Project 2 Data Exploration Project using Python

Jesse Klug, Niranjani Srinivas, Vidisha Patel, and Xinyan Chen

University of Pittsburgh, Department of Health and Rehabilitation Services

Practical Statistics and Programming in Python and R

Dr. Yanshan Wang

December 14, 2021

I. Introduction

The following report is the combined work of team 3 of the HI 2020 SEC1090 class, consisting of team members Jesse Klug, Niranjani Srinivas, Vidisha Patel, and Xinyan Chen. This goal of this project is to demonstrate our proficiency in interpreting data using the Python programming language and learning how to use object-oriented programming and machine learning algorithms to make a meaningful impact in the diagnosis and treatment of real-world conditions. For this project, we chose to investigate the concept of using machine learning models on digital image of breast tumor cells to predict whether the tumor is benign or malignant (cancerous). The data has been used in previous studies and several visualizations and machine learning models have been trained on the data, but the work in this project is our own and any similarities to previous studies is coincidental.

The data set used in our project is titled "Breast Cancer Wisconsin (Diagnostic) Data Set". The data was collected by a team consisting of surgeons and researchers from the University of Wisconsin (Madison) Departments of Computer Science and Surgery. The data consists of 569 patients seen at the university with breast tumors. Samples were collected of the tumor tissue using a fine needle aspiration biopsy, which were then photographed by a color video camera placed on top of a microscope. Images were taken in regions with minimal nuclear overlap. Various measurements were taken of visible cell nuclei for each patient, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. For each variable the team recorded the mean, standard error, and the worst (mean of the three largest values), and recorded these measurements in a data set, totaling 30 features. After one year each patient's final diagnosis was known and the researchers used the measurement data along with the final diagnosis to train a machine learning model to predict

whether the tumor was cancerous or benign. The full study was published in the Journal of the American Medical Association in 1995. Our team obtained the data set from Kaggle, and the original database is still stored in the University of Wisconsin Computer Science ftp server. To better understand the subject of breast cancer and the use of machine learning models on cell images, we conducted research on relevant journal articles.

II. Specific Aims

Our goal for this project is to learn how we can use the Python programming language to recognize trends in the obtained data set and train a machine learning model to use the recorded measurements of tumor cell nuclei to predict whether a tumor is cancerous or benign. The predictor variable is categorical, in the data set a positive case is identified as “Benign” and a negative is identified as “Malicious”. We will be training several machine learning models and then comparing their performance based on metrics such as accuracy, kappa, and F1 score. We would also like to rank the variables used by successful machine learning models to better understand what features have the most impact in determining a diagnosis. The goal of this project is to prove the practical usage of machine learning models in cancer diagnosis, and if our machine learning model is successful, it could be implemented in hospitals to quickly diagnose future cases of breast cancer with minimal invasive procedures.

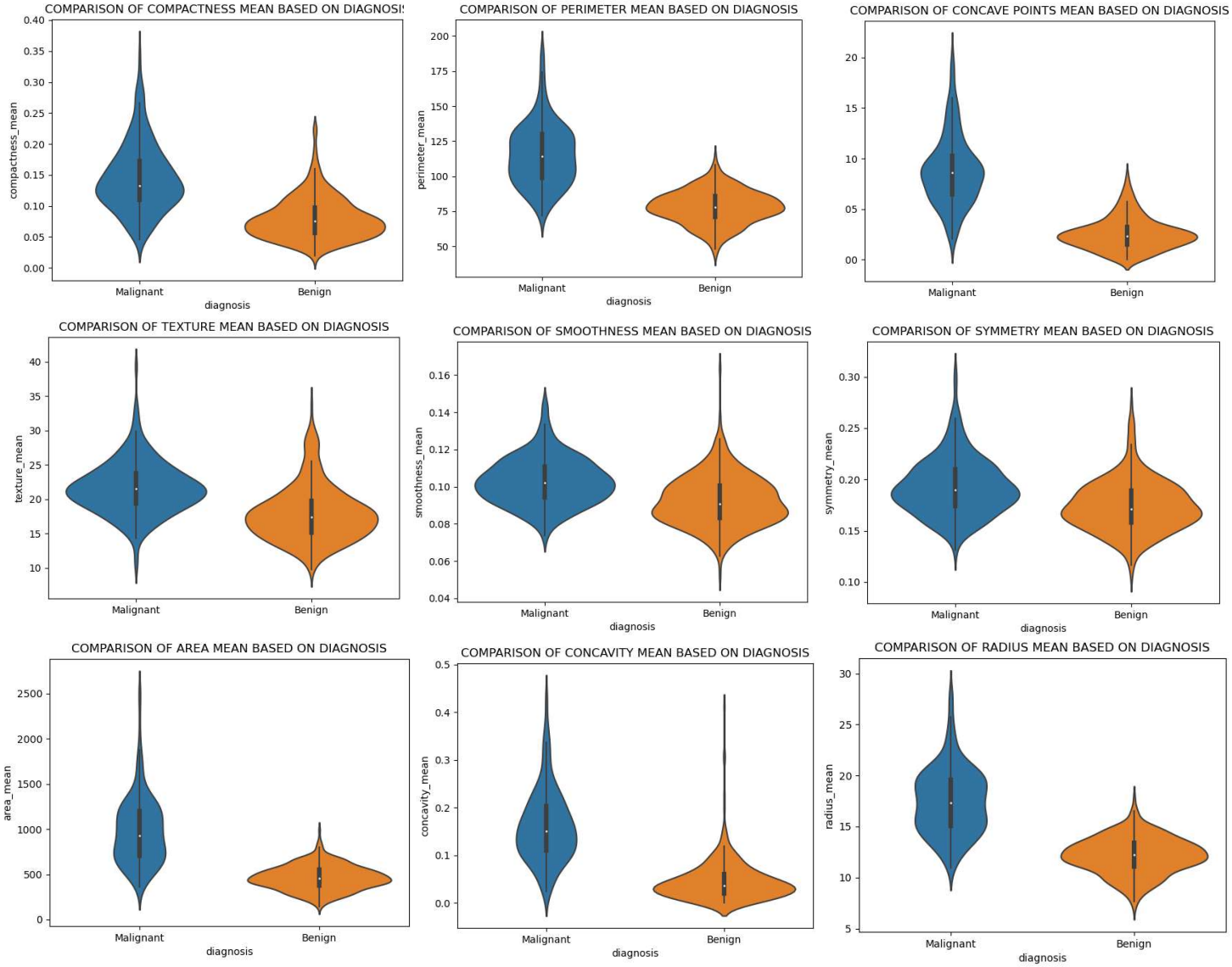
III. Quality Control of Database

The data used for this project was very clean and elegantly prepared. Out of the 569 patients and 30 features, there was no missing data. Each patient had a patient ID number recorded that was removed as it was unnecessary for training our machine learning models but otherwise all data was utilized for training our machine learning models.

IV. Data Analysis Part I: Descriptive Statistics to Analyze Data

We performed statistical analyses of the data to understand how the measurements differ based on the known outcome. We decided to use paired violin plots to compare the means of each of the ten features between “Malignant” and “Benign” to give a visualization of how the data is distributed as well as identify outliers. All measurements on the Y axis are in micrometers.

Figure 1: Violin plots comparing means of malignant and benign breast tumors



V. Data Analysis Part II: Inferential Statistics or Machine Learning

To prepare the machine learning model we separated the data into a training and testing group. We divided 80 percent of the training group, which consists of 455 samples, leaving 114 samples to test the accuracy of our models with. The models we chose to perform our predictions were the Random Forest, Logistic Regression, and Gradient Boosting Classifier. The random forest machine learning uses a bootstrapped data set and running a series of decision trees using random factors to try to predict the known outcome. Though many of the trees will not yield useful data the aggregate of the results should trend towards a meaningful and accurate prediction. Logistic regression is a form of linear regression used for a binary categorical variable. In our data this target variable is the state of the tumor. The shape of the line is an S shaped curve formed using data points from the training date, and the testing data is mapped to the curve. If a data point on the regression curve is on the flat top or bottom then the model has confidently predicted its category, and if it falls along the slope of the curve then the model will assign its prediction with a lower degree of confidence. The Gradient Boosting Classifier is a type of Gradient Boosting Machine and uses decision trees like the random forest and is used to predict categorical variables in new data. The model first determines a base probability of each value of the target variable occurring in new data based on the existing odds of it occurring in the training data. The model then creates a series of decision trees based on the training data and adjusts the odds of each value of the target variable being correct by validating its prediction against the true outcome and using the residual (difference between the observed and predicted probability) to determine how well it performed. After determining that adding new trees does not increase performance, then it is ready to run new data through each tree to classify it with a high degree of accuracy.

The results of our machine learning model in classifying the test data set are listed below in the form of confusion matrices The confusion matrix shows how the model distributed the 114 test rows of the data set as a grid, comparing the predictions to the true data. Top left number of the matrix represents true positives, which indicates the number of rows that the model has successfully predicted as being a benign tumor. The top right cell shows the false positives, where the model incorrectly predicted a tumor as benign when the true outcome was malignant. The bottom right cell represents true negatives, where the model accurately predicted a tumor was malignant, and the bottom left cell shows false negatives, where the model predicted a tumor as malignant when it was truly benign.

Figure 2: Comparison of confusion matrices for each machine learning mode

RF	Reference	
Prediction	Benign	Malignant
Benign	75	4
Malignant	1	34

LogReg	Reference	
Prediction	Benign	Malignant
Benign	73	4
Malignant	3	34

GBC	Reference	
Prediction	Benign	Malignant
Benign	74	4
Malignant	2	34

From these results, it appears that the Random Forest performed best. It correctly classified 75 benign tumors, 34 malignant tumors, yet it misclassified 1 tumor as malignant and 4 as benign. The worst outcome in this scenario is a false positive, as patients require urgent care for treating tumors. In all four models there were four false positives. In addition to the visual interpretation of the confusion matrix we performed several mathematical measures to validate the performance of each model, shown below.

Figure 3: Comparison of accuracy metrics for each machine learning model

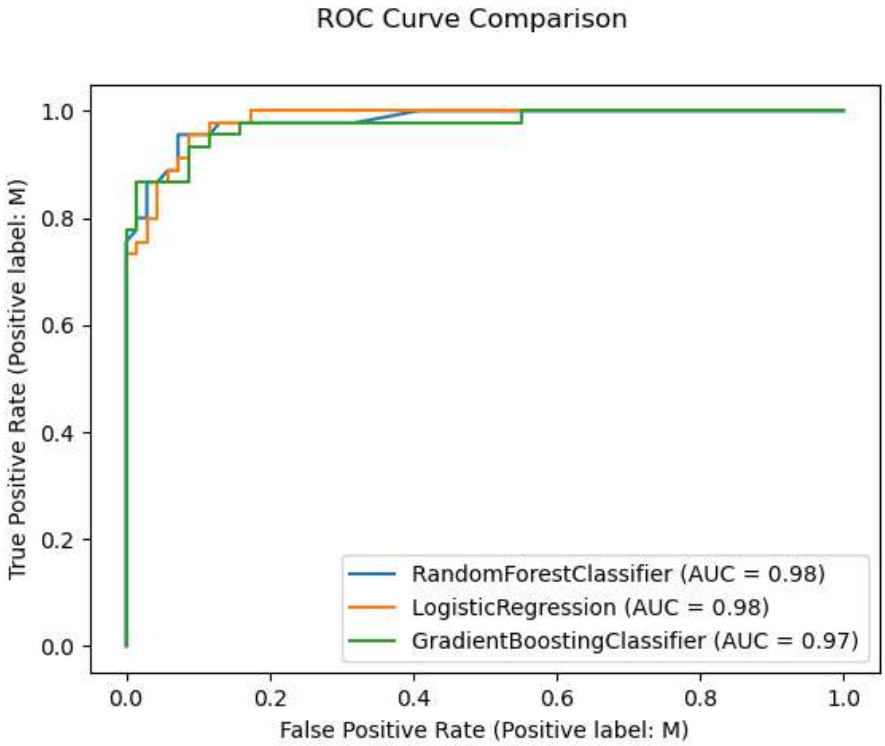
Method	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Kappa	Balanced Accuracy
Random Forest	0.956	0.949	0.971	0.949	0.968	0.899	0.545
Logistic Regression	0.939	0.948	0.919	0.948	0.954	0.88	0.535
Gradient Boosting Classifier	0.947	0.949	0.944	0.949	0.961	0.879	0.54

Each of these measures uses the results of the confusion matrix to show the performance in a way that is easier to compare. The accuracy is a quotient of the data that our model classified correctly against the total number of predictions. Sensitivity and specificity are measurements of what percentage of tumor cells were correctly predicted compared to the total of true classifications for each outcome respectively. Precision is a percentage of patient's tumors correctly classified as having a benign tumor compared to the total number of predicted benign tumors. The F1 Score is another measure of the model's overall accuracy that averages the results of precision and specificity. Kappa is a measure of how well the accuracy compares to a random assignment of categories. Balanced accuracy is the mean of the combined sensitivity and specificity. In every category the Random Forest model performed better or equal to the other models.

Another method we used to compare the model was to plot the models on a Receiver Operator Characteristic curve (ROC). The ROC curve is an aggregate of many different confusion matrices using the models while setting different thresholds of required sensitivity and specificity to show how the model performs at classifying predictions at each threshold. The metric for evaluating how models perform using an ROC curve is by determining the Area Under the Curve (AUC). A larger value for the AUC represents a greater amount of confusion matrices that performed well at predicting the type of tumor at each level of sensitivity and specificity. In the following ROC curve, both the Random Forest and the Logistic Regression

models scored equally well. Despite this, there are still apparent differences in performance along the curve.

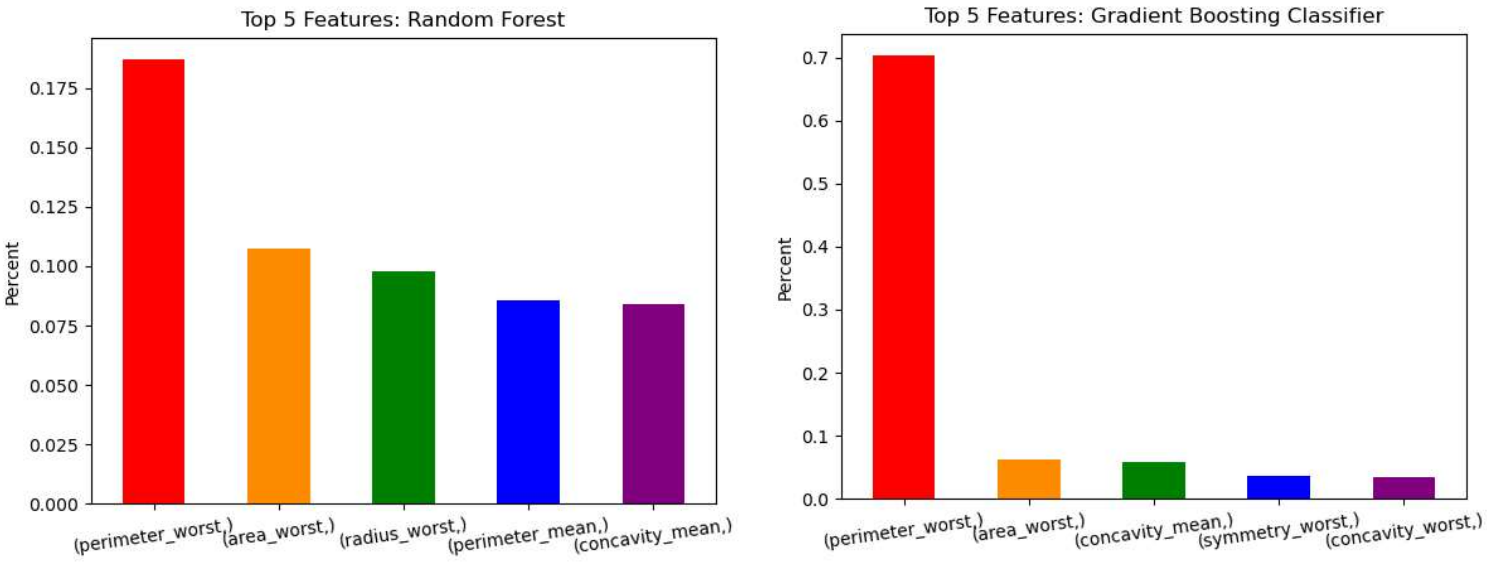
Figure 4: Comparison of machine learning model ROC curve and AUC



Our team also wanted to determine what features from the data set had the greatest impact on the model's predictions. If we can determine that the model performs very well on a smaller subset of variables, then it should reduce the amount of data collection needed to produce a valuable result. The Random Forest and Gradient Boosting Classifier models both contain functions to plot their top five features. The rank is determined by the percentage of added mean square error in the prediction by the variable being omitted. Mean square error is the average squared difference between the predicted outcome and the actual outcome. Both rankings are

listed in the following graphs for comparison, where the Y axis represents the percentage of importance used in the prediction.

Figure 5: Illustration of top five features ranked by importance



In both the Random Forest and Gradient Boosting Classifier models, the most important predictor variable was the worst perimeter. This is the mean of the perimeter of the largest three cell nuclei for each image. For both models the area worst is the second most predictor. Interestingly, the Random Forest allocates much higher weights to the other four features than the Gradient Boosting Classifier, meaning that the GBC may be a more efficient model when only using the perimeter worst feature.

VI. Overall Results from Data Analysis

In our statistical analysis our violin plots showed very distinct differences in the means of the measurements between benign and malignant tumors. The performance of each of our machine learning models was excellent. They predicted the status of the patient's tumor with a

high degree of accuracy. The best model was the Random Forest, though all of them performed considerably well. The results were not perfect as some patients were wrongly predicted as having a benign tumor when the true status was malignant, but we believe the model is a highly effective tool for diagnosing breast cancer and will be effective for future patients if proper samples and measurements are obtained from the tumor tissue as they were with this data set.

VII. Conclusion and Recommendations

We believe the collected data was very accurate and relevant to the final diagnosis, and the results of the machine learning models support this conclusion. The most important part of performing research is obtaining good data, and the researchers at the University of Wisconsin Departments of Computer Science and Surgery collected very relevant and valuable data for the purpose of training a machine learning model. It appears that malignant cells generally have larger nuclei, which sets them apart from benign cells. In the future we believe the study should be repeated for other kinds of tumors that can be sampled using Fine Needle Aspiration biopsies, to train similar machine learning models to identify cancer. The applications for machine learning models are extensive, and it appears they have valuable application for diagnosing malignant tumors.

References

Chaurasia, V., & Pal, S. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3139141

Choe, R., Konecky, S. D., Corlu, A., Lee, K., Durduran, T., Busch, D. R., Pathak, S., Czerniecki, B. J., Tchou, J., Fraker, D. L., DeMichele, A., Chance, B., Arridge, S. R., Schweiger, M., Culver, J. P., Schnall, M. D., Putt, M. E., Rosen, M. A., & Yodh, A. G. (2009). Differentiation of benign and malignant breast tumors by in-vivo three-dimensional parallel-plate diffuse optical tomography. *Journal of Biomedical Optics*, *14*(2). https://doi.org/10.1117/1.3103325

Huang, J.-S., Pan, H.-B., Yang, T.-L., Hung, B.-H., Chiang, C.-L., Tsai, M.-Y., & Chou, C.-P. (2020). Kinetic patterns of benign and malignant breast lesions on contrast enhanced digital mammogram. *PLOS ONE*, *15*(9). https://doi.org/10.1371/journal.pone.0239271

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010). Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research, 2,* 109–126. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3255438/

Turkki, R., Byckhov, D., Lundin, M., Isola, J., Nordling, S., Kovanen, P. E., Verrill, C., von Smitten, K., Joensuu, H., Lundin, J., & Linder, N. (2019). Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Research and Treatment*, *177*(1), 41–52. https://doi.org/10.1007/s10549-019-05281-1

Wolberg, W. H. (1995). Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. *Archives of Surgery*, *130*(5), 511. https://doi.org/10.1001/archsurg.1995.01430050061010