Examining Fairness and Transparency in Medical Automated Decision Systems

Pedro Galarza and Jack Epstein

Background: general information about your chosen ADS

The advent of large scale data collection and storage has given way to a new era in information that is revolutionizing a variety of fields. Due to the sheer volume and availability of data many organizations seeking quantitative solutions have begun employing Automated Decision Systems, or ADS's. There is an obvious intersection between data science and the healthcare industry, and ADS's are already being leveraged to predict diseases, make public health decisions, assess patient risk, allocate specialized care, and much more. While the promises of AI in medicine are exciting, there are numerous risks emerging from these systems. Problems of bias, fairness, equitability and privacy already rampant within our healthcare system, have the potential to be amplified by ADS systems capable of making high-impact decisions about individuals' health at scale [1,2].

One particular application of ADSs that have received much attention and research are machine learning models for the diagnosis of cardiovascular diseases. Cardiovascular diseases are the leading cause of mortality around the globe according to the WHO, with coronary artery disease representing a large portion. Methods for diagnosing coronary artery disease can often be invasive and risky. ADS systems have the potential to provide an accurate, accessible, and non-invasive diagnostic method for CAD. There is a rich literature exploring machine learning techniques that have been developed around data sets for exactly this purpose [3,4].

One of the foundational data sets on which many early CAD machine learning diagnostic systems implemented is the UCI heart disease data set released with the purpose for pioneering predictive models [5]. The data set has 303 patients with 76 associated features, with the "goal" feature being a binary indicator of heart disease. Only 14 medically relevant features have been released, the redacted features are mostly medical attributes less relevant to CAD, but also include identifying information like social security numbers and names. One feature that is conspicuously missing from both the original data set and the redacted on is race. Futhermore the attribute "sex" present in released data set which exposes a natural tension between a feature's utility and its protected status.

In this paper, we seek to perform a case study of a publicly available prediction system called "12 ML Models + Visualization (92% Accuracy)" which employs an array of standard machine learning techniques to predict CAD using the UCI dataset much like an ADS would [6].

While perhaps not as sophisticated as some of the academic models, the notebook is implementable locally and allows for the exploration of the benefits and drawbacks of this diagnostic ADS systems. Furthermore, because the notebook uses the industry standard sci-kit-learn packages, our analysis seeks to show the advantages and drawbacks of these out of box methods which are impletmenets every day on similar problems.

As hinted, we intend to consider sex as a protected class and explore how men and women maybe treated differently by this classifier. There is an inherent tradeoff between accuracy and fairness that is especially contentious in healthcare domains. Sex as feature contains important and relevant medical information for an accurate diagnosis, however in the context of machine learning its status as a protected class may be overlooked and lead to classifieres with large accuracy gaps between the sexes. We're hoping that this ADS system will be a useful test case in exploring this tension and reveal some of the risks diagnostic ADS systems pose toward protected classes.

Furthermore, diagnostic ADS systems in practice are usually deployed under the guidance of medical professionals. For this reason it's not only important these ADS systems to be accurate, but also trustworthy and accessible to its users (doctors, nurses, etc). By implementing different transparency methods and metrics we hope to also further understand the relationship between complexity, fairness, interpretability, and accuracy.

Input and output

As mentioned above, the data for this ADS comes from the UCI Machine Learning Repository, where it undergoes most of the aggregation, anonymization and cleaning. Initially the data comes from hospital patients from a combination of the following: Hungarian Institute of Cardiology, University Hospital in Zurich, University Hospital in Basel and V.A. Medical Center in Cleveland [7]. The original dataset has 76 attributes including personally identifying information (PII), however this has all been removed before the ADS owner accesses this data.

The 13 remaining input features are listed in Table 1 below. All of the categorical features have been numericized and there are no missing values in the dataset once it reaches Kaggle.

Table 1

Feature Name	Description	Input Space	Mean	Range
age	Patient's age	R	54.37	29-77
sex	Patient's sex, M=1	{0,1}	0.68	0-1

cp	chest pain type (0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic)	{0,1,2,3}	0.97	0-3
trestbps	resting blood pressure (in mm Hg on admission to the hospital)	R	131.62	94-200
chol	serum cholesterol in mg/dl	R	246.26	126-564
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	{0,1}	0.15	0-1
restcg	resting electrocardiographic results: 0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy	{0,1,2}	0.53	0-2
thalac	maximum heart rate achieved	R	149.65	71-202
exang	exercise induced angina (1 = yes; 0 = no)	{0,1}	0.33	0-1
oldpeak	ST depression induced by exercise relative to rest	R	1.04	0.0-6.2
slope	the slope of the peak exercise ST segment. 0: upsloping, 1: flat, 2: downsloping	{0,1,2}	1.40	0-2

The target variable in this dataset is binary, with the positive class representing a diagnosis of angiographic disease. More specifically, someone is classified as positive for the disease if their artery diameter has narrowed by more than 50%. The ADS solely outputs the binary prediction {0,1}, however, the underlying code is capable of producing probabilistic outputs, which will help during the evaluation stage. In practice, when this ADS predicts a positive result, this is an indicator for someone to seek immediate medical help. In terms of distribution, this is a relatively balanced dataset, with 54% of the instances in the positive class. This base rate will change considerably when crosstabbed with certain target variables, as we will explain below.

Along with the issue of limited sample size, it is important to note that this dataset is not representative of the general population and while the output of the ADS is to be predictive for all people, the inputs are biased in many ways. First, the data used is of hospital patients meaning these are people who selected to get medical treatment when they felt they needed it. While this could potentially bias towards people who have access to care, the more obvious issue is that this biases the dataset towards people who either have heart problems or feel unwell enough to get treated for heart problems.

This self-selected group leads the data to skew in other, more tangible ways. As we can see in the initial plots [see Figure 3], this dataset skews much older than the average population. The median age is 55, while the global median age is 30 and the median age in the United States is 38, both far lower than our dataset [8]. Another clear sign this data does not represent the general population is the prevalence of chest pain. While a rough estimate, about a quarter of adults experience chest pain [9], which is far lower than the 92.5% experiencing some sort of pain in this dataset. This is unsurprising, as chest pain is likely one of the primary reasons why patients sought medical attention. Perhaps most important, given the context of our paper, is that this dataset is over two-thirds male. The US and global population is just about half male/female, so this doesn't represent the general population. Despite common misconceptions, heart disease is quite prevalent in women [10] and thus could put women at a disadvantage as they are severely underrepresented in this dataset.

We can view these distributions overlaid with the target variable to get additional key insights into this dataset. We can see in figure 2 that while the dataset skews more heavily male, this difference is much more extreme with the negative class distribution than with the positive class distribution. This is caused by differing base rates between the sexes, where men in this dataset have heart disease 45% of the time, compared to 75% for women. In the same figure, we see an interesting phenomenon with chest pain, where the positive cases over-index in having non-anginal chest pain, while the negative class over-index in typical angina. This may seem counterintuitive, but could be explained because people with chest pains are likely to seek medical help yet a typical angina is different from our target variable.

In FIgure 3, we see the pairwise correlations of all continuous features, overlaid with the target variable. While not a specific insight around correlation, we see another interesting phenomenon around the age distributions. While conventional wisdom would say that the positive class should skew older, we see the opposite -- the negative class clearly skews older, while the positive class is relatively more normally shaped. This is likely due to the fact that older people are more likely to suffer from heart disease so are also more likely to seek out preventative care.

Figure 2

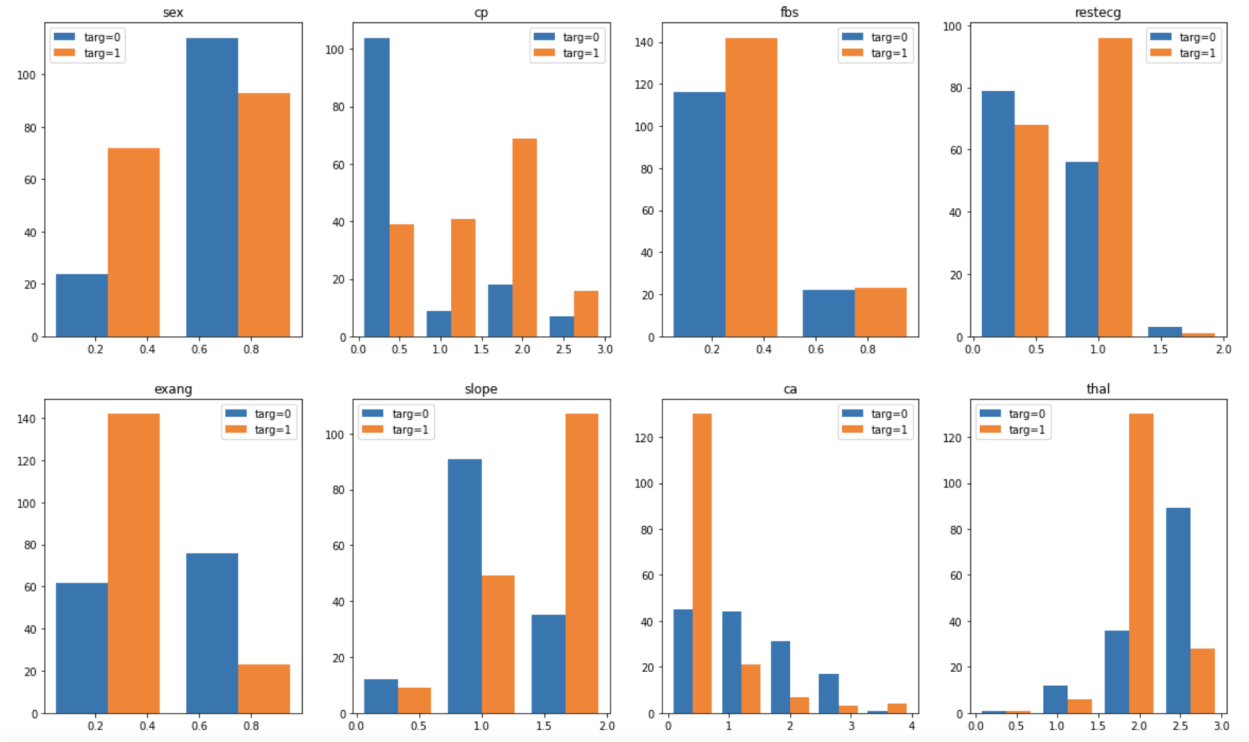
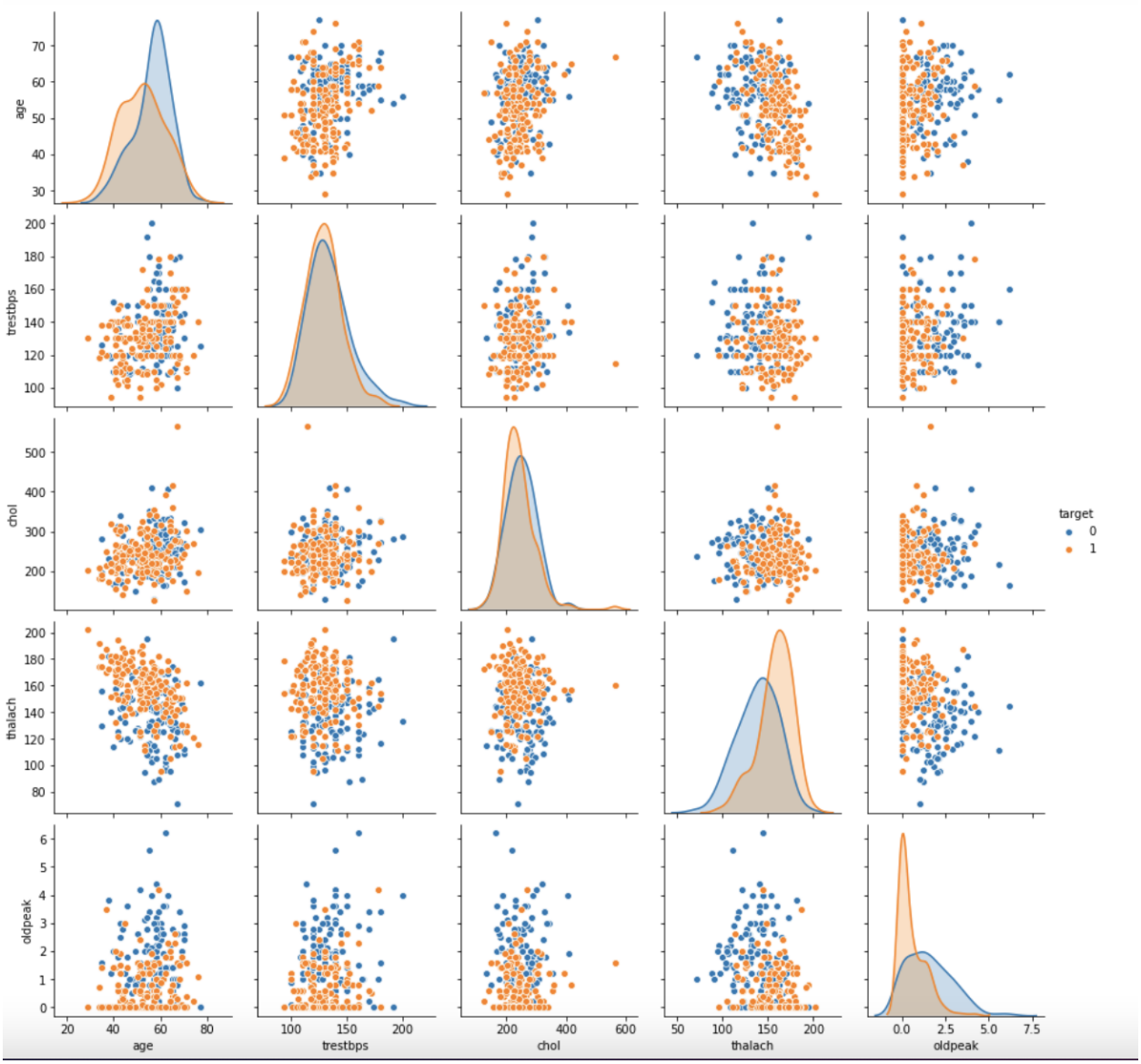

sex
cp
fbs
restecg
exang
slope
ca
thal
targ=0
targ=1

Figure 3



Implementation and validation: present your understanding of the code that implements the ADS. This code was implemented by others, not by you as part of this assignment. Your goal here is to demonstrate that you understand the implementation at a high level.

1. Describe data cleaning and any other pre-processing
    - The original UCI data set contains 76 attributes describing 303 participants, however only 14 features are released. The “goal” feature refers to the presence or absence of coronary artery disease. Missing values are marked with a value of -9.
    - Overall the released data set is very clean and ready for modeling, consequently the only pre-processing done in the ADS system is standard feature scaling.

2. Give high-level information about the implementation of the system
    - System is trained with standart train, validation, and test split. 12 different models of varying complexity are tuned for 0-1 loss. The best performing model is a neural network with the following metrics:
        i. Accuracy: 92
        ii. Precision: 1
        iii. Recall: .857
        iv. F1: .923
        v. Specificity: 1
    - This model uses the out of bag hyperparameters (more detail below)

3. How was the ADS validated? How do we know that it meets its stated goal(s)?
    - Standard train, validation, test splits
        i. It's important to note that there is no parameter tuning in the model selection process. He effectively treats both testing and validation data in the same way
    - He does some feature selection after the fact, but doesn't appear to incorporate this into his ADS
    - Accuracy is the main KPI -- he checks other metrics (recall/precision) but picks the winner based on accuracy.

Outcomes

1. Analyze the effectiveness (accuracy) of the ADS by comparing its performance across different subpopulations.
    a. As described in our Background section, we intend to explore how the model performs differently against male and female subpopulations.
2. Select one or several fairness or diversity measures, justify your choice of these measures for the ADS in question, and quantify the fairness or diversity of this ADS.
    a. Given that this is a diagnostic tool and the stakes are extremely high for false negatives (sensitivity of the model), we intend to measure model performance between sexes by comparing the false negative rate and recall. As a secondary concern we will also look at general accuracy and the false positive rate differences as well.
    b. After performing with metrics we will explore some baseline fairness pre/post processing techniques from the aif package depending on our disparity results.
3. Develop additional methods for analyzing ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME), or any other property that you believe is important to check for this ADS.
    a. There is a medical consensus that there does exist a correlation between biological sex and coronary artery disease, and thus model performance difference between genders may not entirely be due to machine learning bias. Thus we intend to use QII techniques to understand causal relationships between features and predictions and explore the role sex plays.

b. Furthermore, because this model is likely to be used by a medical professional whose expertise may complement the effectiveness of the model, we home to accompany the ADS with an explainable and interpretable companion.

Summary

4. Do you believe that the data was appropriate for this ADS?
    - The intention of the data set seems to allow for the exploration of baseline ML models for diagnostic application. Given how small the dataset already is, it gets even smaller when overlaying with features such as age or sex. This is most apparent in test data which is only 50 instances.
5. Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.
    - We intend to analyze and report on these model characteristics based on our results from the Outcomes section.
6. Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?
    - ADS systems of this sort have the potential to revolutionize medical treatment and accessibility and there is a societal benefit to building systems like these that can complement current diagnostic methods (in person diagnosis, lab testing, etc.)
    - Clearly, this implementation is extremely limited and should not be implemented on a large scale simply from the limited size of the data and features. Results from the Outcomes section should also point to evidence of other potential drawbacks concerning fairness, stability, and robustness.
7. What improvements do you recommend to the data collection, processing, or analysis methodology?
    - This will also be dependent on the outcomes section. The overall goal however is an ADS system that is accurate, stable, fair, and interpretable.