***Male Testosterone Machine Learning Classification Modeling***

**William Justice   
Summer 2021 DSC680-T301   
https://github.com/WilliamJustice/Mlae-Testosterone-Machine-Learning-Classification-Modeling**

**Which Domain?**

This project is centered on the medical industry, specifically in the endocrinology field. Over the last few decades males in the United States and across the globe have seen a significant decrease in testosterone levels. In fact, within the last decade the statistical normal range of testosterone in the blood has been lowered by the medical field in order to meet the recent population’s testosterone levels despite the range being significantly lower than just decades past. Average testosterone levels one hundred years ago were 800-2000ng/DL and the current range is set for 270-700ng/DL. The problem with this is the negative health effects correlated and caused by low testosterone levels to include weight increase, depression, reduced muscle mass, and most importantly heart disease.

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5083173/
2. https://www.usatoday.com/story/news/2021/02/27/falling-sperm-counts-threaten-humanity-chemicals-blame-book-says/6842950002/
3. https://www.wtsp.com/article/news/health/male-fertility-rate-sperm-count-falling/67-9f65ab4c-5e55-46d3-8aea-1843a227d848
4. https://www.mayoclinic.org/healthy-lifestyle/sexual-health/expert-answers/testosterone-level/faq-20089016
5. https://www.medicalnewstoday.com/articles/323085
6. https://www.medicalnewstoday.com/articles/321564
7. https://www.urologytimes.com/view/testosterone-levels-show-steady-decrease-among-young-us-men
8. https://www.healthline.com/health/side-effects-of-low-testosterone
9. https://www.anoptimalyou.com/what-is-a-normal-testosterone-level-for-a-man/#:~:text=Average%20testosterone%20levels%20have%20decreased,range%20was%20800%20to%202000!
10. https://www.sciencedirect.com/science/article/pii/S2352914821000289

**Which Data?**

The data that will be utilized was gathered from a urology clinic in Feira de Santana, Brazil. The patients were between the ages of 40-85 and numbered 3397 patients. Features include age, diabetes, triglycerides, hypertension status, HDL level, waist circumference, and normal or low testosterone status. Thus, the features in this dataset are limited, but that only requires the right questions to be asked of the data.

**Research Questions? Benefits? Why analyze these data?**

The primary research question being proposed is “Is there potential for Machine Learning algorithm to produce a satisfactory model to instantiate a probability that a patient is to be diagnosed with a testosterone deficiency”?

It is important to remember that the goal here is not to diagnose a patient but merely to give a physician a glimpse into a possible diagnosis that would lower the costs associated with testing other possible diagnoses.

The benefit of this will be reduced costs to both the patient and hospital. The most important benefit is adding the ability to quickly and more accurately diagnose a patient which would decrease length of treatment. Another benefit of this is referral bonuses for directing patients to private practice endocrinologists.

**What Method?**

The first method will be to clean the dataset. Part one of the first method will be to create headers that are more interpretable. Although the current dataset variable names may be readable to an endocrinologist, they need renamed for the average audience member.

The second part of the first method will be to add a binary classifier to the classification variables. For example, the variables for diabetes, hypertension, and testosterone levels are all True or False. These will need to be altered to a binary for processing.

The second method will be to perform exploratory data analysis. The third method will be to perform feature reduction. Since there are only 6 features and one target, it is not anticipated that any features will be deemed insignificant, but it is possible.

The third step will be to begin feeding the data into a few different machine learning algorithms to measure initial performance. It is still undetermined which algorithms will be used as it must be researched more to properly fit the slightly imbalanced dataset.

The fourth step will be to take the two- three highest initial performing datasets and begin hyperparameter tuning for optimization.

It is anticipated that the performance metric used for model evaluation will be the f1 score because since the dataset is imbalanced there is a higher likelihood that recall will be very low and precision could be high. Due to this, the F1 score which attempts to balance precision and recall will be the safest bet in measuring the outcome.

**Potential Issues?**

At this time I see no potential issues in the data or the process. Of course there is always potential for more data that could make the outcome better, but I think the data has the ability to answer the overarching hypothesis. I believe this is a pretty straight-forward data science challenge that has is interesting to observe.

**Concluding Remarks**

My primary concern is that this problem does not offer a “business” solution. It could be beneficial to feed this into a data pipeline for physicians to utilize, but would not have much impact on revenue. Yes, it would make for a quicker diagnosis, but for a minimal problem. More importantly a minimal problem that could be exploited for more testing and lab work on other possible diagnosis’s that require more in-depth and costly testing. I think this problem will still show a great approach to an interesting problem, but may not provide a revenue increasing solution.