Binary physical activity does not improve predictions of heart disease

Audrey Krishnadasan and Clayton Greenberg

Abstract

In this paper, we tested the binary indicator of physical activity for its impact on heart disease prediction from machine learning algorithms. We try to determine whether or not a binary physical activity feature improves accuracy when predicting heart disease. When predicting heart disease using machine learning, datasets are used that have features that are often more broad; in this case, physical activity levels are as simple as yes or no. It is important to investigate its effectiveness in prediction of heart disease to see if improvements in data collection methods could be made. In order to investigate this proposal, we tested different variations of our dataset on 3 different machine learning algorithms: K-Nearest Neighbors, Random Forest, and XGBoost. In these experiments we computed accuracy metrics with and without the binary physical activity feature. We found that the binary physical activity feature did not improve accuracy when predicting heart disease; in fact, through all three models we found a decrease in accuracy when we added the binary physical activity feature. These test results negate the notion that the binary physical activity feature improves the accuracy of heart disease prediction and calls for a more robust method of measuring physical activity.

1. Introduction

In this paper, we look at how physical activity measured in the binary impacts our machine learning predictions of heart disease. Looking at this specific feature can help us refine our data collection methods and create more effective prediction models. If there was an effective relationship between heart disease prediction and the binary physical activity feature, we could continue using this method - but if there is not an effective relationship, it calls for a more refined method of measurement of physical activity. This paper works with a large CDC phone collection dataset of over 300,000 people with 18 general health features each. The data includes numerical, binary, and categorical. We used three machine learning algorithms in order to be thorough: K-Nearest Neighbors, Random Forest, and XGBoost. We performed ablation studies on our data with these models including calculation of performance metrics.

2. Background

There is an abundance of work done related to this topic and the questions that are brought up by this paper. Our novel contribution is the focus on physical activity as a predictive feature within the models. For instance, Mohan et. al (2019) included features that were numerical side effects of physical activity - such as the presence of exercise induced angina and ST depression - rather than a binary indication of physical activity that is used in this paper. These methods garnered impressive results, with their hybrid random forest with a linear component model achieving 88.7% accuracy. The limitation of using more scientific values as an indicator of physical activity is that they are not patient-friendly; in other words, potential heart disease patients would have no way of tracking these values for themselves without visiting their physicians.

In another paper by Meng et. al (2020), the defining feature and focus of the paper surrounded activity trackers, such as a Fitbit watch, to track daily activity levels. Although their approach had a major shortfall in relation to this specific paper, the activity tracking was monitored on patients who had already been diagnosed with heart disease - so it does not give much insight into the prediction side of this paper. Instead this work suggests a more innovative (yet expensive) method of measuring a patient's physical activity level, which could be implemented in the lives of patients without heart disease to further monitor their health.

3. Dataset

The dataset used in this paper is called "Personal Key Indicators of Heart Disease," sourced from the 2020 annual CDC survey data. The dataset is made up of a telephone survey of 319,795 people regarding their health status and includes 18 features/variables.

Feature Name	Feature Type	Description	Feature Name	Feature Type	Description
Heart Disease	Binary (Y/N)	Respondents that have ever reported having coronary heart disease (CHD)* or myocardial infarction (MI)	Kidney Disease	Binary (Y/N)	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
Smoking	Binary (Y/N)	Have you smoked at least 100 cigarettes in your entire life?	Skin Cancer	Binary (Y/N)	(Ever told) (you had) skin cancer?
Alcohol Drinking	Binary (Y/N)	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	BMI	Numerical	Body Mass Index
Stroke	Binary (Y/N)	(Ever told) (you had) a stroke?	Physical Health	Numerical	For how many days during the past 30 days was your physical health not good?
Difficulty Walking	Binary (Y/N)	Do you have serious difficulty walking or climbing stairs?	Mental Health	Numerical	For how many days during the past 30 days was your mental health not good?
Sex	Binary	Female/Male	Sleeptime	Numerical	On average, how many hours of sleep do you get in a 24-hour period?
Diabetic	Binary (Y/N)	(Ever told) (you had) diabetes?	Race	Categorical	Imputed race/ethnicity value
Physical Activity	Binary (Y/N)	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	General Health	Categorical	Self-rated grade of general health: 'Excellent', 'Very good', 'Good', 'Fair', 'Poor'
Asthma	Binary (Y/N)	(Ever told) (you had) asthma?	Age Category	Categorical	Categorical Age Categories (eg. below 20, 20-25, 25-30, ... , 75-80, above 80)

Fig 1: A table of all the features included in the dataset and their descriptions.

In order to normalize all of these features into forms that would be better recognizable by the machine learning models we had to reformat some of the features. The yes/no responses were turned to binary (yes - 1, no - 0), the Age Categories (ie 50-54, 70-74) were changed to a scale of 0-18, *'General Health'* strings (ie. "Excellent", "good", "poor") were changed to a scale of 1-5, and '*Sex*' was turned into a binary indicator (Female - 1, Male - 0).

About 91% of surveyed people in the dataset did not have heart disease (CDC, 2020). This caused some issues during our preliminary testing on the models; when we used the full dataset, the majority class baseline was very high at 91.4%, making further improvements very difficult. Due to the high majority class baseline of the original dataset, we decided to sample from the dataset, with 270,000 points per class. Therefore, the new dataset had a 50% majority class baseline - which allowed for greater improvement of accuracy through machine learning.

After sampling the new dataset, we experimented with the models and suspected that some of the features were not useful and/or degrading the accuracy of the prediction. Through ablation, we created a "partial" set including the following features: "GenHealth", "Asthma", "Sex", "SkinCancer", "Smoking", "KidneyDisease"," PhysicalHealth", "PhysicalActivity", "Diabetic", "Stroke", "DiffWalking", "Age"; the partial set excluded the following features: "SleepTime", "MentalHealth", "AlcoholDrinking", "BMI". Through testing we found that when the excluded features were not included, our accuracy improved on the model predictions.

Our research question focuses on how physical activity can be used in the predictions of heart disease; This dataset in particular is rather limiting in how it measures physical activity, in a binary yes/no of whether the patient has exercised in the last 30 days aside from their job. This metric is not very specific and does not give us the full picture of their physical activity. Due to this generality, we shifted our focus to whether it is beneficial to include the general binary physical activity feature in these heart disease prediction models. This could mean, as one option, does it improve accuracy. As another option, does it improve accuracy enough to offset the increase in model complexity.

4. Methodology/Models

We used three different machine learning algorithms: K Nearest Neighbors, Random Forest, XGBoost.

4.1 K-Nearest Neighbors

K Nearest Neighbors is an algorithm that predicts a feature of a particular datapoint based on the same feature from the unclassified point's k nearest neighbors. We used all the default values from the SKlearn API. The number of neighbors (k) default values is k = 5, so the algorithm looks at the 5 nearest data points to the prediction point to determine the value. The 'weights' value determines whether or not these neighboring data points are weighted in a certain fashion. This metric could be useful in some cases, but the default value and the one that we used was "uniform" - all of the neighboring data points were weighted and considered equally. The 'metric' value is used to calculate distances when looking at the nearest neighbor - the default value uses a standard Euclidean distance where p=2. In hindsight, due to our data having varying ranges of numerical values, using the default euclidean distances was probably not the most efficient method.

4.2 Random Forest

The Random Forest algorithm is a supervised algorithm made up of multiple decision trees. Rather than focusing on the "best" prediction feature, RF creates more steady accuracy by randomizing the features it focuses on. RF uses "bagging" in its process; all of the training subsets are treated equally, the model tries to pick out the best subsets for uses of prediction. Random Forest also prevents overfitting by creating random subsets of certain features on trees. Again, all of the default values were used in the parameters for the RF initialization. The parameter "n_estimators" is the number of trees that are used in the forest, 100 is the default value. Another important parameter is "max_features" which determines the number of features to consider when looking at the best split, we used the default value of "sqrt". It may have been a better choice to use the actual number of features in the dataset being tested as the max as feature selection was a key feature in this research paper.

4.3 XGBoost

XGBoost is considered the "strongest" machine learning algorithm out of the three, it stands for *Extreme Gradient Boosting.* While RF uses bagging in its algorithm, XGBoost uses boosting. Essentially, this model makes duplicates of the training subsets by improving on previous subsets until it settles on the most optimal version. This improvement method is largely the reason why XGBoost is considered the strongest algorithm between the three mentioned in this paper. When initializing XGBoost, we used all the default values for parameters. One of these parameters is "max_depth", which in default is set to 6. This is the maximum depth of a tree, increasing this value will make the model more complex and more likely to overfit.

4.4 Remaining Methodology

To split the data into train/testing portions, we used the standard sklearn train/test to split the data, which randomly and efficiently splits the data into training and testing subsets. After splitting the data we moved forward with prediction via the three different machine learning algorithms: K Nearest Neighbors, Random Forest, and XGBoost.

We then used the three models to predict the binary feature 'Heart Disease' on the test set. In order to provide a better baseline of our testing, we repeated this process for the dataset including all 18 original features, a dataset including partial features that included physical activity as a feature, and finally a dataset including the same partial features but excluded physical activity. This step is important as it helped us gain a greater understanding of what the dataset would predict without taking out some of the less important features. By including and excluding physical activity from two separate partial datasets, we would be able to confirm that the sole reason for any improvement or change in prediction would be due to the removal/addition of the binary physical activity feature.

5. Results and Discussion

Through our testing we found that the inclusion of physical activity as a binary indicator does not improve prediction of heart disease. There are several metrics we used to measure the differences in accuracy of prediction. It is important to note that these metrics are only relevant when considered together, as their measurements work inversely with each other. We calculated: Sensitivity, Specificity, F-1 Score, and Accuracy. We calculated all of these metrics in a standard manner for all three configurations of the data sets (full features, partial features w/physical activity, and partial features without physical activities) for all three machine learning algorithms.

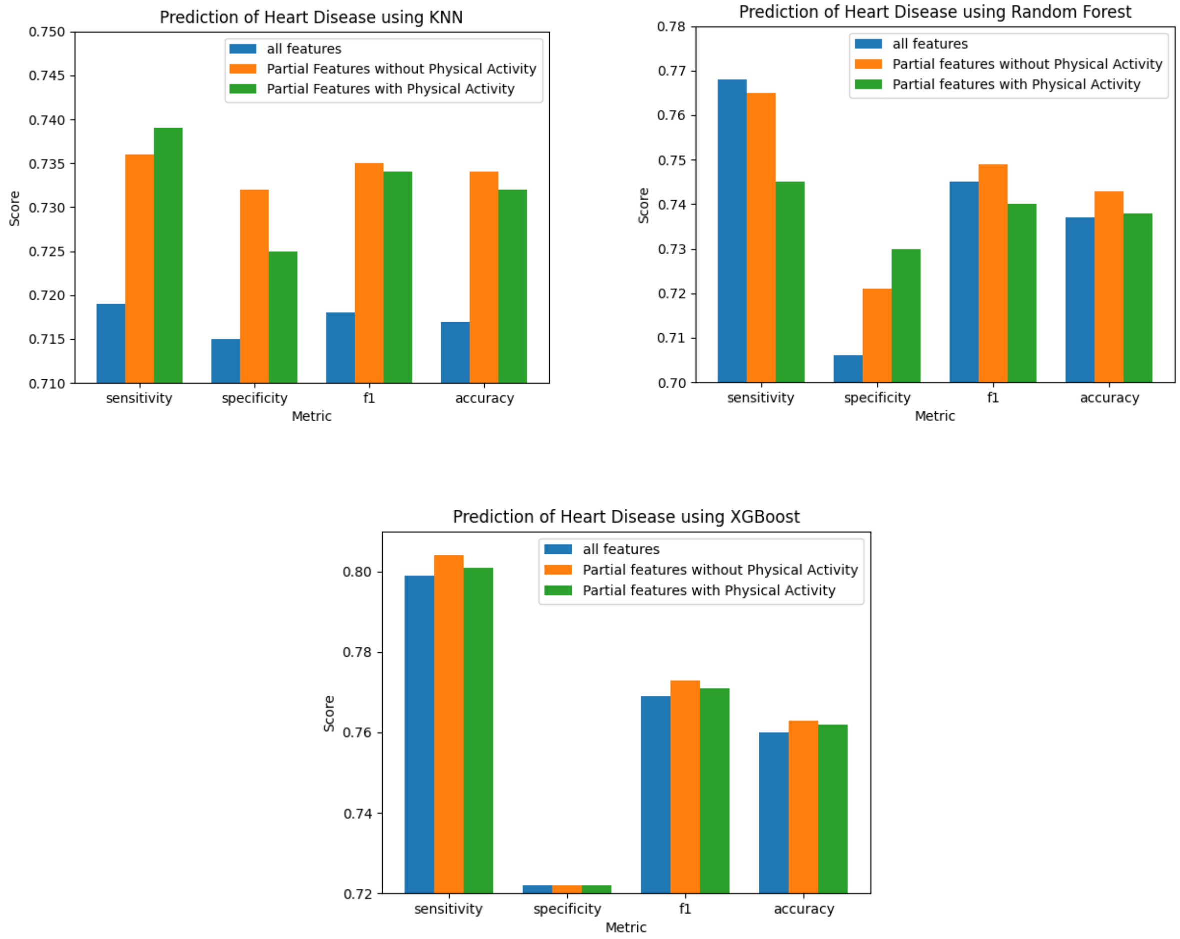

Fig 2-5: Each graph shows the metrics sensitivity, specificity, F1 score, and accuracy tested on three different versions of the same dataset. The varying factor between the three datasets were the included features - namely Physical Activity.

In the figures above, the main metrics are displayed for each machine learning algorithm and for each feature configuration. The model accuracy with the full features is shown in order to show a gauge of how much the model improved by removing a few features from the dataset. The partial dataset is a specific subset of the dataset in which certain features were removed in order to improve the performance of the model.

It can be seen that both the F-1 score and accuracy of all three machine learning models decreased from the partial dataset without physical activity as a feature to the partial dataset with physical activity. This shows that the use of physical activity as a binary feature does not improve the prediction of heart disease because these numbers are decreasing rather than increasing. Although the differences and decreases are quite small, it is clear that the binary physical activity indicator does not improve heart disease prediction.

6. Conclusion

In this paper, we investigated the effectiveness of a binary physical activity feature on machine learning predictions of heart disease. Our goals were to highlight a more user-friendly method of patient involvement in prevention and prediction of heart disease; if providers can predict heart disease with physical activity levels, patients could have a more active and measurable role in their health. We had to weight the dataset differently as we had a rather high majority class baseline in the original dataset, so in our testing set we weighted the data 50/50 (with and without heart disease). We utilized three different machine learning algorithms: Random Forest, K Nearest Neighbors, and XG Boost. We tested each model on three different datasets with particular features of the data removed in order to improve accuracy metrics. In the future, we could perform further hyperparameter tuning. When physical activity was included as a feature, we found a detrimental effect on our predictions of heart disease. Perhaps by including a more nuanced or fine-grained measure of physical activity as in Meng et al. (2020), we could see a significant improvement. If there was a more specific measure of physical activity, perhaps the feature could be more helpful in prediction of heart disease and could be a greater influence on patients' activity levels.

Acknowledgements

We would like to thank Inspirit AI for providing us with many resources in order to aid our research.

References

Centers for Disease Control and Prevention (CDC). 2020. Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Meng Y, Speier W, Shufelt C, Joung S, E Van Eyk J, Bairey Merz CN, Lopez M, Spiegel B, Arnold CW. 2020. A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients With Heart Disease Using Activity Tracker Data. IEEE Journal of Biomedical Health Information, 24(3):878-884.

Mohan S, Thirumalai C, Srivastava G. 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7;81542-81554.