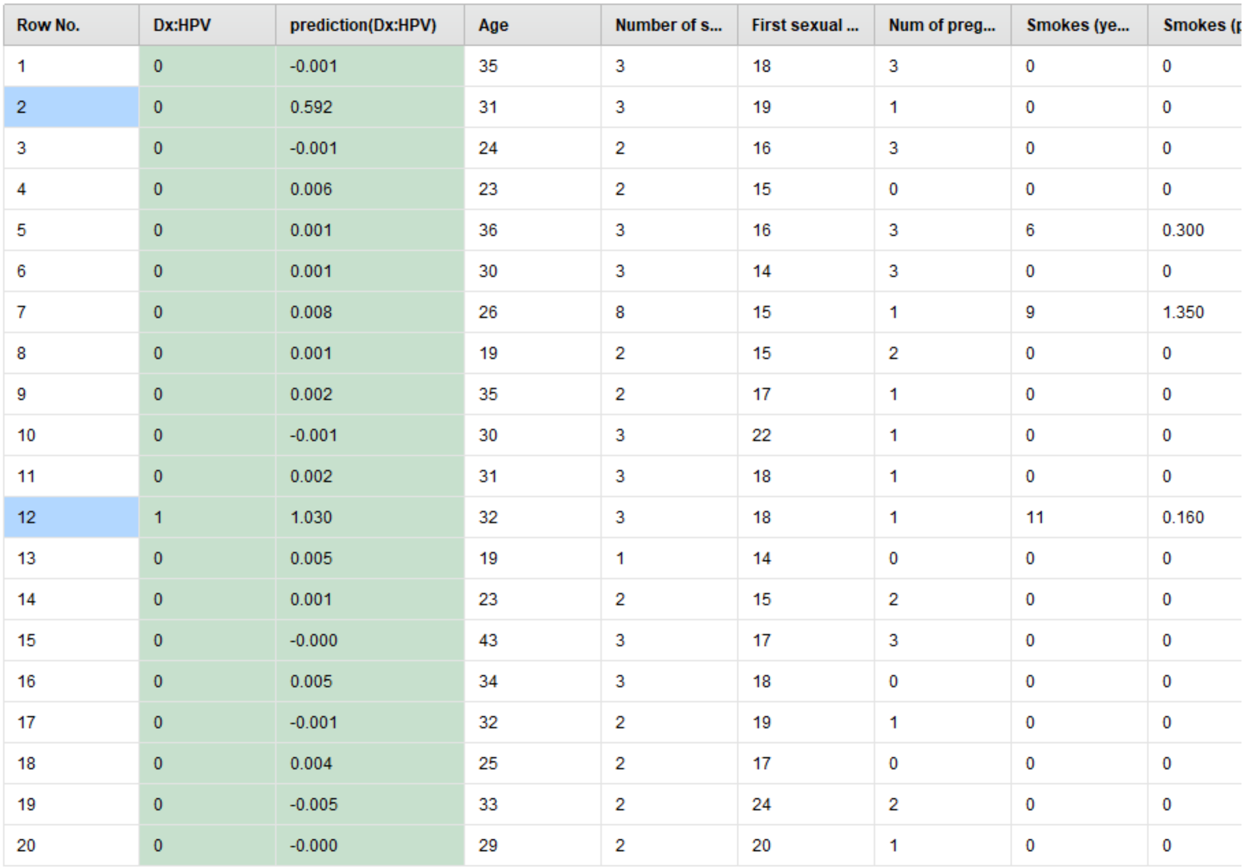
Figure 16. Dx: CIN Testset Prediction

2.2.3 Dx: HPV (Human Papillomavirus) Prediction

For Dx HPV, the algorithm predicted two positives from the inputs instead of one which was expected at row-12. Hence,

Correct Prediction Score: 19

Accuracy: 19/20 = 95 %



Row No.	Dx:HPV	prediction(Dx:HPV)	Age	Number of s...	First sexual ...	Num of preg...	Smokes (ye...	Smokes (p
1	0	-0.001	35	3	18	3	0	0
2	0	0.592	31	3	19	1	0	0
3	0	-0.001	24	2	16	3	0	0
4	0	0.006	23	2	15	0	0	0
5	0	0.001	36	3	16	3	6	0.300
6	0	0.001	30	3	14	3	0	0
7	0	0.008	26	8	15	1	9	1.350
8	0	0.001	19	2	15	2	0	0
9	0	0.002	35	2	17	1	0	0
10	0	-0.001	30	3	22	1	0	0
11	0	0.002	31	3	18	1	0	0
12	1	1.030	32	3	18	1	11	0.160
13	0	0.005	19	1	14	0	0	0
14	0	0.001	23	2	15	2	0	0
15	0	-0.000	43	3	17	3	0	0
16	0	0.005	34	3	18	0	0	0
17	0	-0.001	32	2	19	1	0	0
18	0	0.004	25	2	17	0	0	0
19	0	-0.005	33	2	24	2	0	0
20	0	-0.000	29	2	20	1	0	0

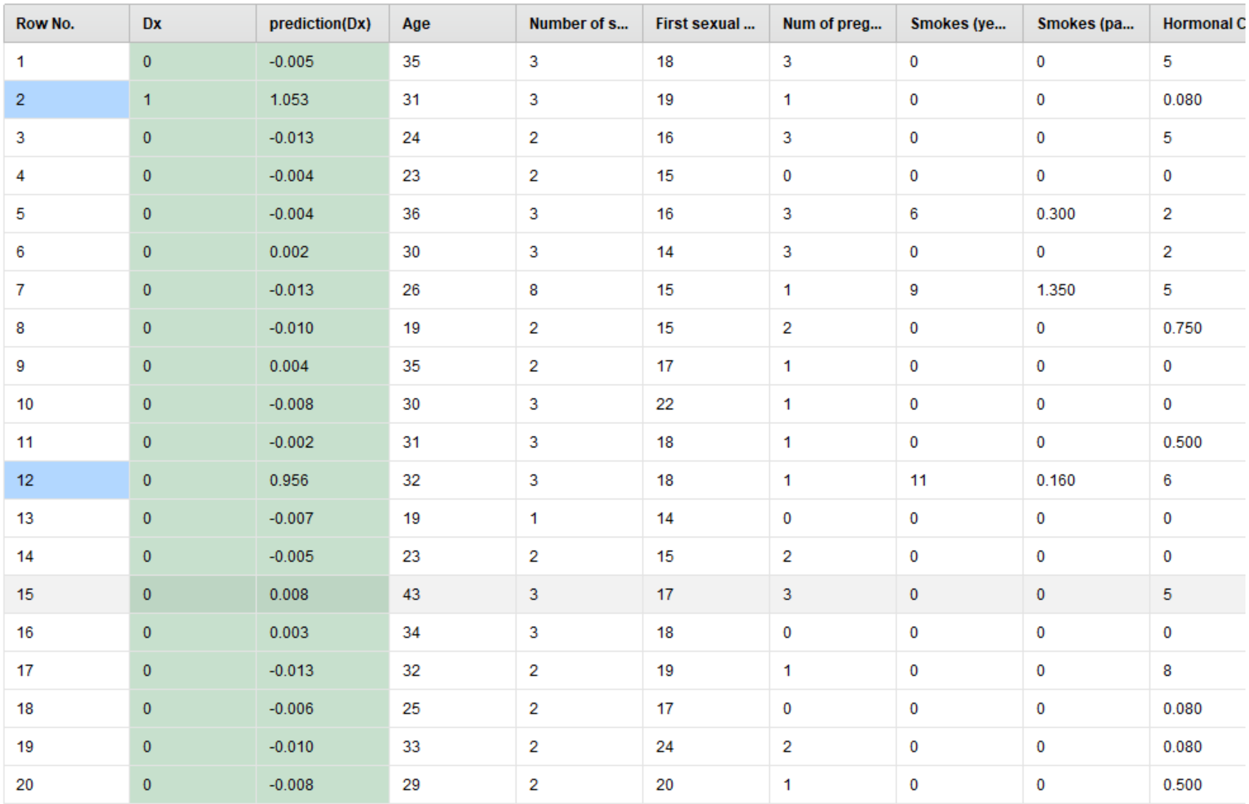
Figure 17. Dx: HPV Testset Prediction

2.2.4 Dx: Others Predictions

For the last type of cervical cancers which is Dx (others), the algorithm predicted one case successfully (row-2) but detected a false-positive case (row-12).

Correct Prediction Score: 19

Accuracy: 19/20 = 95 %



Row No.	Dx	prediction(Dx)	Age	Number of s...	First sexual ...	Num of preg...	Smokes (ye...	Smokes (pa...	Hormonal C
1	0	-0.005	35	3	18	3	0	0	5
2	1	1.053	31	3	19	1	0	0	0.080
3	0	-0.013	24	2	16	3	0	0	5
4	0	-0.004	23	2	15	0	0	0	0
5	0	-0.004	36	3	16	3	6	0.300	2
6	0	0.002	30	3	14	3	0	0	2
7	0	-0.013	26	8	15	1	9	1.350	5
8	0	-0.010	19	2	15	2	0	0	0.750
9	0	0.004	35	2	17	1	0	0	0
10	0	-0.008	30	3	22	1	0	0	0
11	0	-0.002	31	3	18	1	0	0	0.500
12	0	0.956	32	3	18	1	11	0.160	6
13	0	-0.007	19	1	14	0	0	0	0
14	0	-0.005	23	2	15	2	0	0	0
15	0	0.008	43	3	17	3	0	0	5
16	0	0.003	34	3	18	0	0	0	0
17	0	-0.013	32	2	19	1	0	0	8
18	0	-0.006	25	2	17	0	0	0	0.080
19	0	-0.010	33	2	24	2	0	0	0.080
20	0	-0.008	29	2	20	1	0	0	0.500

Figure 18. Dx: Others Testset Prediction

2.3 N-Fold Evaluation Test

The N-Fold Test method was used to validate the results obtained from the prediction model designed. A 4-Fold Test was carried on the training and test datasets in because it was found to be suitable with the number of records used in this study.

As such the whole original dataset was used to carry out this experiment such that the 858 records were split into 4 folds of equal size, i.e. 858/4 = 214 (remainder 2).

Fold Combinations		Cervical Cancers Accuracy							
Training Folds	Test Fold	Dx: Cancer		Dx: CIN		Dx: HPV		Dx: Others	
		Correct	Acc	Correct	Acc	Correct	Acc	Correct	Acc
1 + 2 + 3	4	205	95.8%	208	97.2%	213	99.5%	208	97.2%
1 + 2 + 4	3	211	98.6%	214	100%	213	99.5%	214	100%
1 + 3 + 4	2	213	99.5%	213	99.5%	213	99.5%	213	99.5%
2 + 3 + 4	1	203	94.9%	212	99.1%	213	99.5%	208	97.2%

Table 1. N-Fold Evaluation Results.

From the 5-Fold test evaluation conducted above, the average accuracy of the neural network algorithm was calculated from the mean of the 16 benchmark values obtained.

Mean Accuracy = 1576.5% / 16 = 98.5 %

The 4-Fold test confirms that the model designed is able to predict the expected class variables of cervical cancer cases with an average accuracy of 98.5 % which is a very good score. Even when different combinations of folds were processed, the neural network algorithm model was successfully able to determine the correct expected value. Hence it is safe to say that the model has passed the N-Fold test with the high accuracy values obtained in table 1.

2.4 Precision and Recall Evaluation

A precision and recall method was also used to evaluate the neural network algorithm based on the results obtained in section 2.2. This method utilizes the confusion matrix to determine the parameters for calculating the precision and recall. Since there are four different class variables, the values were summed when calculating the true positives, true negatives, false positives, and false negatives to calculate Precision, Recall values, specificity, and sensitivity.

	Dx Cancer		Dx CIN		Dx HPV		Dx: others	
	Class Positive	Class Negative	Class Positive	Class Negative	Class Positive	Class Negative	Class Positive	Class Negative
Actual Positive	1	1	0	0	1	0	1	0
Actual Negative	0	18	0	20	1	18	1	18

Table 2. Precision and Recall Evaluation Results

Using $Precision, p = \frac{TP}{\text{TP+FP}} = \frac{1+1+1}{(1+1+1)+(1+1)} = \frac{3}{5} = 0.60$

And using $Recall, r = \frac{TP}{\text{TP+ FN}} = \frac{1+1+1}{(1+1+1)+(1)} = \frac{3}{4} = 0.75$

Then the harmonic mean of the results obtained have been calculated using

$$F_1 = \frac{2pr}{\text{p + r}} = \frac{2\ \times 0.60\ \times 0.75}{0.60 + 0.75} = \frac{0.9}{1.35} = 0.67$$

The $True\ Positive\ Rate = \frac{TP}{\text{TP+FN}} = \frac{1+1+1}{(1+1+1)+(1)} = \frac{3}{4} = 0.75$

The $False\ Positive\ Rate = \frac{FP}{\text{TN+FP}} = \frac{1+1}{(18+20+18+18)+(1+1)} = \frac{2}{76} = 0.0263$

And the Specificity, $True\ Negative\ Rate = \frac{TN}{\text{TN+FP}} = \frac{18+20+18+18}{(18+20+18+18)+(1+1)} = \frac{74}{76} = 0.97$

And finally the $False\ Positive\ Rate, FPR = 1 - \text{Specificity} = 1 - 0.97 = 0.03$

Different performance metrics have been calculated based on the results obtained in table 2. The precision of the model had the worst rating and this shows that the neural network algorithm was not trained properly and this could be due to the lack of positive cervical cancer cases in the training set. The majority of expected class variables were negative and this is also reflected in the high value of True Negative Rate and low False Positive Rate. However based on the current dataset, a precision of 0.60 was a decent score and a recall value of 0.75 confirms this statement. The model generated a True Positive Rate of 0.75 and False Positive Rate of 0.0263 which is very good considering the clear lack of training data for the positive cases. The results show that the model has the potential to be even more precise and give better predictions if a higher quality and more abundant training data tuples are used to train the model.

3 Conclusion

3.1 Overall Results

In the data visualization process, the different patterns of the different cases of cervical cancers have been investigated to see how the different input variables historically affect the outcome of the class variables. Based on the graphs generated, it could be seen that patients from different age groups were equally vulnerable to contracting cancer. More surprisingly, there were more cases for middle aged women. The study shows that all the input variables have an impact on cancer risks due to a considerable positives cases related to each of them. But the trends show that most cancer cases were related to patients with a considerable amount of Hormonal Contraceptives use and a number of Pregnancies between 1 and 4. The visualization of the dataset proved to be a challenging task since the different tuples in the dataset were not unique because most of the variables were binary values which offers very little room in terms of modelling. If more patient data is collected for the research, more patterns could have been generated and with a higher level of complexity. The majority of tuples had negative class variables as the number of positive cancer cases in the dataset were low as compared to the number of tuples.

The research indicates that in order to properly predict the contracting of cervical cancers, a model where all the different inputs could impact on the prediction results was best suited to this particular scenario since each of the inputs contribute to the expected value of class variable to a certain degree. A decision-based algorithm was not suitable since the contraction of cancer relies on the value of all inputs equally and the nature of this system to too probabilistic. Due to the influence of multiple inputs in this dataset, the neural network algorithm was found to be the best data mining algorithm to be most appropriate for creating a model which would be able to predict the occurrence of cervical cancers. With high accuracy results, it is safe to say that the model devised was successfully able to classify a subset of the original dataset. The model has also been evaluated using N-Fold test and performance metrics to see how well it can generate predictions. A minimum rate of 60% for True Positive Rate (TPR) obtained shows that the model devised was able to predict both positive and negative cancer cases reasonably well but the model produced much better results for negative cases due to the low cases of positive cases in the dataset for it to be trained properly.

3.2 Data Mining Model Results

The model devised for the cervical cancers dataset was found to be ideal since it was able to properly classify the inputs with a minimum accuracy score of 95.0% on 20 tuples test data and 838 training tuples dataset. A mean average of 98.5% of the class variables from the test dataset in the N-Fold test carried out which was impressive since a larger ratio of testing data was used against training data. Given that the model was trained with only 642 tuples of patient cases, the model was shows that it has the potential to generate much better results if a larger pool of training tuples is utilized because neural networks work better when it can learn with more training samples. Considering the model had to be able to classify four different types of cervical cancers, the results show that the model can be used in datasets where various class variables.

The second evaluation carried out based on precision and recall shows that despite that the high accuracy scores obtained, the model's ability in predicting positive cancer results was precise at only 60% and its recall ability was only 75% due because the model was not able to detect all of the expected positive cancer output as intended. This issue was caused because the dataset had very few occurrences of positive cases against negative cancer cases, thus the algorithm was not trained enough for the proper detection of positive cases. This has also been reflected in the score for true positive rate of 0.75, which is not bad but could be improved. However the low value of False Positive Rate (FPR) of 3% and True Negative Rate

(TNR) of 97% shows that the model is much more efficient in classifying negatives. The results were due to the large presence of negative cancer cases in the whole dataset. In order to improve the models behavior against expected positive cases, the further data needs to be collected from more patients which are positively detected with cancer until an equilibrium between positive and negative cases is reached so that the neural network is better able to handle true positive cases as it needs to learn more on the detection of positive cases. This argument is also supported by the low False Positive Rate of 2.6%.

In order to adapt to a more balanced training dataset, the neural network devised might have to be adjusted to cater for under-fitting and over-fitting based on the results outputted from the network. Hence, the model designed has been able to learn from the cervical dataset and produced very good predictions based on the expected class variable values and association rules expected but it could be improved through the use of a more diverse and balance training dataset.

3.3 Business Intelligence based on the Model

Throughout this study, the data mining explored can now enable stakeholders involved in the medical industry to make use of this tool to improve their quality of service by reducing the time it takes to carry out diagnostics from future patients since it is known that some cancer tests can take a long time to process. In order to assess how Business Intelligence can help to improve the healthcare industry through the creation of a model for predicting cervical cancers, the Metter's Framework for BI (Business Intelligence (2009) was used to evaluate how the model developed. At the same time it can help physicians to make a preliminary assessment of patients by applying the model to a patient's data. The model could help countries which cannot afford to invest massively in high quality healthcare to carry out more tests. It can even be used to provide a cheaper alternative to medical laboratory tests. The use of business intelligence in the detection of cancer also helps to ease the workload stressed upon medical practitioners and medical laboratories since the model is able to diagnose patients with a high level of accuracy.

The model can also be used in the media to advise the public on which input variables (smoking, sexual habits, and sexually transmittable diseases …) tend to pose a higher risk in developing cervical cancers. As seen from the visualization of historical data, the use of hormonal contraceptives and IUD birth control have a huge impact on cervical cancers, hence this model could push pharmaceutical companies to improve the quality of their product and devise new methods to regain the trust of consumers. Pregnancies were also a huge factor in the model, so this can encourage the development of new medicines to reduce cancer risks linked to this variable.

The model designed could help researchers to understand the causes of different types of cervical cancers and how the discovery of new remedies can help to reduce the risks of contracting cancers. As such it also help to bring down the cost of researches by helping them to focus on specific data and reduce the amount of resources required. This would in turn reduce the cost of medical treatment in the long term of the costs are usually associated with the amount of resources invested to detect and treat cancer. A reduction is cost of healthcare would then help to increase customer satisfaction as the customers would no longer have to fear of the financial burden of cancer treatment.

Any country which lacks a proper healthcare system or that needs to modernize its healthcare would benefit from an affordable intelligent cancer detection system as it helps the government to focus on other areas of the economy, thus the healthcare system would be operating with lower costs in the long run and more efficiently. If this technology is applied to the data all citizens of a country, the government would be able to have an overview of the healthiness of its population and be able to provide the necessary cancer treatment facilities with the necessary equipment and personnel. Moreover if the geographical location of the different patients diagnosed is recorded, it would be possible to help the government in determining the most suitable location where a cancer treatment facility needs to be built and how many patients the treatment facility must be able to accommodate.

Hence the use of business intelligence will act as a decision-support system which would help the different stakeholders involved in the cervical cancer treatment to perform their roles using data driven facts generated by the business intelligence model. The system would enable the latter to make timely decisions with confidence and increase the efficiency of each of their different business activities. However there are initial cost factors which are involved initially as the system requires the participation of data scientists to design and improve the current model developed for the scope of this study but the eventual benefits outweigh the investments in the long run against other traditional business intelligence tools.

4 References

Fernandes, K., Cardoso, J.S. and Fernandes, J., 2017, June. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis* (pp. 243-250). Springer, Cham.

Mettler, T. and Vimarlund, V., 2009. Understanding business intelligence in the context of healthcare. *Health informatics journal*, *15*(3), pp.254-264.

5 Bibliography

Foshay, N. and Kuziemsky, C., 2014. Towards an implementation framework for business intelligence in healthcare. *International Journal of Information Management*, *34*(1), pp.20-27.

Hočevar, B. and Jaklič, J., 2010. Assessing benefits of business intelligence systems–a case study. *Management: journal of contemporary management issues*, *15*(1), pp.87-119.

Odom, M.D. and Sharda, R., 1990, June. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks* (pp. 163-168). IEEE.