Data Mining

Final assessment

Dataset: Wisconsin Breast Cancer

Klaus Konadu-Finke

Fontys Venlo

Contents

Introduction:

The dataset use is the Wisconsin Breast Cancer dataset from UCI (https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin). This dataset is used for the Data mining module in order to get more familiar with machine learning techniques and do independent research on the chosen topic. This report will firstly describe the personal learning goals, describe the problem afterwards and state a few research questions. Next there will be a short analysis, data description and prediction of results. The analysis will discuss a few visualizations. After that there will be a description on the procedure and methods used to build the neural network. Then the results are evaluated and other possible modes are discussed. Finally a reflection on the achieved earning goals are made.

Learning Goals:

In this project the aim is to predict the output value Y from input values Xn given a test and training dataset. The prediction on if the mutated cells are benign or malignant given the attributes Xn. Using a machine learning engine from Microsoft Azure, the data is first loaded, manipulated and then the model is trained to predict the nature of the cancer cells. The goal is to predict the result as accurately as possible (hopefully more than 70%) by comparing the test data output with the predicted output.

Problem Description:

The problem is that the number of diagnosis made is very little compared to the amount of patients. The cell mutation of cancer grows the longer one waits, hence a fast diagnosis and a fast treatment is essential to cause as little damage to the human body. Machine learning is used to evaluate previously made diagnosis on whether the found cancer cells are benign or malignant and uses this training data to predict future outcomes.

Research Questions:

To what extent may mutated cells be predicted to be benign or malignant?

What is the accuracy of the chosen classifiers as compare to others?

Data Description:

The dataset describes cell scans of breast cancer. The cells have the attribute Size, shape the cell tissue size.

Epitheliy cells are human primary cells https://www.promocell.com/products/human-primary-cells/epithelial-cells/. Basically this is the cell size of the scanned tissue.

Bare nuclei shows the amount of electrons either deprived or constructed to the cell normal count of electrons. This leads to a smaller cytoplasm off the cell

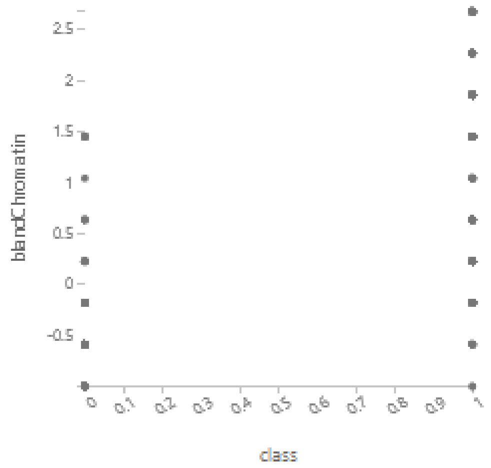
Chromatin show the level of Chromatin in the genetically mutation of the cells. If at least three genes are turned off, a normal cell can be converted into a cancer cell. https://www.nature.com/articles/1204322

Mitosis is the process which describes cell division, producing copies of themselves. Cancer is essentially mitosis that is out of control. So the larger the mitosis count, the higher the possibility of a malignant cancer tissue. https://www.biology.iupui.edu/biocourses/N100H/ch8mitosis.html

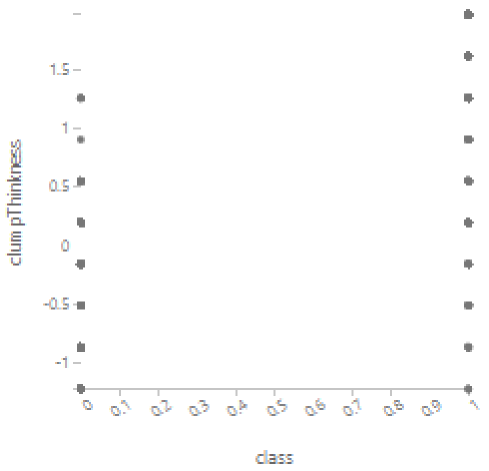
Prediction:

A development of cancer cells would show a lager cell size, chromatin and mitosis count. An irregular shape and a high rate of cell adhesion may also indicate malignant cells.

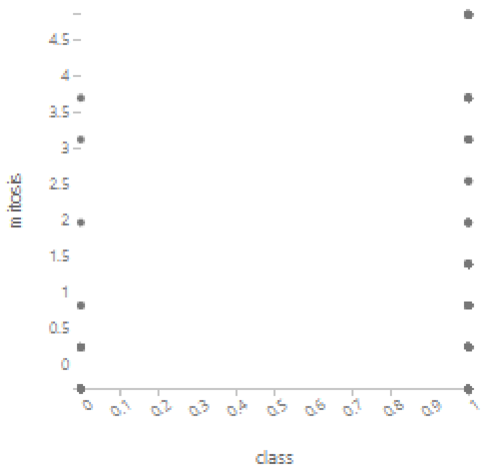
Analysis:



The figure above shows that the higher the bland Chromatin levels is, the ore likely the cell is to develop malignant cell tissue. However, there are also malignant cells which have relativly low chomatin levels.



Generally know for brest cancer is that there might be lumps resulted from mutated lobes. The table above shows the clump thikness against the class or cancer.

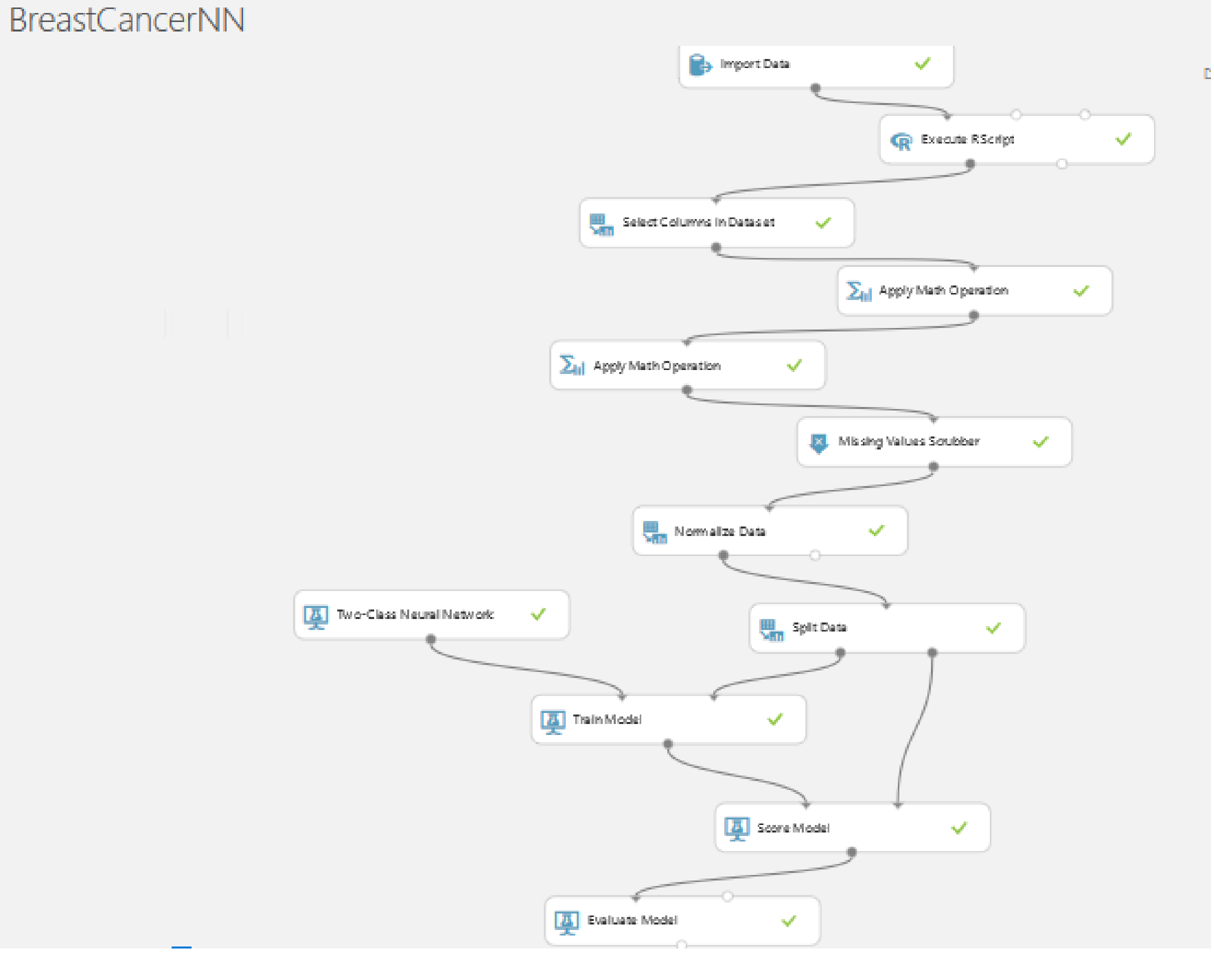


The analysis concludes that the larger and more deformed the cells are and the greater the cell mitosis, the ore likely the chance of a malignant cancer cell.

Development Steps:

1. Import the breast cancer dataset
2. Added column names via an R script
3. Deleted unnecessary columns to make the program more efficient (deleted the index column)
4. Applied two math operation which classified the score in either 1 or 0
5. Looked at missing values and decided to delete those. Only one row had a missing value
6. Normalize all columns except of the class column
7. Split the data into training (80%) and test data (20%)
8. Train the model using a two class neural network
9. Assess the accuracy by comparing the training data with the test data

Here is the setup of the steps produced in designing the neural network



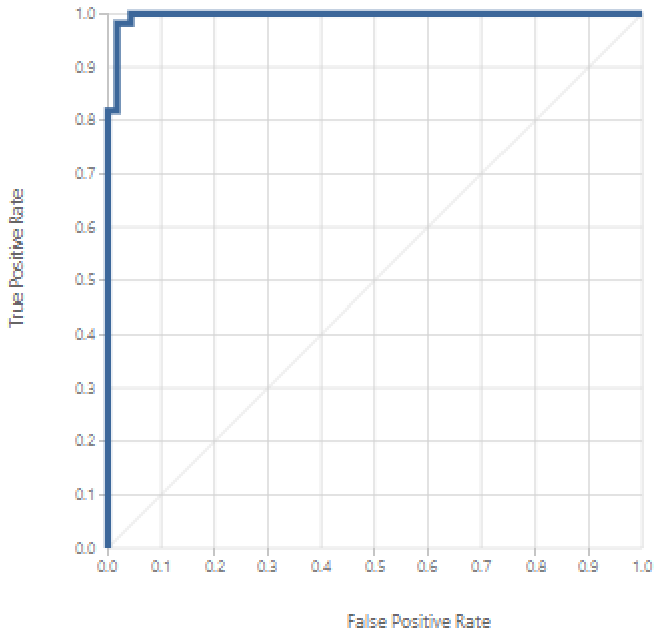
Model Decision

This problem is a prediction of labeled data which concludes to a regression problem. Different regression models were used and compared against each other to make a decision.

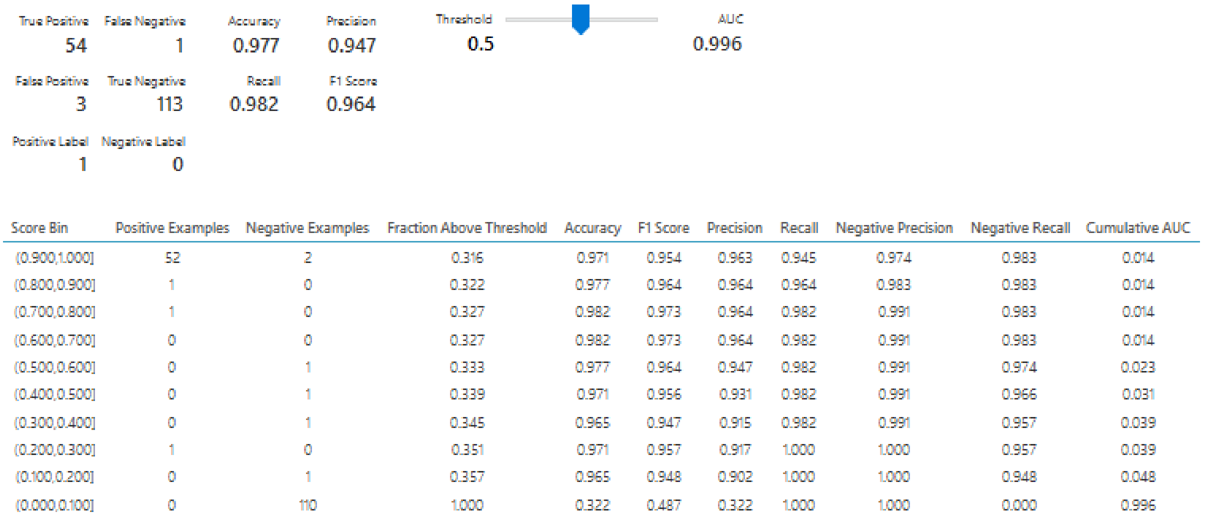
Different classifier models were used and compare against their accuracy and their time to complete. Below, the model with the most accuracy score is the *two classed averaged perceptron* classifier with 98.8%. However it is one of the slowest and if the input data increases, performance issues arise. The model with the fastest algorithm is the 2 class neural network which took a total of 24 seconds to complete. The experiment was done three times und the same conditions to ensure that the anomaly count was kept low.

Model	Average accuracy	Average time to complete
Multiclass neural network	0.976608	39
Two class neural network	0.982	23
Muliclass decision forest	0.97076	28
Two classed averaged perceptron	0.988	35
Two class boosted decision tree	0.982	34

Evaluation



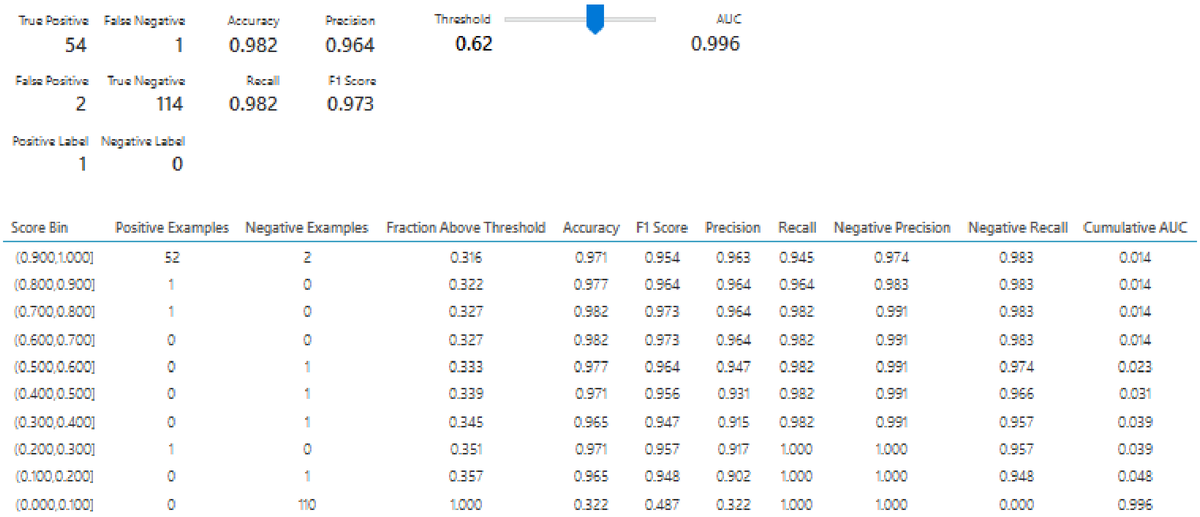
The figure below shows the first results of the training model. The accuracy is better than expected but the threshold was adjusted to create a higher accuracy.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
54	1	0.977	0.947	0.5	0.996
False Positive	True Negative	Recall	F1 Score		
3	113	0.982	0.964		
Positive Label	Negative Label				
1	0				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	52	2	0.316	0.971	0.954	0.963	0.945	0.974	0.983	0.014
(0.800,0.900]	1	0	0.322	0.977	0.964	0.964	0.964	0.983	0.983	0.014
(0.700,0.800]	1	0	0.327	0.982	0.973	0.964	0.982	0.991	0.983	0.014
(0.600,0.700]	0	0	0.327	0.982	0.973	0.964	0.982	0.991	0.983	0.014
(0.500,0.600]	0	1	0.333	0.977	0.964	0.947	0.982	0.991	0.974	0.023
(0.400,0.500]	0	1	0.339	0.971	0.956	0.931	0.982	0.991	0.966	0.031
(0.300,0.400]	0	1	0.345	0.965	0.947	0.915	0.982	0.991	0.957	0.039
(0.200,0.300]	1	0	0.351	0.971	0.957	0.917	1.000	1.000	0.957	0.039
(0.100,0.200]	0	1	0.357	0.965	0.948	0.902	1.000	1.000	0.948	0.048
(0.000,0.100]	0	110	1.000	0.322	0.487	0.322	1.000	1.000	0.000	0.996

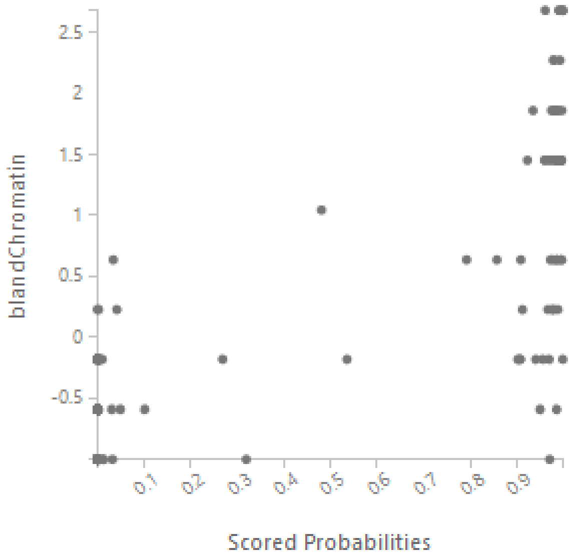
Adjusting the threshold to 0.62, increasing the accuracy to 98.2% and reducing the false positives by 1.

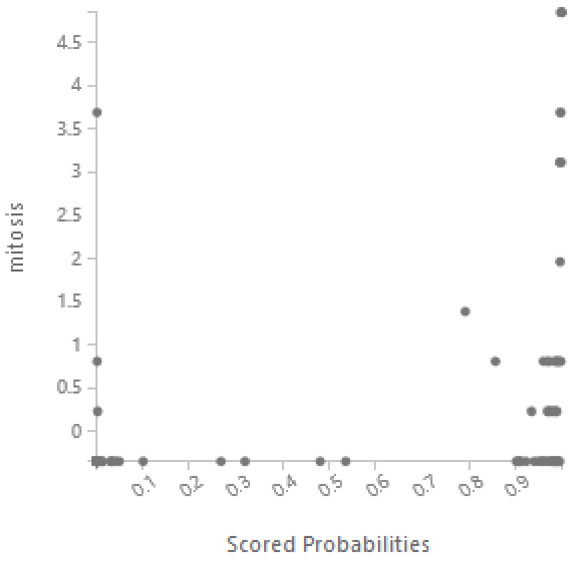


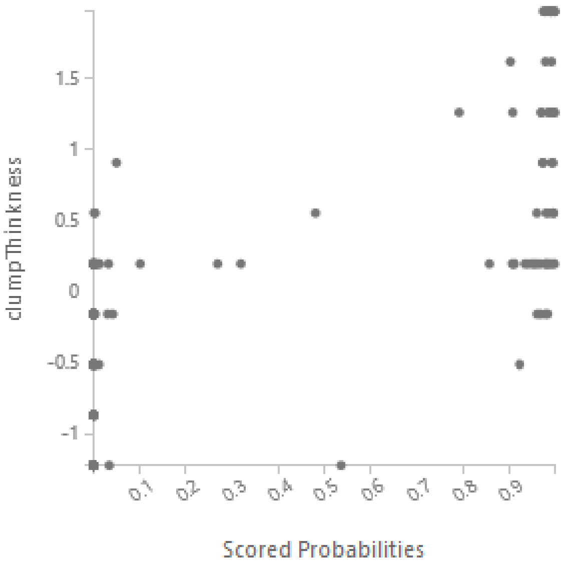
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
54	1	0.982	0.964	0.62	0.996
False Positive	True Negative	Recall	F1 Score		
2	114	0.982	0.973		
Positive Label	Negative Label				
1	0				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	52	2	0.316	0.971	0.954	0.963	0.945	0.974	0.983	0.014
(0.800,0.900]	1	0	0.322	0.977	0.964	0.964	0.964	0.983	0.983	0.014
(0.700,0.800]	1	0	0.327	0.982	0.973	0.964	0.982	0.991	0.983	0.014
(0.600,0.700]	0	0	0.327	0.982	0.973	0.964	0.982	0.991	0.983	0.014
(0.500,0.600]	0	1	0.333	0.977	0.964	0.947	0.982	0.991	0.974	0.023
(0.400,0.500]	0	1	0.339	0.971	0.956	0.931	0.982	0.991	0.966	0.031
(0.300,0.400]	0	1	0.345	0.965	0.947	0.915	0.982	0.991	0.957	0.039
(0.200,0.300]	1	0	0.351	0.971	0.957	0.917	1.000	1.000	0.957	0.039
(0.100,0.200]	0	1	0.357	0.965	0.948	0.902	1.000	1.000	0.948	0.048
(0.000,0.100]	0	110	1.000	0.322	0.487	0.322	1.000	1.000	0.000	0.996

Below are the equivalent of figures as in the Analysis but for the scored model. The scores between 2.0 and 0.8 show indecisive results.


blandChromatin
2.5
2
1.5
1
0.5
0
-0.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Scored Probabilities


mitosis
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Scored Probabilities


clumpThinkness
1.5
1
0.5
0
-0.5
-1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Scored Probabilities

Reflection on achievement learning goals

The data could be successfully imported and manipulated to fit the machine learning models. Different models could be evaluated and the best for this problem was chosen. The best suitable model for this problem is the 2 class neural network, due to its performance and speed. A goal of an accuracy higher than 70% was achieved with as little as 2 false positives and 1 false negative.