Is it Possible to use Machine Learning Software to predict if cancers are benign or malignant?

A case study using K-Nearest Neighbors Algorithm on Breast Cancer

By: Topu Saha

New Jersey City University

May 2022

Table of Context

1. Background
2. Descriptive Analysis of Dataset
3. K-Nearest-Neighbors Algorithm
4. Procedure
5. Results and Model Evaluation
6. Conclusion
7. References
8. GitHub Link

Background

Cancer is one of the worst disease known to mankind. It is when cells in the body are growing uncontrollably. This process can be extremely harmful for the individual causing pain and can be life threatening if not properly treated. It is a growing problem for medical experts to fight as “in 2020, an estimated 1,806,590 new cases of cancer will be diagnosed in the United States and 606,520 people will die from the disease.” [2] With growing numbers, testing for cancer is very important. Doctors are not able to treat what they do not know exists.

The process of cancer testing starts with the patient. They come in expierencing symtomps that has the doctor concerned that they might have some form of cancer. This can be a lump they feel on the body which is a tumor. A tumor is a group of cancer cells that grew enough to be large. Often times these cells are visible and disgintuishable. To diagnose cancer several techniques are used. It starts with some laboratory exams to “identify abnormalities that can be caused by cancer” [3]. Next is imaging such as X-rays, CT scans and MRIs which are non-inavsive ways for doctors to look inside your body and organs. The last test a doctor can do to test for cancer is called a biospy. A biopsy is when “your doctor collects a sample of cells for testing in the laboratory.” [3]

A biopsy will be able to tell if your cancer cells are benign or malignant. A test result of benign means that the cancer is not cancerous or harmful and removing the cells will relive the symptoms and danger. However, the result comes back maligant, it means the cancer is dangorus as it can spread to other parts of the body and cause harm. With the process of detecting and identifying cancer, we can see how important the biopsy results are as they are life changing and gives both the doctor and patient important information about their health. However a negative of biopsys are they the results take time.

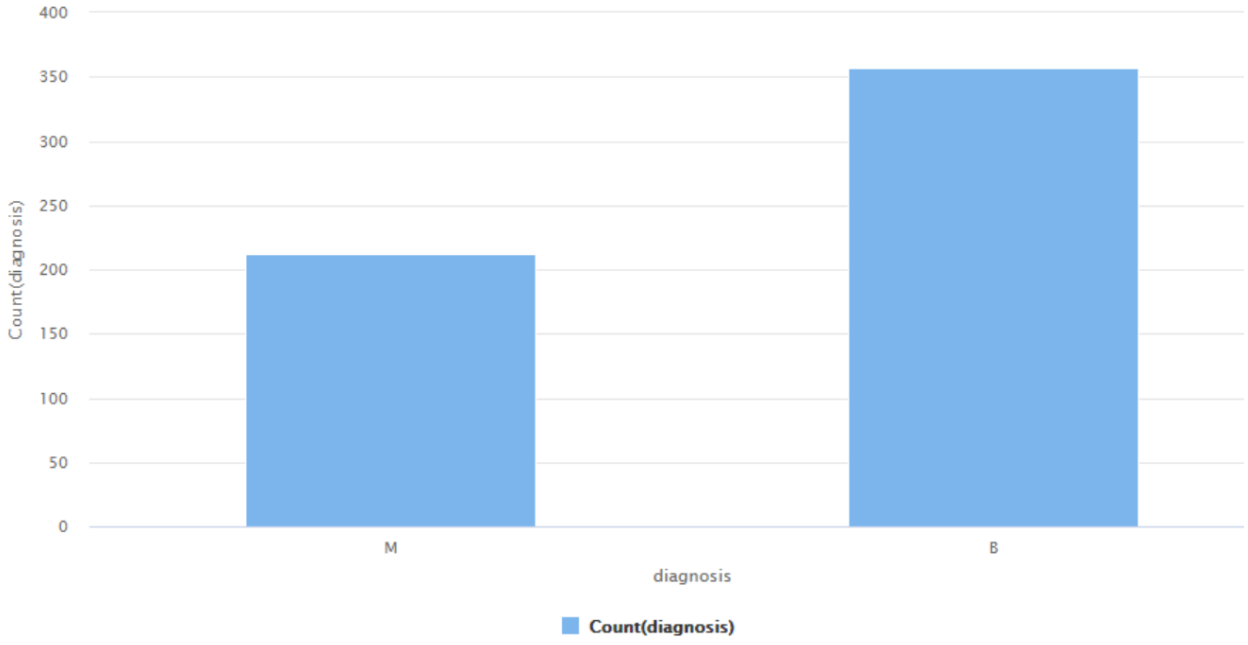
> “The amount of time it will take for you to receive the results of the biopsy depends on how many tests are needed on the sample to make a diagnosis. A result can often be given within 2 to 3 days after the biopsy. A result that requires a more complicated analysis can take 7 to 10 days. ” [3]

This is where the problem lies. Often times the days waiting can seem daunting. It creates unneeded stress for the patient. The goal of this project is to create a machine learning model to predict if cancers are benign or malignant. This model is designed to speed to the process of

diagnosing cancer, giving valuable information to the doctor and patient faster, limiting waittime and speeding up decision making. The model is designed to automate the biopsy testing and examining phase to help all pirates involved and save critical time.

Descriptive Analysis of Dataset

The dataset I am using in this analysis is from kaggle. It is a dataset that contains information on 569 breast cancer cells. The attrubutes are 33 attrubutes which are id, diagnoses, radius_mean. Texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, and att33. Besides the id and att3 attriiubte which are the values of Interand Normal, the rest are the type Real. Most importantly, the dataset is split up into 212 being malignant and 357 being benign. This dataset was used because of all the different attributes included to help classify breast cancer cells of being malignant or benign. In theory, the more details we have about the cells, the easier it is to classify them as malignant or benign.



K-Nearest-Neighbors Algorithm

The algorithm I have chosen to use to predict the dignaoses of these breast cancer cells is K-Nearest-Neighbors. K-Nearest-Neighbors is a supervised machine learning algorithm which means that it "relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data." [5] It can be used for classification or regression, but in this case we are using its classification abilities. It works by first plotting all the attributes and data is going to be plotted on a multidimensional graph which I am hoping can show a clear difference between the malignant and bengign breast cancer cells. After the difference is noticible by the computer, any new inputted data point will be classified into one of the group based on how close it is to the surrounding data points. For example, if a new data point is close to three malignant breast cancer cells and two benign breat cancer cells, then it will be classified as a malignant breast cancer cell. The amount of points the new point would have to be close to make it of that group. In our case we are using ten because after several trails, setting K equal to ten maximaized the accuracy. With this process, it is possible for a cell to be classified wrong due to it being an outlier. However, the probability of that happening should be low if the machine learning model is good.

Procedure

1. Download Breast Cancer Dataset provided on github repo linked at the end.
2. Retrieve Breast Cancer Dataset into a blank process .
3. Use the Set Role Opperiation and set the attribute name to diagnosis and target role to label. No additional role has to be set.
4. Use the Select Attributes Operation to select all the attributes used for the classification
    a. The attributes we are using are the following radius_mean. Texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst.

5. Next use the Split Operation to split the data into the ratio 70/30.
    a. 70% of the data will go towards training the model.
    b. 30% of the data will go towards testing the model.
6. Use the KNN operation and set the value of K to 10.
7. Use the Apply Model Opperation on the Test Dataset and the KNN Model.
8. Use the Performance Opperation to verify that the model is working properly.

Results and Model Evaluation

accuracy: 93.57%

	true M	true B	class precision
pred. M	56	3	94.92%
pred. B	8	104	92.86%
class recall	87.50%	97.20%	

PerformanceVector

```
PerformanceVector:
accuracy: 93.57%
ConfusionMatrix:
True:    M       B
M:       56      3
B:       8       104
```

Conclusion

The performance evaluation shows that the model is strong at classifying unlabeled data of breast cancer cells into the groups of miligant or benign. With an accuracy of 93.57%, this means that the model accurately classified the unlabeled data points from the test dataset 93.57% of the time. Giving confidence to this model, it is ready to be deployed.

Since our model is able to successfully predict if breast cancer cells are migliant or benign, we can assume that it is able to do this with any other cancer cell types as long as the information requirements are met. This proves our question that machine learning can predict if cance ris benign or malignant and if this is implemented in hospitals it will drastically reduce time to get results from biopsy and decrease stress of the patients. It will give doctors less time to diagnose, and come up with a treatment plant for the patient. It will also free up time of pathalogists, as they can spend their time doing activities that cannot be automated by current technology.

In future research, we can use other unsupervised machine learning algorithms to predict or classify benign or malignant cancers. These algorithms include Decison Trees, Logistic Regression, and Naive Bayes. All these algorithms are unsupervised and can be used to classify cancers in the way we are referring too. After running our analysis with these different machine learning algorithms, we can see which one has an higher accuracy and is more efficient.

Another idea is that we try to classify cancer cells into categories such as their king. So when doctors collect a biopsy, they can know what type of cancer it is since they're different types of cancers such as liver, breast, brain, and kidney. This can be an enitrely different process to help save time of doctors. They can use this model to detect what type of cancer it is and then use a second model to determine if it is malignant or benign.

Additionally, some improvements on this project would be to have more data. In statistics, generally the more data you have the more accurate results you will receive. Even though a 93.57% accuracy rate is very high, having a larger dataset could bump this accuracy score up into the 95 percent range. Another consideration to have is as the size of the dataset increases we can also increase the K value for the algorithm which can also increase the accuracy rate.

In conclusion, this project showed that it is possible to use machine learning to classify cancers into their groups of being benign or malignant. This is a start as the process can get more complex and be more beneifitcal in the real world setting. Using a sample dataset of breast

cancer was a general way of testing if the idea would work. Future work for this is important as it will greatly impact medicine and the world.

References

1. https://www.cancer.gov/about-cancer/understanding/statistics
2. https://www.mayoclinic.org/diseases-conditions/cancer/diagnosis-treatment/drc-20370594#:~:text=Imaging%20tests%20used%20in%20diagnosing,for%20testing%20in%20the%20laboratory.
3. https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/biopsy#:~:text=A%20biopsy%20is%20the%20main,place%20in%20your%20doctor's%20office.
4. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Github Link: https://github.com/Topusaha/Breast-Cancer-Classification