Theran Meadows

DSC 550 Final Project

November 18, 2023

Prostate Cancer Predictions

Introduction

The problem that I would like to address is prostate cancer. From the Center of Disease Control, about 13 men out of every 100 will develop prostate cancer in their lifetime. I personally do not have prostate cancer but my father was diagnosed some years ago with it. He has since been testing negative for prostate cancer ever since. Nonetheless, With age being the most common risk factor and having a family member be diagnosed with it does give an increase to being diagnosed with it yourself in the future.

I would like to do some simple correlation graphs to see if there are any factors that correlate together that might be worth putting more research into. With any correlating features, I would like to make a prediction model, train it with some of the data, and then test it with a larger portion of the data.

The data set contains 100 observations and 10 features. Features range from "radius", "texture", "perimeter" and other size measures to one feature that is categorical for a diagnosis result whether a patient result was "B" for Benign or "M" for Malignant that I plan on using to test the model's accuracy. I know I will have to change this categorical data set to binary so that I can do a logistic regression model.

The way I would pitch this project to stakeholders is to share with them is using the above statistics. Perhaps one of them could relate directly or indirectly knowing someone who has or has had prostate cancer. This is a study that would deal heavily in medical data. Normally I would like to get it from a hospital or an official medical database. In this project, I got the data from Kaggle.

In conclusion, I hope that by exploring this data set along with creating a prediction model can shed some light on possible prediction indicators whether or not a man has prostate cancer.

Exploratory Data Analysis

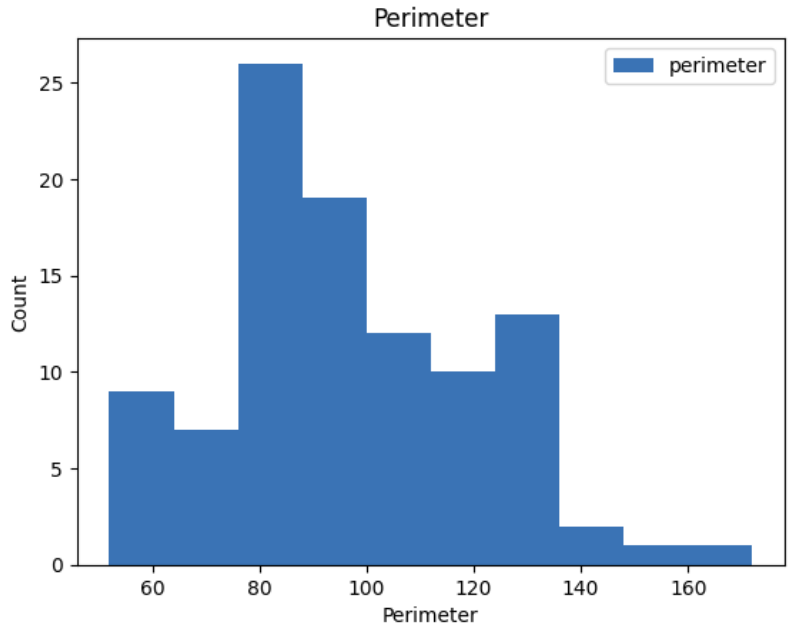
Here is a preview of the data:

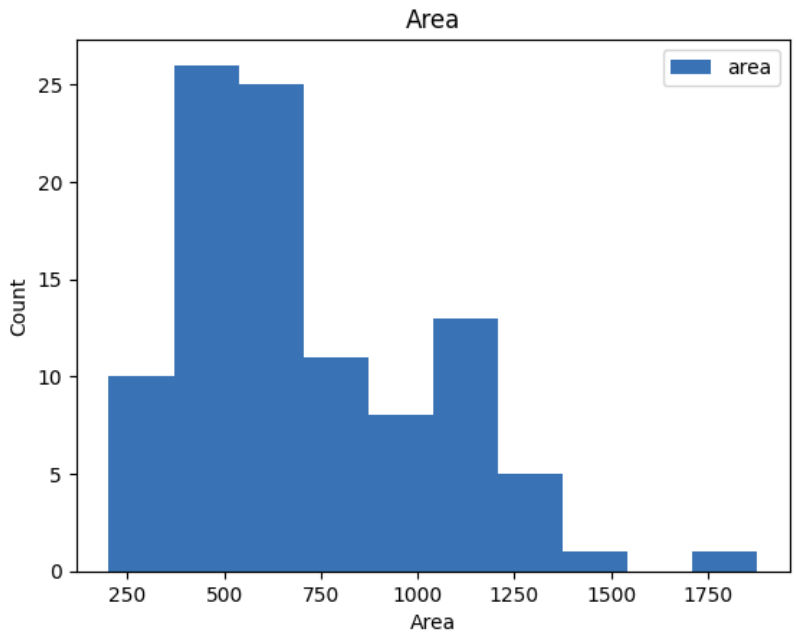
ID	Diagnosis Result	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Symmetry	Fractal Dimension
1	M	23	12	151	954	0.143	0.278	0.242	0.079
2	B	9	13	131	1326	0.143	0.079	0.181	0.057
3	M	21	27	130	1203	0.125	0.160	0.207	0.060
4	M	14	16	78	386	0.070	0.284	0.260	0.097
5	M	9	19	135	1297	0.141	0.133	0.181	0.059

Using a correlation matrix with all the variables, I found that there were some variables that had stronger correlations between them than others. Ultimately, I was trying to find out what variables would be good indicators of a “B” or Benign diagnosis result.

For diagnosis result, perimeter and area had the highest correlations with perimeter being 0.60 and are having 0.56. Even though those are the highest of this data set, perimeter and area only have a moderate correlation strength.

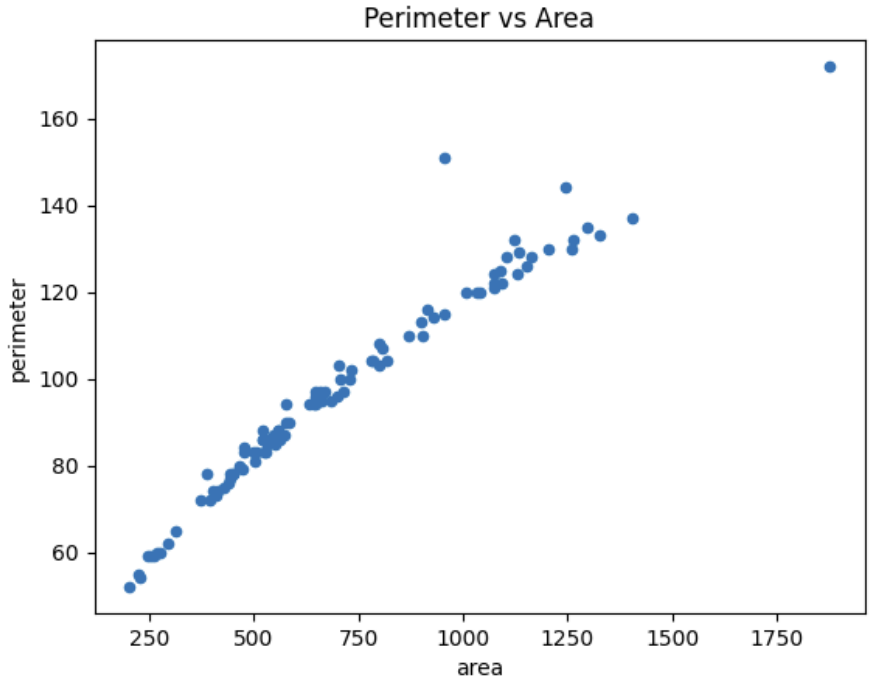
Below is a histogram of both of these variables.

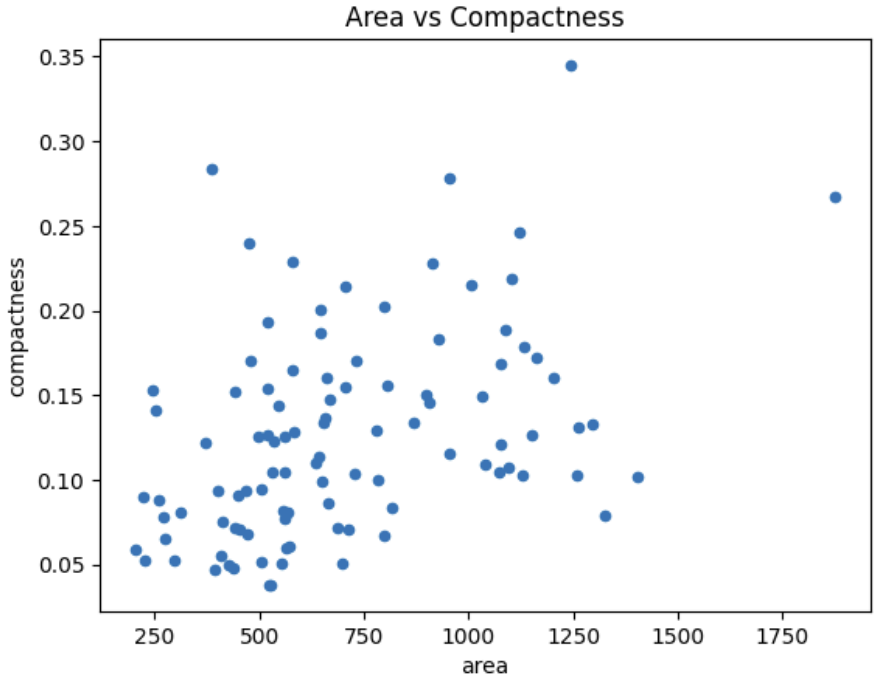




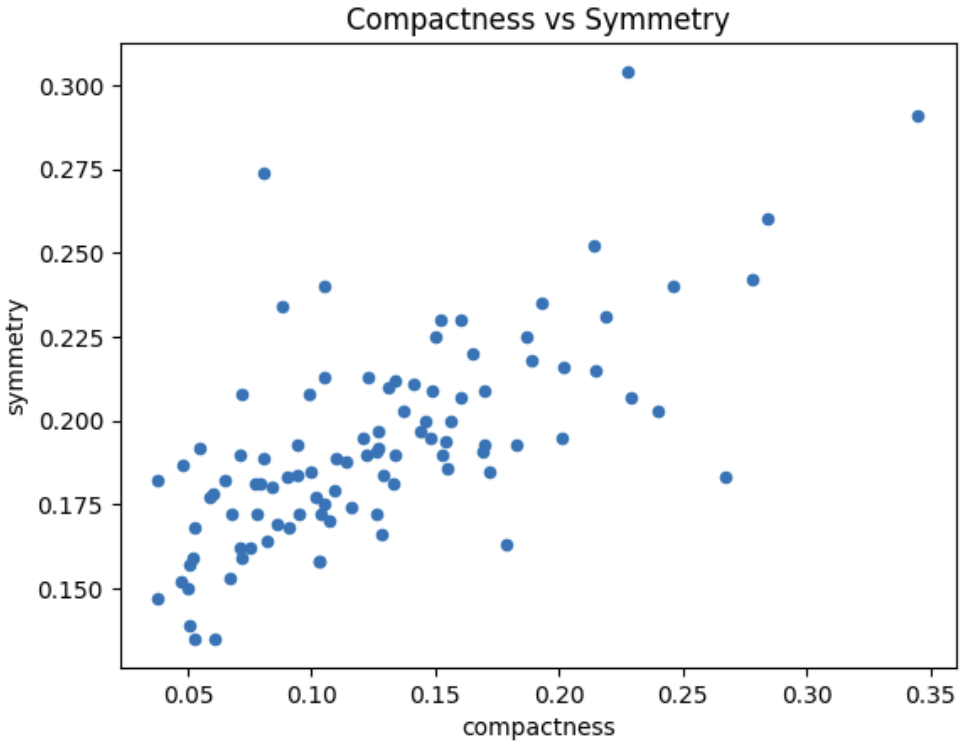
Both of the distributions do have similarities. There is a hint of a positive skew for both graphs.

These next scatter plots show the correlation strengths of the variables in between each other. You can see that other than perimeter and area, the correlation strengths really took a nosedive.





Almost no correlation is seen in area v compactness and a weak correlation in compactness and symmetry.



Data Preparation

During this phase, there was little to be done with the data to have it be cleaned before modeling. The ‘diagnosis_result’ column was changed to having boolean values instead. The data was then split into an 80:20 ratio for testing and training.

Model Building and Evaluation

I selected a linear regression model since my goal was to see patterns that could possibly predict if a patient was going to be diagnosed with prostate cancer. This did not turn out

how I was hoping as there was a high variance in the r-squared metrics ranging from 0.01 t0 0.60. This means there were low to moderate results that could be predicted by the variables.

I then tried a polynomial regression model which may have been impacted by overfitting the data. The training set did too well and the test set performed horribly.

Conclusion

In conclusion, these models have not provided me with adequate evidence to suggest that prostate cancer could be predicted based on the variables tested above. The correlations to begin with were not as strong as they could have been suggesting this outcome was likely. I would not recommend deploying this model and instead recommend more variables be measured, recorded, and tested. Potential challenges could be gathering enough data on the subject from various locations in the US. Because of that, there could be other factors not accounted for that could affect the results again.