CPS 844 Assignment 1

Classification Methods

Omar Syed - 500809837

A S M Rubayet Ahmed - 500962603

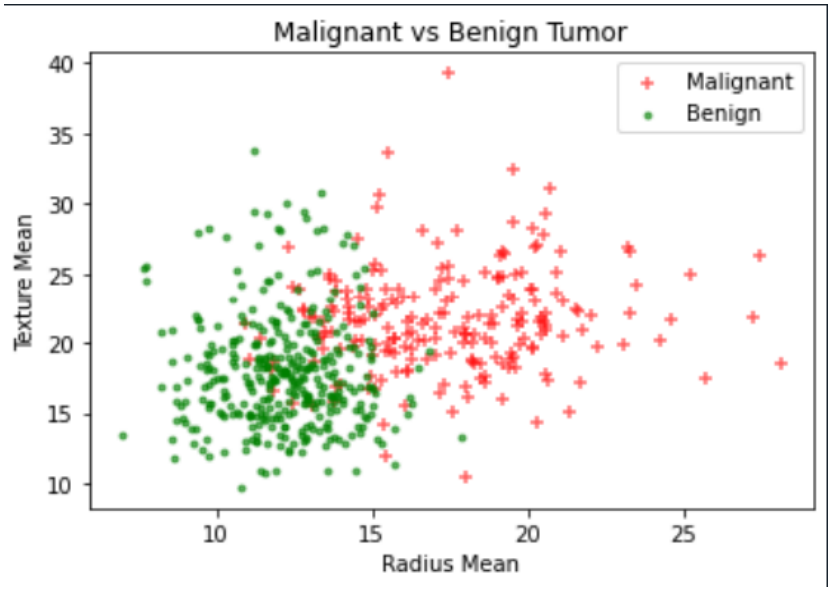
The Dataset:

For our assignment we are using a Breast Cancer Wisconsin (Diagnostic) dataset from Kaggle. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. It contains the following: ID number, Diagnosis, Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal Dimension. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All the values in the data are recorded to four significant values. The data classifies breast cancer tumor using two training classification:

- 1 = Malignant (Cancerous) – Present
- 0 = Benign (Not Cancerous) – Absent

The dataset contains 569 examples of benign and malignant tumors and 32 columns. The class distribution is as follows: 357 Benign, 212 Malignant with a total of 17639 data points. [3]

Using the data we plotted a graph that helped us visualize the Radius-Mean (x-axis) and Texture-Mean (y-axis) of the Benign and Malignant tumors.



Classifiers:

For the dataset we decided to use the following classification methods:

1. Gaussian Naïve Bayes
2. Support Vector Machine (SVM)
3. K-Nearest Neighbor Algorithm (KNN)
4. Decision Tree Classifier
5. Artificial Neural Network (ANN)

Gaussian Naive Bayes Classification

The Naive Bayes classifier separates data into different classes according to the Bayes' Theorem:

$$P(A|B) = \frac{P(B \mid A) \, . \, P(A)}{P(B)}$$

In the equation for Bayes' Theorem the values A and B are two events. P(A) is the independent probability of A and P(B) is the independent probability of B. P(A|B) defines the probability of event A provided event B has occurred already. P(B|A) defines the probability of event B provided event A has occurred already.

We can further expand the above equation by taking 'B' as independent predictors or attributes or data and taking 'A' as class.

It assumes that all the predictors are independent of one another and assumes that a particular feature in a class is not related to the presence of other features. Gaussian Naive Bayes is employed when the predictor values are continuous and are expected to follow a Gaussian distribution.

For our dataset we used the attributes 'radius_mean' and 'texture_mean' as the most important attributes. After dropping the irrelevant data, we plotted a graph to visualize the data after which

we preprocessed the data, where malignant and benign were assigned values of ‘1’ and ‘0’ respectively. Then we divided the data into x components (this variable includes independent predictor factor) and y components (this variable provides the diagnostic prediction). After normalizing the data, we used the train_test_split module from the package sklearn so we can divide the data into testing and training sections. Then, we used the Gaussian Naive Bayes module from sklearn on our data and got an accuracy of 93.567. This means that this module can help us determine based on the attributes whether or not a tumor is benign or malignant with 93.567 accuracy.

There are a lot of advantages to using Naive Bayes classifier as it is very simple and fast to implement. It works very well with large datasets and does not take too much time for training. Furthermore, despite recent major breakthroughs in machine learning, Naive Bayes is still among the most powerful algorithms. However, typically to achieve a good result it requires a large dataset. It also assumes that all features are independent which is rarely the case as in most situations, features show some dependency. The classifier also sometimes suffers from zero probability problems.

Support Vector Machine Classifier:

Support Vector Machines (SVM) are a set of supervised learning models with associated learning algorithms that are used for classification, regression, and outlier detection. In SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. After that we perform classification by finding the hyper-plane that differentiates the two classes. [7][8]

For our dataset, we dropped the irrelevant values and assigned target and used the attributes ‘radius_mean’ and ‘texture_mean’ as the most important attributes. We then used train_test_split module from the package sklearn so we can divide the data into testing and training sections. For kerner we used ‘linear’ and made a prediction model. We then plotted a confusion matrix and at last got an accuracy of 94.7368. This means that this module can helps us determine based on the attributes whether a tumor is benign or malignant with 94.7368 accuracy.

```
[1 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 1 2 2 1 2 2 1 2 1 1 2 2 1 2 2
1
 2 1 2 2 2 1 1 1 1 1 2 2 2 1 2 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 2 2 1 1
2
 2 1 1 1 2 1 1 2 1 1 1 2 2 2 2 1 2 2 1 2 2 2 2 2 2 1 1 2 2 1 2 2 1 1 2 2
2
 2 2 1 1 2 1 2 2 1 1 2 2 2 2 1 2 2 2 1 2 2 2 2 1 2 2 2 2 2 1 1 2 2 2 2 2
1
 2 2 2 1 2 1 2 1 2 2 2 1 1 1 2 1 1 2 1 2 2 2 1]
Confusion Matrix
[[ 56   4]
 [  5 106]]
Test Set: 171
SVM Accuracy: 94.73684210526315 %
```

SVM works relatively well when there is a clear margin of separation between classes and is very effective in high dimensional spaces. It is quite memory efficient and is effective in cases where the number of dimensions is greater than the number of samples. However, SVM algorithm is not suitable for large datasets as it does not perform well when the target classes in the data are overlapping. It will underperform if the number of features for each data point exceeds the number of training data samples. [6]

KNN Classification -

The KNN or K-Nearest Neighbor algorithm is a widely implemented supervised machine learning algorithm which is used to solve classification problems. K-Nearest Neighbor algorithm uses neighboring data points in a larger dataset that present similarity in specific groups to generate results with high accuracy. The KNN algorithm generates classification values based on the Euclidean Distances between the neighboring values with cases relevant to class with the most number of commons amongst its neighbors.

In our breast cancer data set, one of the classification methods we applied was the K-Nearest Neighbor algorithm. With our key attributes among many were the radius and the texture value of the tumors represented within our dataset. We were able to use the K-Nearest Neighbor algorithm to create a machine learning program that needs minimal to no training and is able to represent and divide the relevant date almost immediately. Due to the fact that this algorithm is able to operate at such quick speeds because of the menial to no training, this

algorithm is perfect for softwares that require instant results, allowing it to work in strong conditions at real time (Kumar, 2019). Since it is able to work with minimal to no training, K-Nearest Neighbor algorithm works best with smaller data sets, with accuracy significantly decreasing when compared to other algorithms in larger data sets (Kumar, 2019). Since our data set does not classify as a massive dataset but not a small one either, with the size of 17639 data points, we were able to run the K-Nearest Neighbor algorithm with a decent classifier accuracy of almost 93.21%. With a generated 93.21% accuracy from our KNN algorithm, we know that the algorithm was able to determine with 93.21 percent accuracy whether a not a data point, or tumor in our case would be labeled as a benign tumor or a malignant tumor.

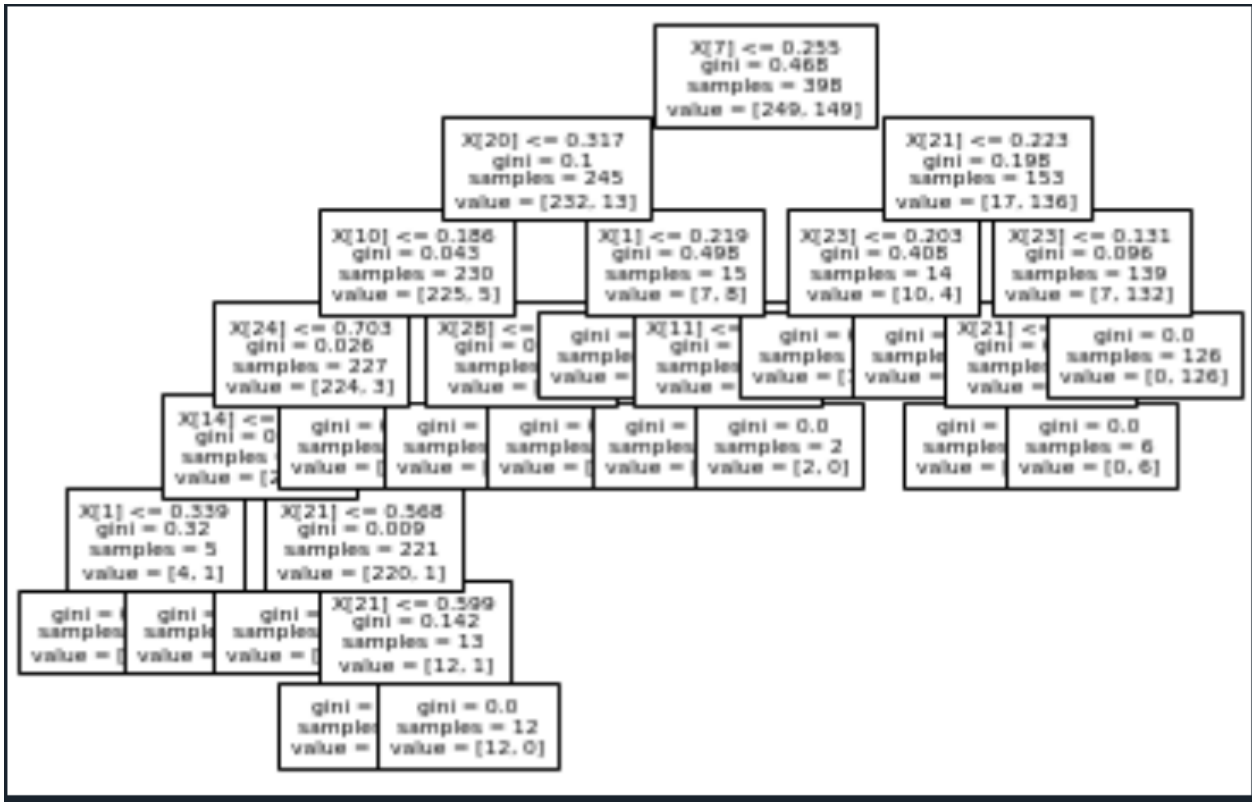
```
    diagnosis  radius_mean  ...  symmetry_worst  fractal_dimension_worst
0           1        17.99  ...          0.4601                  0.11890
1           1        20.57  ...          0.2750                  0.08902
2           1        19.69  ...          0.3613                  0.08758
3           1        11.42  ...          0.6638                  0.17300
4           1        20.29  ...          0.2364                  0.07678
5           1        12.45  ...          0.3985                  0.12440
6           1        18.25  ...          0.3063                  0.08368
7           1        13.71  ...          0.3196                  0.11510
8           1        13.00  ...          0.4378                  0.10720
9           1        12.46  ...          0.4366                  0.20750
10          1        16.02  ...          0.2948                  0.08452
11          1        15.78  ...          0.3792                  0.10480
12          1        19.17  ...          0.3176                  0.10230
13          1        15.85  ...          0.2809                  0.06287
14          1        13.73  ...          0.3596                  0.14310

[15 rows x 31 columns]
The accuracy of the classifier is 0.9321608040201005
17639
```

With our dataset being larger than that recommended for KNN, and lack of scaling features that would allow the modified algorithm to operate with a much larger dataset, the algorithm is only able to hit a 93.21% accuracy threshold constantly, unable to scale properly. The K-Nearest Neighbor algorithm also struggles with outlines and any data discrepancies that could generate noisy data or even missing values since it is generally untrained. These types of data points have the potential to significantly affect the accuracy of the classification (Kumar, 2019).

Decision Tree Classification -

The Decision Tree Classification Algorithm is a machine learning algorithm that is a supervised learning method. Decision Trees can be used for classification problems as well as regression problems, but are suited best for classification problems. A decision tree uses nodes to hold values known as Leaf Nodes and Decision Nodes. These branches of the tree are split according to their parameters, whether or not data points within a dataset fall in either direction of the tree. When looking at classification trees generated from the Decision Tree Classification Algorithm, an entropy value is needed to determine the amount of randomness in the attributes for the classification. This entropy value allows the algorithm to train the machine to discriminate values for classification near entropy. This training model then can be used to predict values for target variables by applying simple decision rules inferred from the previous data (used in training).

In our breast cancer data set, we used the Decision Tree Classification Algorithm which was able to train the program to determine with almost 93.5% accuracy to determine whether the respective data represents a tumor that is benign or malignant. Decision trees are very simple and easy to understand and work with since they can easily be visualized in the shape of an actual tree, as shown in the visual representation breakdown of our decision tree being applied to the breast cancer dataset.

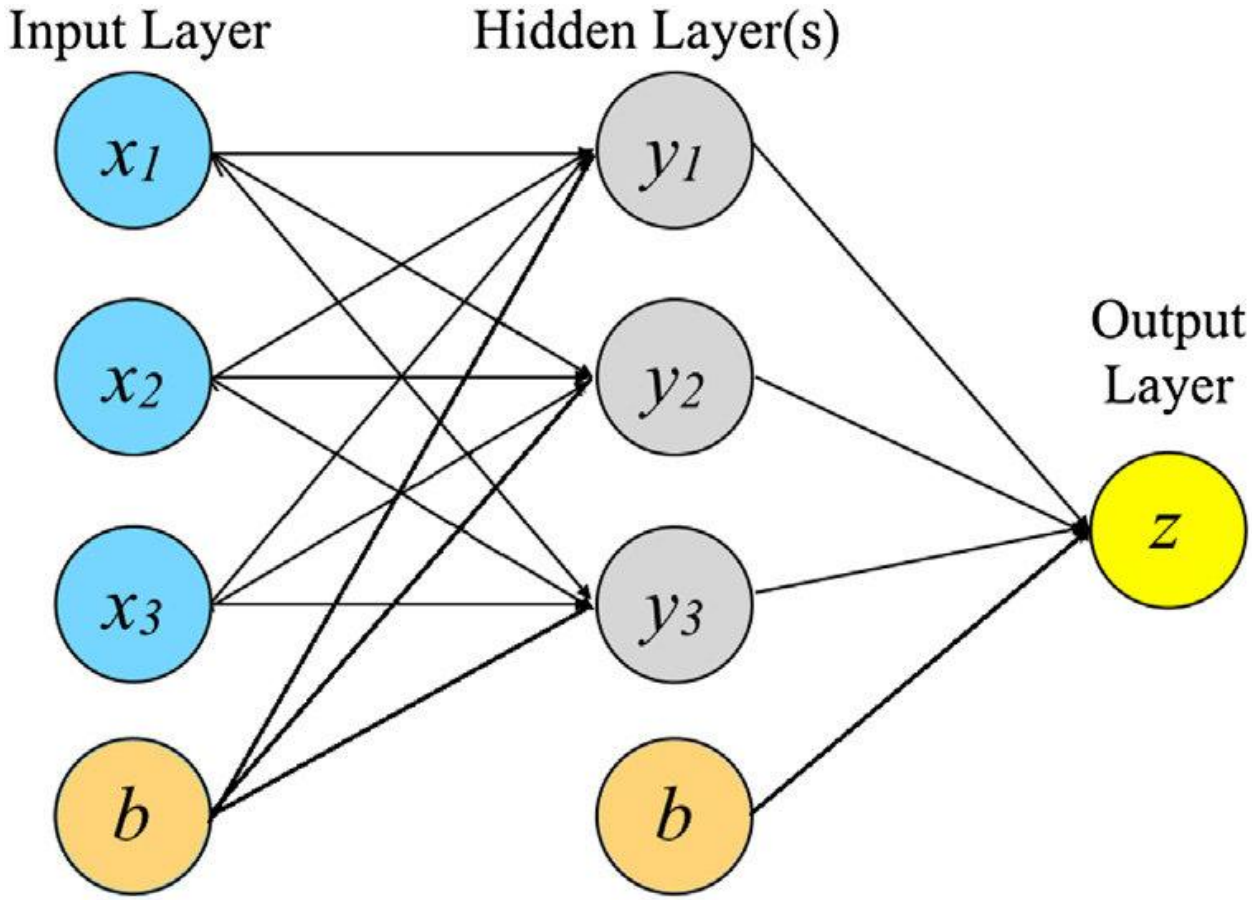


Since all our data set was too big to go through manually, the benefits of the Decision Tree Algorithm is that it does not require us to predetermine any hot coding values or dummy variables, as well as no assumptions about the shape of the data. With a dataset the size of 17639 data points, the decision tree is automatically about to determine feature selections and is able to find values that are unimportant to skew the classification. Our Decision Tree Classification was able to generate an accuracy of 93.5 % which is fairly stable considering the program needs to be trained at each run. With values overlapping and certain outliers, our decision tree struggles to exceed that accuracy without further algorithm modifications such as limiting tree depth to prevent overfits.

ANN Classification -

The ANN or Artificial Neural Network Algorithm is a complex algorithm that can be used for classification problems such as ours. If we are attempting to determine binary outputs, in our case whether a tumor is benign or malignant. We would use the Artificial Neural Network algorithm that has one ‘neuron’ at the output layer, either 1 or 0 value will be outputted as

referred above, 1 = Malignant (Cancerous) – Present, 0 = Benign (Not Cancerous) – Absent. When using an Artificial Neural Network Algorithm, the code generates layers for each data, these layer layouts vary based on what kind of dataset you have and often requires changing of multiple variables to result in the highest possible accuracy. Neural Networks work similar to how the human brain works, each neuron performing a set of functions carrying data to determine results that have the highest probability of succeeding when repeated multiple times.



For mining large sets of data, Artificial Neural Network is a supervised classification model that is commonly used as they can use MLP Classifiers. Also known as Multi-layer Perceptron classifiers which are able to connect to the Artificial Neural network. MLP Classifiers consist of three layers, the input layer for the data, a hidden layer where the neural network is able to judge functions and probability and then the output layer. This output layer in our case generates a binary representation of the tumor - malignant or benign. With our representation of

the Artificial Neural Network using the Multi-Layer Perceptron Classifier for our breast cancer classification problem, we are able to generate an accuracy of approximately 91,812 %. This may be lower than a few of the simpler classification methods as well as being a lot more complex. But due to the complexity of this algorithm, the applications of it are vast as it has quickly become one of the best known machine learning algorithms.

Since Neural networks are flexible, they operate with both regression and classification problems, essentially any type of data can be represented in numerical form. Since our breast cancer data set is purely numerical with a binary output representation, it is a reliable approach to model our data. With over 17000 data points, the artificial neural network is able to be trained relatively quickly and once trained is able to generate predictions repeatedly very quickly with good accuracy under optimal conditions. Only issue with such neural networks is that they are nearly impossible to track since it is almost impossible for us to determine how each variable is influenced by others.

Conclusion:

Although all of the five classification methods we used for our dataset had an accuracy of greater than 90, SVM had the highest accuracy with 94.74, while ANN had the lowest accuracy with 91.88. Thus, SVM is the best algorithm for our dataset.

Sources

1. Gupta, P. (2017, March 12). *Decision trees in machine learning*. Medium. Retrieved February 28, 2022, from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
2. Hashmi, F. (2021, November 26). *How to use artificial neural networks for classification in python?* Thinking Neuron. Retrieved February 26, 2022, from https://thinkingneuron.com/how-to-use-artificial-neural-networks-for-classification-in-python/
3.
4. Kumar, N. (2019, February 23). *Advantages and disadvantages of KNN algorithm in Machine Learning*. Advantages and Disadvantages of KNN Algorithm in Machine Learning. Retrieved February 26, 2022, from http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html
5. Learning, U. C. I. M. (2016, September 25). *Breast cancer wisconsin (diagnostic) data set*. Kaggle. Retrieved March 1, 2022, from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
6. MLNerds. (2021, August 9). *Naive Bayes classifier : Advantages and disadvantages*. Machine Learning Interviews. Retrieved March 1, 2022, from https://machinelearninginterview.com/topics/machine-learning/naive-bayes-classifier-advantages-and-disadvantages/
7. *Naive Bayes classifier: Pros & Cons, applications & types explained*. upGrad blog. (2021, December 14). Retrieved March 1, 2022, from https://www.upgrad.com/blog/naive-bayes-classifier/
8. K, D. (2020, December 26). *Top 4 advantages and disadvantages of support vector machine or SVM*. Medium. Retrieved March 1, 2022, from https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
9. Wikimedia Foundation. (2022, January 7). *Support-Vector Machine*. Wikipedia. Retrieved March 1, 2022, from https://en.wikipedia.org/wiki/Support-vector_machine

10. *SVM: Support Vector Machine Algorithm in machine learning*. Analytics Vidhya. (2021, August 26). Retrieved March 1, 2022, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/