Subject: Understanding Cancer with Metastasis Data
Date: April 28, 2022

Introduction and Research Question

Cancer: the #2 cause of death in the United States among adults. Over a half-million Americans die from cancer every year, and around 40% of people develop cancer at some point in their lives (cancer.gov). One sentence that no cancer patient wants to hear is: "Your cancer is metastatic." Metastasis means the cancer has spread from its origin to other parts of the body, via the bloodstream or lymph nodes. One specifically bleak classification, distant metastasis, means cancer cells have traveled through the body to distant organs, such as from the liver to the brain. Oftentimes these cases are fatal. It is vital for patients and medical professionals to know whether one's cancer is metastatic, so that they may rapidly deploy the necessary treatments to mitigate the spread.

But while it is crucial for treatment purposes to test for metastasis, does these test results give any clearer insights into the outcomes of cancer patient? Studies have yet to give conclusive answers one way or the other. This investigation aims to determine whether the outcome of cancer cases is more predictable once the cancer is classified with or without distant metastasis. The outcome of this research will help cancer patients understand the clarity of their future once they undergo metastasis testing. It will also assist medical providers in making more accurate decisions on when to rule a cancer case as terminally ill or curable.

Research Population

The aim of this research is to investigate our understanding of cancer case outcomes based on our knowledge of the cancer being metastatic or not. Therefore, this research has two populations of interest that coincide with one another: cancer patients who do not yet know if their cancer is metastatic, and cancer patients who know whether their cancer is metastatic. We seek to compare the predictability of the case outcomes for each population.

Data

This study incorporates three datasets available to the public on Kaggle that are all derived from the same sample. The first dataset consists of observations of esophageal cancer patients around the world, including their demographics, health habits, family history and case information. Notably, this data includes information on whether the cancer is ruled as distant metastasis and the patient's survival or decease several years after diagnosis. The second and third datasets contains the same information, but for colorectal and prostate cancer, respectively. Esophageal, colorectal and prostate cancer are all in the top seven cancers with the highest fatalities nationally (cancer.gov). Therefore, these datasets provide cancer data that represent a large number and sufficiently diverse range of patients. The observations have remained anonymous from the beginning of the investigation to avoid privacy concerns.

The datasets had sporadic missing values in various categories. These values are imputed via the Miss Forest method in R Studio. This method predicts the missing value based on similar observations. Missing values are imputed to preserve the number of observations used for analysis.

One aspect to note is that the cancer cases in the three datasets are from 2010 or later, so the modern treatment methods used for these cases closely reflect today's survival patterns for cancer patients.

Research Design

The first action for the research process is to merge the three datasets. Necessary changes are made to transform categorical variables into binary variables. To determine the change in predictability of cancer case outcomes once the metastasis classification is known, the data analysis will be split into two cases. The first case involves predicting the outcome of cancer cases based on all available case information minus the distant metastasis indicator variable (feature). The second case involves the same method and includes the distant metastasis feature. The accuracy of the two predictions will be compared to investigate whether the distant metastasis feature significantly changes the predictability of cancer case outcomes.

The data came from expansive observations that include dozens of variables about the patients. Therefore, the only risk of confounding variables in the predictions is the treatment method. The data source does not indicate whether the treatment of metastatic versus non-metastatic patients changed drastically. However, this should not be a large concern. Since the majority of cancer cases are treated with chemotherapy, there are not expected to be different treatment methods that may impact other features in the data.

Data cleaning and imputing missing values are performed in R Studio, while the predictive analysis is performed in Python. All programs and the datasets are provided for your reference. As a note, the workflow of the analysis progresses from DataCleaning.R to ResolvingNAs.R to PatientPredictions.py.

Analytic Technique

The analytic technique of this research involves an integrative approach of prediction and explanation. Supervised machine learning is deployed to classify case outcomes as survival or death based on over a dozen features. Since the data does not include an excessive number of observations, computing time and cost is not a grave consideration. Therefore, the analysis uses the ensemble method comprised of the Gaussian Naïve Bayes, Random Forest and X-Gradient Boosting machine learning models to achieve the highest possible accuracy.

Two ensemble methods are built from the three above models. These models are chosen based on the data's number of observations and features. Accuracy scores and feature importance are all used to investigate the effect of the distant metastasis feature on the ensemble classifier's performance.

Results

The ensemble methods of classification with Gaussian Naïve Bayes, Random Forest and X-Gradient Boosting models are fitted on the training dataset. Once the two ensemble methods are established, the first step is to compare the performance of the classifiers on the testing data.

Figures 1 and 2 provide the confusion matrices for the ensemble classifiers.

Figure 1. Confusion Matrix of Ensemble Classifier without Metastasis Feature

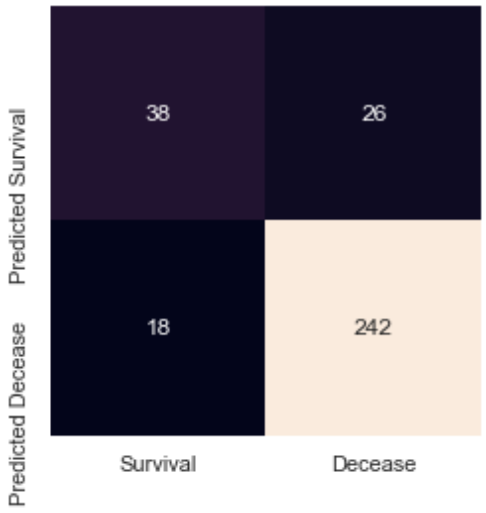


Figure 2. Confusion Matrix of Ensemble Classifier with Metastasis Feature

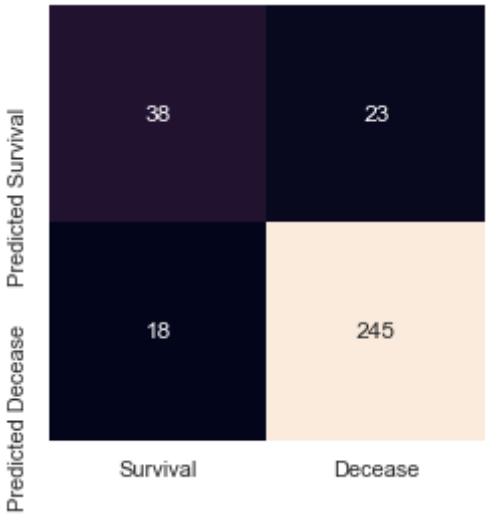


Table 1 provides the test statistics for the ensemble classifiers with and without distant metastasis as a feature, derived from the data in Figures 1 and 2.

Table 1. Results of Ensemble Classifiers

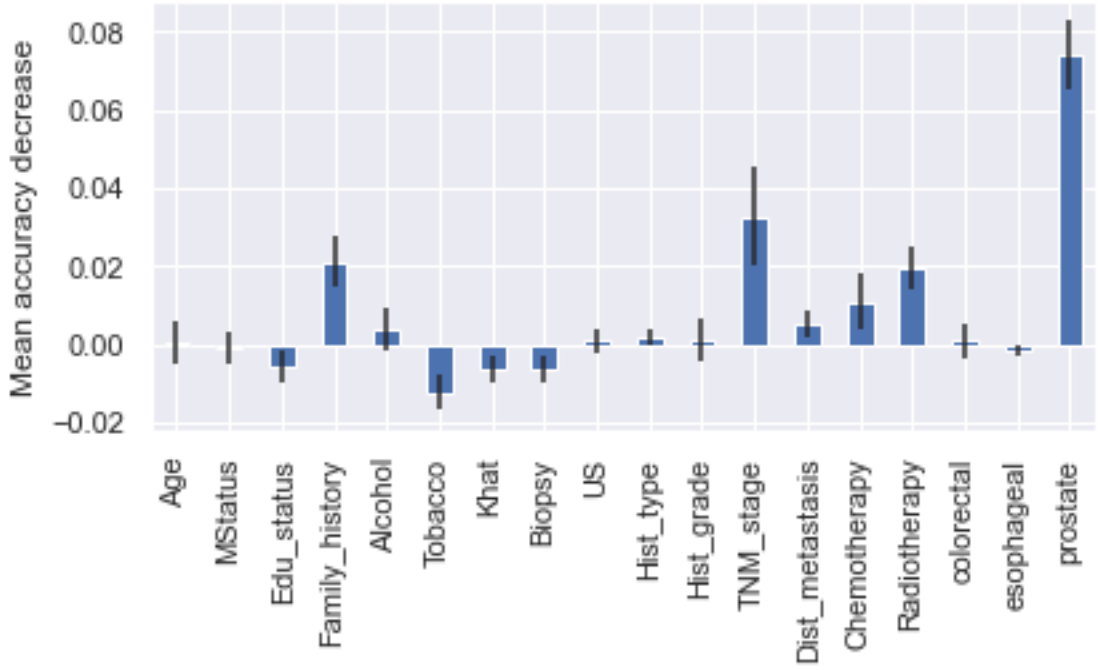
	Classifier without Distant Metastasis Feature	**Classifier with Distant Metastasis Feature**
Accuracy Score	0.864	0.873
F1 Score	0.9167	0.9228
Sensitivity/Recall	0.9030	0.9142
Specificity	0.6786	0.6786
Positive Predictive Value/ Precision	0.9308	0.9316
Negative Predictive Value	0.5938	0.6229

The classifier with the distant metastasis feature has a slightly higher accuracy score, f1 score, precision and recall. For overall accuracy, the classifier including distant metastasis correctly predicted a higher percentage of case outcomes than the classifier excluding distant metastasis. While the accuracy score is the main measurement to compare the two classifiers, it is important to consider the sensitivity and specificity. In this scenario, the sensitivity (also called recall) refers to the ability of the classifier to correctly identify cancer patients who will not survive. Specificity is the classifier's ability to correctly identify patients who will survive their cancer battles. Ethically, we look for the classifier with the highest specificity. This limits the number of situations in which patients are told they will not survive when they actually will survive. These false positives cause immense unnecessary distress for patients and family members. The ensemble classifier with the distant metastasis feature has the higher sensitivity of the two classifiers, meaning it falsely classifies future survivors as deceased less often than the other classifier. This metric, combined with a higher overall accuracy score, f1 score and precision, gives the classifier with the distant metastasis feature the advantage over the classifier without the feature. In applicable terms, we initially see that knowing whether distant metastasis is present in a cancer case increases our ability to accurately predict the case's outcome.

However, we must consider possible random variation accounting for the increase in classification accuracy. Perhaps the increased accuracy from one classifier to the next is significant, but it could also be due to the random error of the testing dataset. We conduct a hypothesis test to determine whether the difference in the accuracy of the classifiers is statistically significant. Each prediction in the classifier with the distant metastasis feature is treated as a trial of a binomial random variable with probability of success equal to the accuracy of the classifier without the distant metastasis feature ($p_{success} = 0.864$). Then, the binomial random variable is replicated over 324 trials, which equals the number of samples in the testing dataset. The probability of observing a total success rate of 0.873 or greater, the accuracy of the classifier with the distant metastasis feature, over 324 trials is calculated. This method is used since we are trying to determine if the success rate (accuracy score) of the binary classifier with the distant metastasis feature is significantly higher than the success rate of the binary classifier without such feature. The calculation yields a p-value of 0.349, which is above any reasonable significance level of a hypothesis test. Therefore, we do not have enough evidence to conclude that the distant metastasis feature significantly increases the ensemble classifier's accuracy.

One other method for checking the impact of the distant metastasis feature on the classifier's accuracy is feature importance. While feature importance cannot be calculated for the ensemble classifier, it is calculated for the random forest model, which gives us an idea of the impact of the metastasis feature. Figure 3 shows the importance of each feature on the random forest model with the distant metastasis feature.

Figure 3. Feature Importance of Random Forest Model Using Permutation



The feature for distant metastasis, named Dist_metastasis, represents less than a 0.01-point accuracy decrease when excluded in the random forest model. Therefore, the inclusion of the feature accounts for a <0.01-point increase of the model's accuracy. While the feature does technically improve the accuracy of the random forest model, it is a rather trivial increase. This agrees with the results from the hypothesis test, which also shows a statistically insignificant increase in the accuracy of the ensemble classifier.

Finally, when comparing the classifier performances, one may consider the threat of model overfitting. The classifiers can be evaluated for overfitting by comparing their accuracy scores on the testing versus training datasets. For the classifier without distant metastasis as a feature, the testing accuracy is 0.864, while the training accuracy is 0.994. For the classifier with the distant metastasis feature, the testing accuracy is 0.873 versus a training accuracy of 0.994. For both classifiers, the training accuracy is significantly higher, signaling an overfitting of at least one of the Naïve Bayes, Random Forest or X-Gradient Boosting models. However, overfitting is not of grave concern for this investigation. This study solely focuses on comparing two classifiers, rather than applying the classifiers to specific future predictions. So, while there is indication of overfitting for certain models, it does not affect the validity of the study.

In summary, while the addition of the distant metastasis feature increases the accuracy of the case outcome predictions, it is not a statistically significant increase. Therefore, there is not enough evidence to show that knowing whether a cancer case is distantly metastatic improves our insight into the case's outcome.

Ethical Considerations

The investigation determines that we do not have enough evidence to conclude that the distant metastasis feature significantly increases the classifier's accuracy. Therefore, the results of testing for distant metastasis do not give us any clearer insight into the probable outcome of cases. This finding has a couple ethical considerations when interpreting results.

First, the study does not compare the survival rates of distantly metastatic versus non-metastatic cancer cases, and its aim is not to predict an individual patient's case outcome. Therefore, medical professionals must not use this study for any predictive inferences. It must only be used to conclude that distant metastasis information does not significantly increase the predictability of cancer cases.

Second, there is a tradeoff in this study's result. The ultimate takeaway from this study is that medical professionals must exercise ultimate caution when ruling cancer patients as terminally ill based solely on the presence of distant metastasis. This could lead to medical professionals being more reluctant to rule a cancer patient as terminally ill. The tradeoff that arises with the study's conclusion is that it could provide more confusion for cancer patients with distant metastasis. The study essentially shows that predicted outcomes of cancer cases should not hinge on whether they are distantly metastatic. Therefore, the study may lead to less predictions by medical professionals, leaving patients in the dark on their likely case outcomes.

Works Cited

"Cancer Statistics." National Cancer Institute. National Institute of Health, September 25, 2020. https://www.cancer.gov/about-cancer/understanding/statistics

"Survival Patterns of Cancer." Saurabh Shahane. Kaggle Inc, 2021. https://www.kaggle.com/datasets/saurabhshahane/survival-patterns-of-cancers