**Breast Cancer Data Analysis with Machine Learning**

1. Research Question: To develop a machine learning model to predict and classify breast cancer tumor cells into benign and malignant and predict the degree of malignancy of each tumor.

   I was able to develop a random forest for predicting the class of the breast cancer tumor (benign/malignant) with 95% accuracy and developed a linear regression model for predicting the degree of malignancy of tumor explaining 89% of the variance in our dataset through clump thickness, bland chromatin, and uniformity in the cell size.

   Our logistic regression model for the dataset had a very high deviance even with optimization whereas random forest model had similar performance parameters but 60% lower deviance. Therefore, random forest was our preferred choice as the final model to be used in further applications.

2. My findings agree with the existing literature and confirm that accurate machine learning models must be developed for early breast cancer diagnosis. This project provides a highly accurate and reliable machine learning model for predicting breast cancer tumor class and its degree of malignancy. Studies have shown that detecting breast cancer early through screening techniques can reduce the associated mortality by 13-16%.[1] Therefore, machine learning models can help speed up the process of diagnosis and the right intervention can prevent a benign tumor from becoming malignant if diagnosed early. In addition to this, we can use the measure of the degree of malignancy to further classify the breast tumors into more refined

Submitted by: Monika Sahai
1007645345

categories and the right intervention can slow the progression and possibly prevent breast cancer, if diagnosed early.

3. The original dataset was collected over time using the fine-needle aspiration (FNAs) technique, and malignant tumors were confirmed histologically. The original literature utilizing this dataset used a mathematical method called multisurface pattern separation which is a linear programming-based method used to distinguish between 2 different patterned datasets.[2] The authors used the 11 parameters of breast cytology to show and confirm that the method is also applicable to the diagnosis of breast cancer and can help classify them. Using linear programming, they correctly classified 369/370 samples (~99% accuracy), out of which 201 of them were benign and 169 malignant. Since then, there have been various other computational analyses on this dataset, using high-level machine learning algorithms such as support vector machine,[3] convolutional neural network,[4] to classify these tumor types with their respective accuracy being around ~95-97%. In this project, I developed a simple machine learning model using well-known methods like logistic regression model and random forest to achieve an appreciable accuracy (~95%). In addition, I was able to come up with a novel prediction of the degree of malignancy using PCA and linear regression under the guidance of Dr. Brian Cox. Therefore, findings of this project are in sync with the previous analyses done on this dataset.
4. Overall workflow

Submitted by: Monika Sahai
1007645345

**Breast Cancer Data Analysis with Machine Learning**

a. Entire code base has been divided into sections and code chunks with appropriate titles. In R-Studio, chunk navigator at the bottom can be used to easily navigate through titled sections.
b. Discussion for the figures is provided after its code and before proceeding to the next step.
c. Data cleaning, EDA
    i. These sections include removing missing or zero values, checking normality of all columns and transformations, correlations between various attributes. I haven't removed the outliers in this section yet and will do so as needed.
d. Machine learning analysis
    i. I first made Logistic regression model for the outcome variable *class* which showed very high deviance. Then I made random forest to find the most important variables and get rid of the outliers to improve performance, essentially bringing down the deviance from 1000 to 400 without affecting performance.
    ii. Then I proceeded to make a new column called *degree of malignancy* using PCA analysis. Linear regression model was made and 3 variables namely: blandChromatin, uniformityCellSize, clumpThickness(sqrt) explained 89% of the variance.

Submitted by: Monika Sahai
1007645345

**Breast Cancer Data Analysis with Machine Learning**

References


1. Duffy, S. W., Vulkan, D., Cuckle, H., Parmar, D., Sheikh, S., Smith, R. A., Evans, A., Blyuss, O., Johns, L., Ellis, I. O., Myles, J., Sasieni, P. D., & Moss, S. M. (2020). Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. *The Lancet Oncology*, *21*(9), 1165–1172. https://doi.org/10.1016/S1470-2045(20)30398-3
2. Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(23), 9193–9196. https://doi.org/10.1073/pnas.87.23.9193
3. Akinnuwesi, B. A., Macaulay, B. O., & Aribisala, B. S. (2020). Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in Medicine Unlocked*, *21*, 100459. https://doi.org/10.1016/j.imu.2020.100459
4. Zhu, C., Song, F., Wang, Y., Dong, H., Guo, Y., & Liu, J. (2019). Breast cancer histopathology image classification through assembling multiple compact CNNs. *BMC medical informatics and decision making*, *19*(1), 198. https://doi.org/10.1186/s12911-019-0913-x



Submitted by: Monika Sahai
1007645345