A Heart Disease Multi-Dataset Analysis With Random Forest Classification and Reduced Dimensionality of the Most Positive Correlated Data

Gerald Rigdon
mail@rigdonhouse.com

Abstract

This paper explores whether better predictive machine learning models are achieved by removing redundant positive correlated data as applied to heart disease datasets. The goal is to identify the 3 most positive correlated dataset features from each dataset and proceed empirically by practice over theory to discover whether the accuracy, precision, and recall of a predictive model can be maintained or improved by retaining only 1 of these 3 features.

I. Introduction

Per the Centers for Disease Control (CDC) National Center for Health Statistics [1] heart disease related deaths have generally increased by 43% since 1950. Moreover, a 2017 report [2] revealed that deaths resulting from heart attacks exceeded 700,000 from the years 2010-2015. Given heart disease remains the leading cause of death in the U.S. [3], and that more than half of Americans are unaware of this fact [4], then ongoing analysis of existing data and demand for new datasets should be a top priority for exploration by data scientists using advanced machine learning tools. The benefits of this research may include better predictive models which in turn can be leveraged to discover solutions that can be employed to significantly raise awareness as well as reduce the risk of heart disease. Instead of comparing multiple methods of classification with various machine learning models, this analysis opted for two related datasets and a single machine learning technique known as Random Forrest Classification (RFC) and explores the more fundamental and yet least intuitive aspects of machine learning, namely, overfitting data.

II. Methodology

The analysis is based on two related heart disease datasets where one of the feature attributes is the target (is or is not heart disease). The first, DataSet1 [5], is comprised of 303 rows or instances and 14 columns or feature attributes. The second, DataSet2 [9], is comprised of 1190 rows or instances and 12 columns or feature attributes. DataSet1 (which combines data from Cleveland, Hungary, Switzerland, and Long Beach VA) is a feature superset of DataSet2, where DataSet2 aggregates much more data (from Statelog) to the original DataSet1 but with reduced features. RFC is employed on both datasets with a 70/30 data split where 70% of the data set is used to train the RFC model and a prediction of heart disease is established for the remaining 30% which serves as test data.

The initial prediction baseline comprises all 14 feature attributes for DataSet1 and 12 feature attributes for DataSet2 defined in Table 1 as follows:

Table 1

Feature **DataSet1**	**Feature** **DataSet2**	**Description**
age	age	The number in years
sex	sex	A binary conversion where: 0 = female 1 = make
cp	chest pain type	Chest pain category type
trestbsp	resting bp s	Resting blood pressure at hospital admission
chol	cholesterol	Cholesterol
restecg	resting ecg	Resting Electrocardiogram
fbs	fasting blood sugar	Fasting blood sugar
thalach	max heart rate	The maximum heart rate achieved
exang	exercise angina	Exercise induced angina. A binary conversion where: 0 = no 1 = yes
oldpeak	oldpeak	ST segment depression
slope	slope	The slope of the peak exercise ST segment
ca	does not exist	Number of vessels colored by fluoroscopy
thal	does not exist	Thalassemia category type
target	target	0 = normal 1 = heart disease

Then a correlation matrix is constructed to identify the 3 most positive correlated features for each dataset. Finally, the experiment is repeated multiple times for each dataset, where initially only the most positive correlated features are evaluated followed by iterations that retain only 1 of these 3 most positive correlated features along with the remaining features. In each experiment we present the following metrics [7] as follows where:

TP = True Positive (Predicted positive and is positive)
TN =True Negative (Predicated negative and is negative)
FP = False Positive (Predicted positive but is negative)
FN = False Negative (Predicted negative but is positive)

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP+TN}{TP+TN+FP+FN}$$

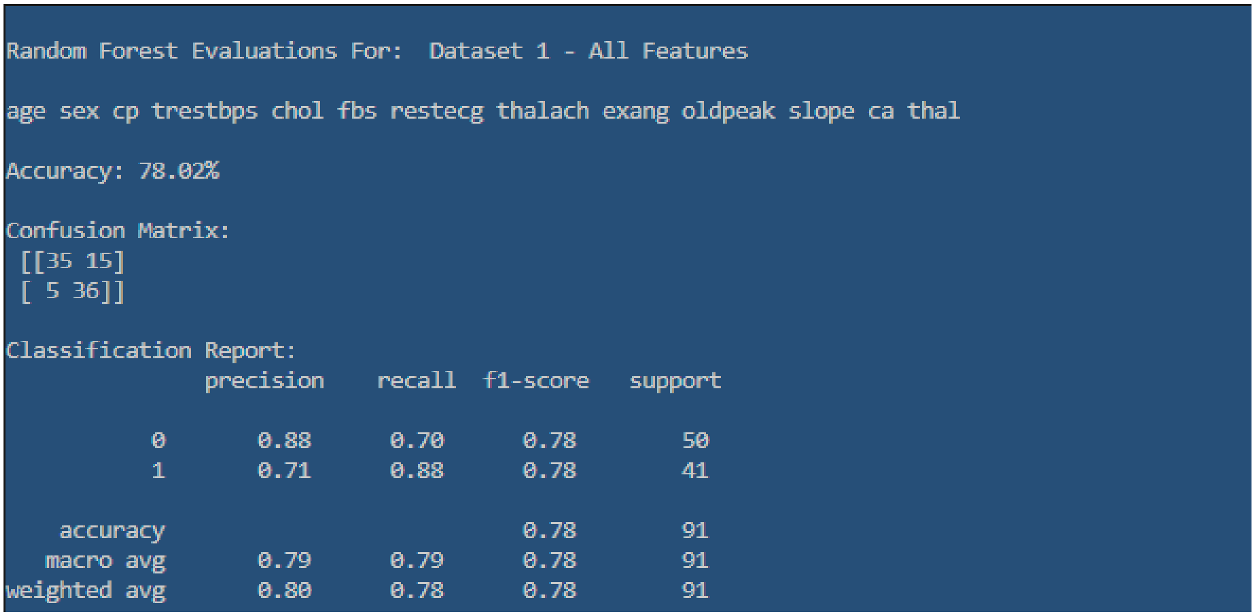
$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP+FP}$$

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP+FN}$$

III. Results and Discussion

Per Fig. 1 we establish baseline metrics using the RFC method. Using all feature attributes in heart DataSet1 where we trained the RFC model based on 70% of the data, the 30% test set shows a model that is 78.02% accurate. As established in the prior equation, accuracy is the percentage of all classifications (both positive and negative) that are correct. Next, the precision score weighted average is 80% which captures the percentage of the model's positive classifications that are in fact positive. Then, the recall weighted average is 78%, a metric that considers False Negatives or the number of positive heart disease predictions that were missed. The big take-away in all these metrics is that 100% precision and 100% recall would be the ideal perfect model and 100% accuracy in this context would reflect that all the predictions were useful.

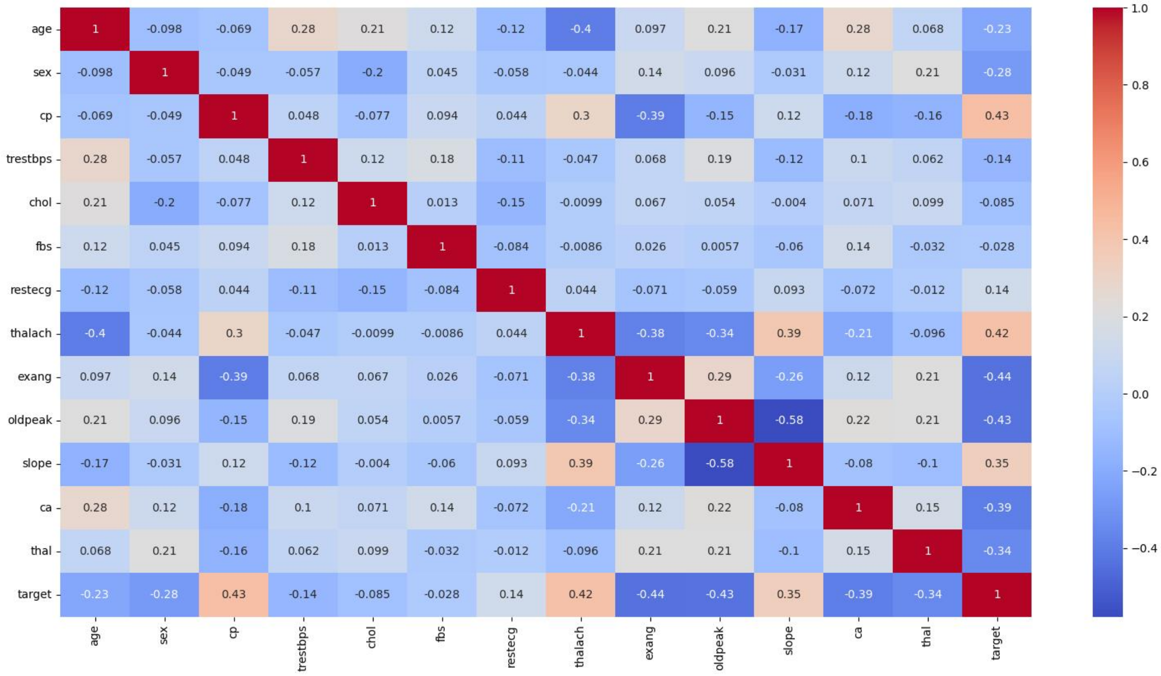
Fig. 1



```
Random Forest Evaluations For:  Dataset 1 - All Features

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal

Accuracy: 78.02%

Confusion Matrix:
 [[35 15]
 [ 5 36]]

Classification Report:
               precision    recall  f1-score   support

           0       0.88      0.70      0.78        50
           1       0.71      0.88      0.78        41

    accuracy                           0.78        91
   macro avg       0.79      0.79      0.78        91
weighted avg       0.80      0.78      0.78        91
```

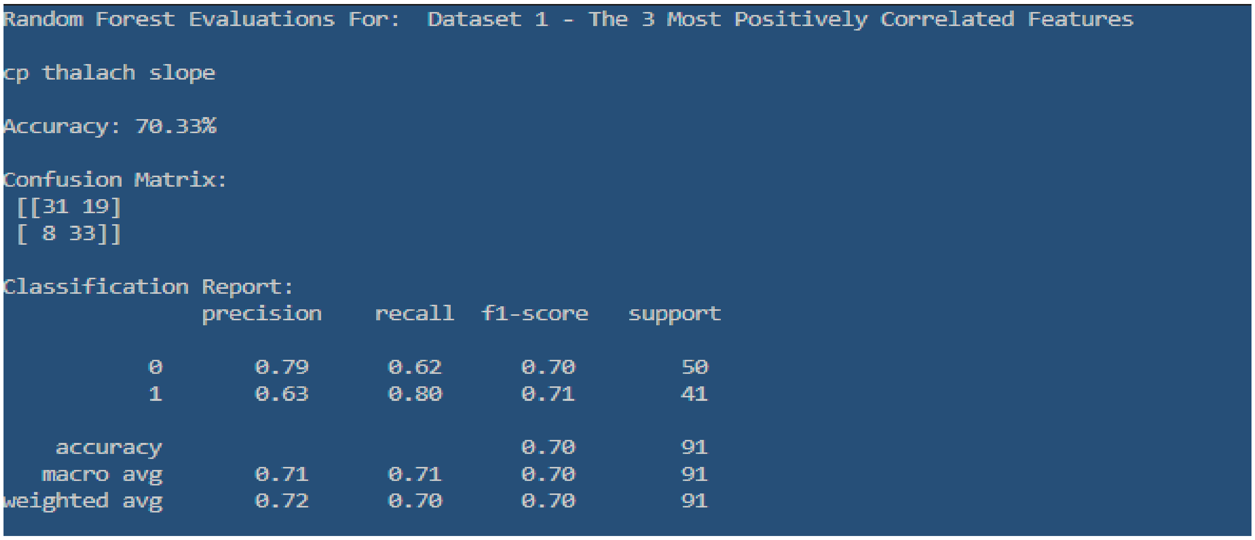
Next, we present the DataSet1 correlation matrix in Fig 2. Using a cool/warm heat map where the deep red color is the most positive correlation corresponding to the number value of 1 and deep blue is the most negative correlation corresponding to the number value of -1. Thus, the 3 features with the most positive correlations to the target are cp (0.43), thalach (0.42), and slope (0.35).

Fig. 2



Intuitively, one might think that if the RFC were trained again on only those 3 most positive correlated features that it would be a more effective model predictor. Yet, when generating the results we present the findings in Fig. 3:

Fig. 3



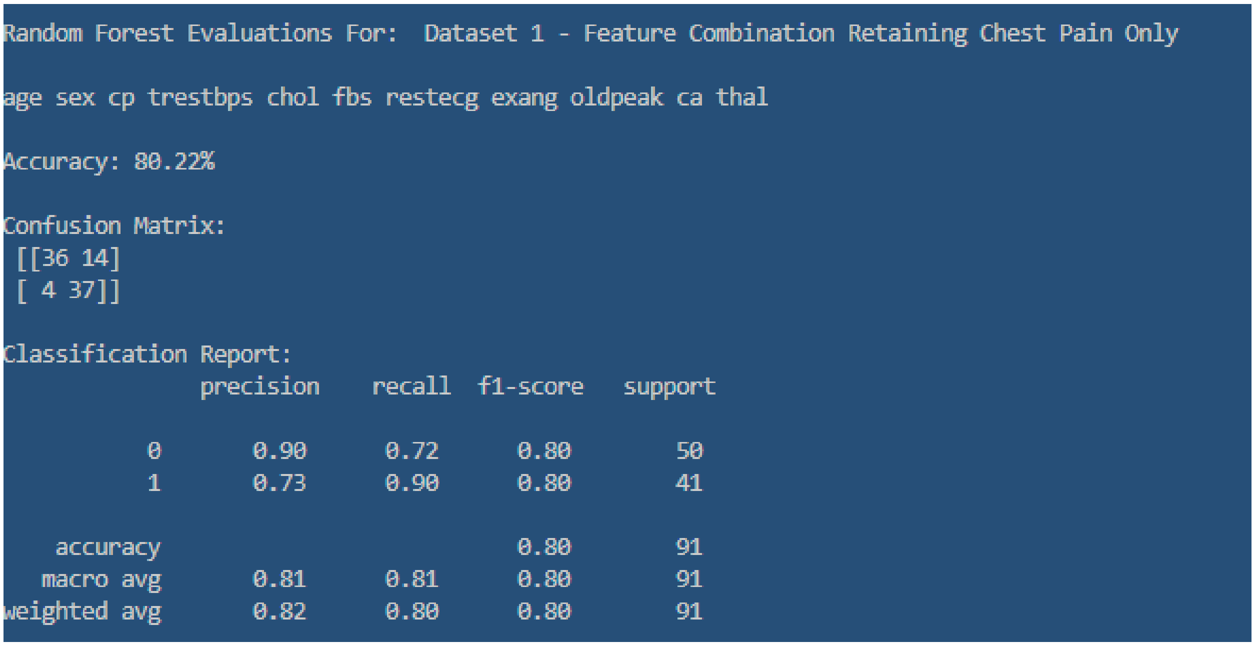
```
Random Forest Evaluations For:  Dataset 1 - The 3 Most Positively Correlated Features

cp thalach slope

Accuracy: 70.33%

Confusion Matrix:
 [[31 19]
 [ 8 33]]

Classification Report:
               precision    recall  f1-score   support

           0       0.79      0.62      0.70        50
           1       0.63      0.80      0.71        41

    accuracy                           0.70        91
   macro avg       0.71      0.71      0.70        91
weighted avg       0.72      0.70      0.70        91
```

He we find the model is less accurate, less precise, and suffers a lower recall. Multiple reasons for these results are discussed in [8] where we highlight some issues that arise with highly correlated features:

- Redundancy doesn't add unique information and can increase complexity without adding value.
- Multicollinearity can occur which can lead to model instability, for example, sensitivity to small changes in data.
- Lack of model generalization results in overfitting data making the model less robust and unable to generalize to new data.

Moreover, as shown in Fig. 4 reducing DataSet1 by eliminating features thalach (0.42) and slope (0.35) and retaining the single highest correlated feature cp (0.43) along with all the original remaining features, the model is more accurate, more precise, and achieves a higher recall.

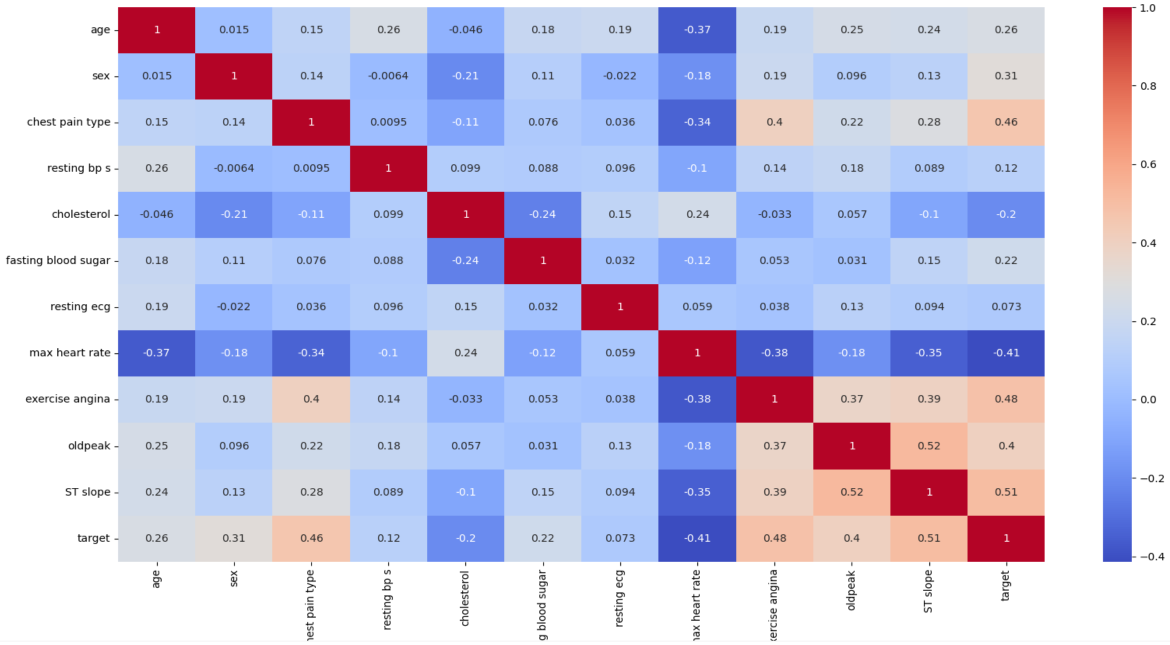
Fig. 4



```
Random Forest Evaluations For:  Dataset 1 - Feature Combination Retaining Chest Pain Only

age sex cp trestbps chol fbs restecg exang oldpeak ca thal

Accuracy: 80.22%

Confusion Matrix:
 [[36 14]
 [ 4 37]]

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.72      0.80        50
           1       0.73      0.90      0.80        41

    accuracy                           0.80        91
   macro avg       0.81      0.81      0.80        91
weighted avg       0.82      0.80      0.80        91
```

In Fig. 5 we turn attention to DataSet2. Interestingly, when adding nearly 300% more data instances we find a change in the correlation matrix. The 3 features with the most positive correlations to the target are chest pain type (0.46), exercise angina (0.48), and ST Slope (0.51). While chest pain type and ST slope are common across the two datasets, in DataSet2 exercise angina replaces thalach or max heart rate as a positive correlated feature.

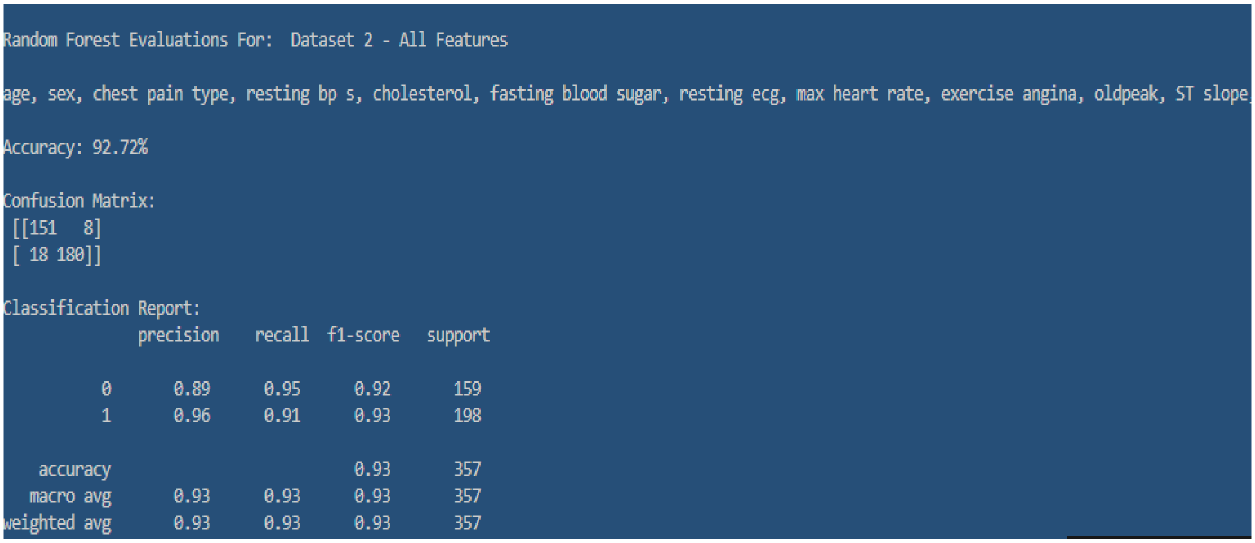
Fig. 5



	age	sex	hest pain type	resting bp s	cholesterol	g blood sugar	resting ecg	ax heart rate	xercise angina	oldpeak	ST slope	target
age	1	0.015	0.15	0.26	-0.046	0.18	0.19	-0.37	0.19	0.25	0.24	0.26
sex	0.015	1	0.14	-0.0064	-0.21	0.11	-0.022	-0.18	0.19	0.096	0.13	0.31
chest pain type	0.15	0.14	1	0.0095	-0.11	0.076	0.036	-0.34	0.4	0.22	0.28	0.46
resting bp s	0.26	-0.0064	0.0095	1	0.099	0.088	0.096	-0.1	0.14	0.18	0.089	0.12
cholesterol	-0.046	-0.21	-0.11	0.099	1	-0.24	0.15	0.24	-0.033	0.057	-0.1	-0.2
fasting blood sugar	0.18	0.11	0.076	0.088	-0.24	1	0.032	-0.12	0.053	0.031	0.15	0.22
resting ecg	0.19	-0.022	0.036	0.096	0.15	0.032	1	0.059	0.038	0.13	0.094	0.073
max heart rate	-0.37	-0.18	-0.34	-0.1	0.24	-0.12	0.059	1	-0.38	-0.18	-0.35	-0.41
exercise angina	0.19	0.19	0.4	0.14	-0.033	0.053	0.038	-0.38	1	0.37	0.39	0.48
oldpeak	0.25	0.096	0.22	0.18	0.057	0.031	0.13	-0.18	0.37	1	0.52	0.4
ST slope	0.24	0.13	0.28	0.089	-0.1	0.15	0.094	-0.35	0.39	0.52	1	0.51
target	0.26	0.31	0.46	0.12	-0.2	0.22	0.073	-0.41	0.48	0.4	0.51	1

Repeating the initial experiment per Fig 6. we again establish the baseline metrics using the RFC method for all feature attributes in the heart DataSet2. Here we find a model that is 92.72% accurate with precision and recall at 93%.

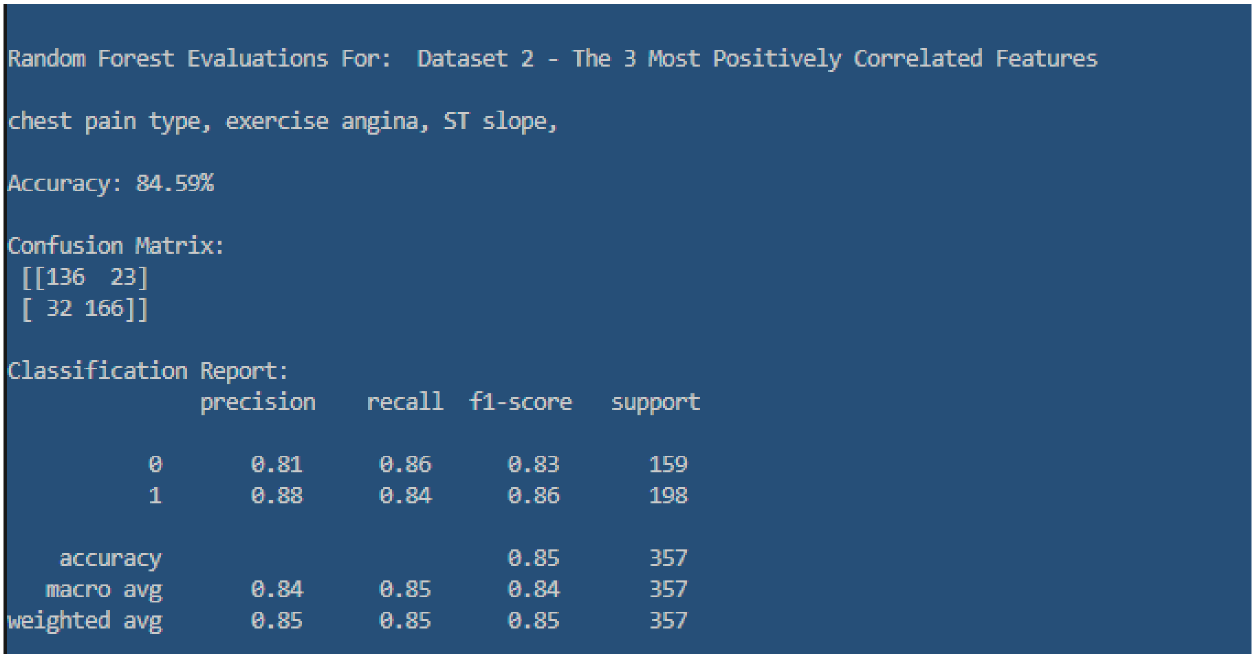
Fig. 6



```
Random Forest Evaluations For:  Dataset 2 - All Features

age, sex, chest pain type, resting bp s, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise angina, oldpeak, ST slope

Accuracy: 92.72%

Confusion Matrix:
 [[151   8]
 [ 18 180]]

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       159
           1       0.96      0.91      0.93       198

    accuracy                           0.93       357
   macro avg       0.93      0.93      0.93       357
weighted avg       0.93      0.93      0.93       357
```

Again as captured in Fig.7, true to form, intuition breaks down as we find that using only the 3 most positive correlated features of DataSet2 produces a less effective model predictor scoring lower in accuracy, precision, and recall compared to the baseline set of all features.

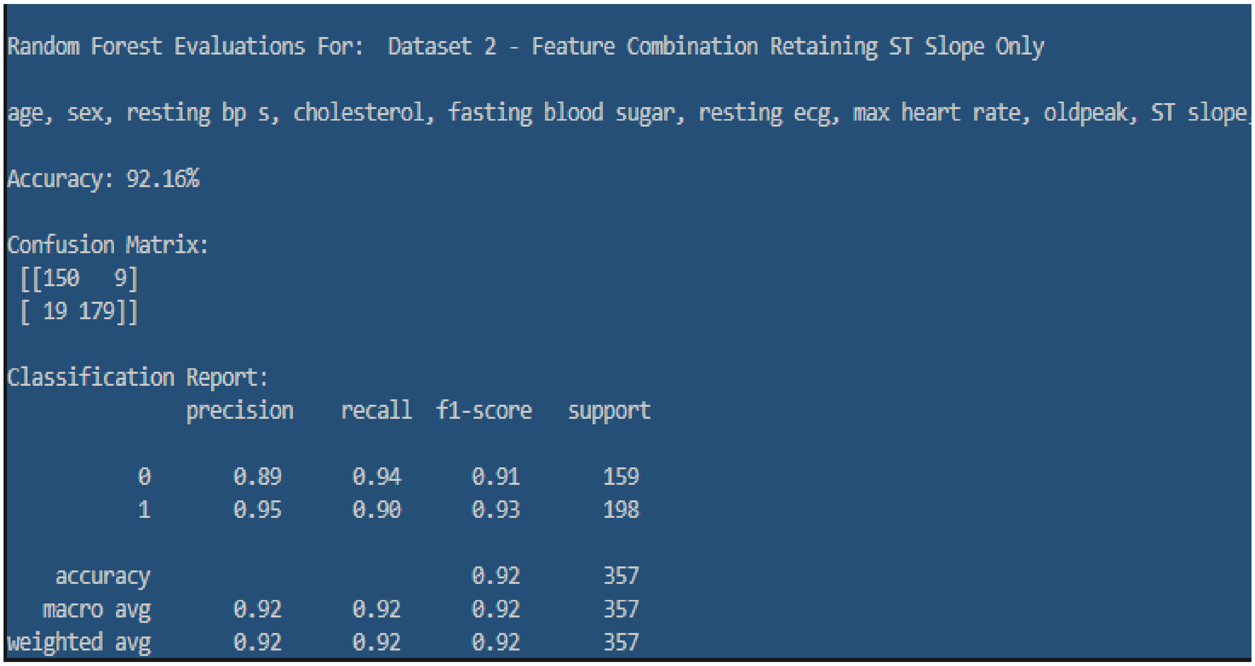
Fig. 7



```
Random Forest Evaluations For:  Dataset 2 - The 3 Most Positively Correlated Features

chest pain type, exercise angina, ST slope,

Accuracy: 84.59%

Confusion Matrix:
 [[136  23]
 [ 32 166]]

Classification Report:
               precision    recall  f1-score   support

           0       0.81      0.86      0.83       159
           1       0.88      0.84      0.86       198

    accuracy                           0.85       357
   macro avg       0.84      0.85      0.84       357
weighted avg       0.85      0.85      0.85       357
```

Finally, per Fig. 8, analyzing the entire original feature set while retaining only the single highest correlated feature ST slope (0.51) this time doesn't produce a better predictor than the baseline but is certainly comparable being within 1 percent on all metrics.

Fig. 8



```
Random Forest Evaluations For:  Dataset 2 - Feature Combination Retaining ST Slope Only

age, sex, resting bp s, cholesterol, fasting blood sugar, resting ecg, max heart rate, oldpeak, ST slope

Accuracy: 92.16%

Confusion Matrix:
 [[150   9]
 [ 19 179]]

Classification Report:
               precision    recall  f1-score   support

           0       0.89      0.94      0.91       159
           1       0.95      0.90      0.93       198

    accuracy                           0.92       357
   macro avg       0.92      0.92      0.92       357
weighted avg       0.92      0.92      0.92       357
```

IV. Conclusion

As counter to the conclusions drawn in [8], Manjit [10] contends that conventional wisdom and assumptions that multiple features with higher correlation results in redundancy and overfitting does not always hold true and that “removing highly correlated features might not always lead to better model performance…”. For example, high correlation is a measure of the relationship between features, but this does not necessarily reduce to redundancy given it is possible that each of the highly correlated features may still provide unique information to the model.

With respect to this analysis what do the results reveal? First, it must be conceded that this analysis was limited to only two datasets and one type of machine learning technique namely, Random Forest Classification or RFC. Moreover, this analysis focused on the 3 ***most positive correlated*** features which does not necessarily mean these are strongly correlated in the context of the correlation scale ranging from -1 to 1. In fact, the most positive correlated features over both datasets ranged from .35 to .51, which suggests moderate rather than strong positive correlation. Yet, this experiment yielded results more in line with the conclusions found in [8] or the general belief that features that are more positively correlated tend to be redundant and that model performance is not significantly degraded by maintaining only a subset of a group of the most positive correlated features, especially those with moderate to strong positive correlation.

References

[1] Changes in the Leading Cause of Death: Recent Patterns in Heart Disease and Cancer Mortality. (2016, August). National Center for Health Statistics. cdc.gov. https://blogs.cdc.gov/nchs/2016/08/24/3197/

[2] U.S. Heart Attacks Deaths from 2010-2015. (2017, February). National Center for Health Statistics. cdc.gov. https://blogs.cdc.gov/nchs/2017/02/15/3439/

[3] The Heart Truth. (n. d.). National Heart, Lung, and Blood Institute. nhlbi.nih.gov. https://www.nhlbi.nih.gov/health-topics/education-and-awareness/heart-truth

[4] Mastroianni, B. (2024, March). Most Americans Don’t Know Heart Disease is Leading Cause of Death. health.com. https://www.health.com/most-americans-unaware-heart-disease-leading-cause-death-8550933

[5] Rezaei, A. (2023). Exploring the Heart Dataset. kaggle.com. https://www.kaggle.com/datasets/arezaei81/heartcsv

[6] Heart Disease. (n. d.) UC Irvine Machine Learning Repository. uci.edu. https://archive.ics.uci.edu/dataset/45/heart+disease

[7] Classification: Accuracy, recall, precision, and related metrics. (2024, November). developers.google.com. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

[8] Mudadla, S. (2023, November). Why we have to remove highly correlated features in Machine Learning. medium.com. https://medium.com/@sujathamudadla1213/why-we-have-to-remove-highly-correlated-features-in-machine-learning-9a8416286f18#

[9] Siddhartha, M. (2020, November). Heart Disease Dataset. dataport.org. https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

[10] Baishya, M. (2024, June). Does Removal of Highly Correlated Features Always Improve Model Performance? medium.com. https://medium.com/@datacodedesign/does-removal-of-highly-correlated-features-always-improve-model-performance-8d820d30b71d#