CS 249, Winter 2024

Final Project

William Zhong

Author	Version	Date
William Zhong	Milestone 1.0	Feb 27, 2024
William Zhong	Final	Mar 11, 2024

Abstract:

This project is attempting to look for relevant features to find practical method to predcit hypertension. After statistical analysis I filtered un-correlated and negatively correlated features for ML to predict the result of hypertension. The outcome was good (accuracy was above 95% with best models)

Introduction:

Hypertension is one of the leading causes of death for Americans. According to the final data in 2017 from the National Vital Statistics Report, nearly 2.9 million Americans were reported having hypertension in the United States. Death ratio of hypertension was 731.9 : 100,000 in the U.S. standard population. This ratio increased 0.6% from the 2016 rate, and the overall/average life expectancy has decreased by 0.1% from the 2016 rate. The age-adjusted death rate decreased from 2016 but increased in 2017. Also according to the NCHS Data Brief, Number 364, 2017-2018, hypertension increased with age 22.4% for age from 18 to 39, 54.5% from 40-59, and 74.5% for age 60 and over.

Data Description:

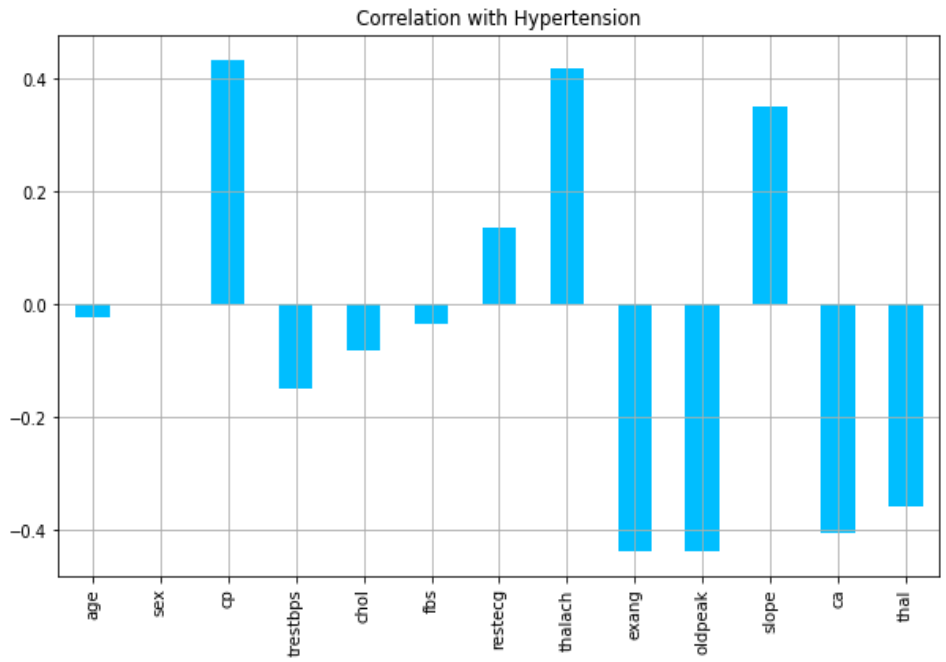
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	
count	26083.000000	26058.00000	26083.000000	26083.000000	26083.000000	26083.000000	26083.000000	26083.000000	26083.000000	26083.000000	26083.0(
mean	55.661389	0.50000	0.958594	131.592992	246.246061	0.149753	0.526512	149.655024	0.326573	1.039512	1.4(
std	15.189768	0.50001	1.023931	17.588809	51.643522	0.356836	0.525641	22.858109	0.468969	1.165138	0.6
min	11.000000	0.00000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.0(
25%	44.000000	0.00000	0.000000	120.000000	211.000000	0.000000	0.000000	133.000000	0.000000	0.000000	1.0(
50%	56.000000	0.50000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.0(
75%	67.000000	1.00000	2.000000	140.000000	275.000000	0.000000	1.000000	166.000000	1.000000	1.600000	2.0(
max	98.000000	1.00000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.0(

slope	ca	thal	target
26083.000000	26083.000000	26083.000000	26083.000000
1.400299	0.721849	2.318752	0.547253
0.616513	1.011608	0.604659	0.497772
0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	2.000000	0.000000
1.000000	0.000000	2.000000	1.000000
2.000000	1.000000	3.000000	1.000000
2.000000	4.000000	3.000000	1.000000

Step-by-step Methodology:

1. Correlation between each feature and hypertension

Findings from step 1:



This analysis will help me to select all postitve correlation features to hypertension. There are few features like ‘cp’, ‘thalach’ , ‘slope’, which are significantly correlated to hypertension. Only ‘restecg’ is weakly correlated.

Split the group 0 for non hypertension and group 1 for hypertension, I have the ATE result:

negative chest pain average outcome: 0.47023456685578796
positive chest pain average outcome: 1.3626173462239035
Estimated ATE is: 0.8923827793681156

Maximum heart rate achieved average outcome: 139.12439664662546
Maximum heart rate achieved average outcome: 158.36710102283874
Estimated ATE is: 19.242704376213283

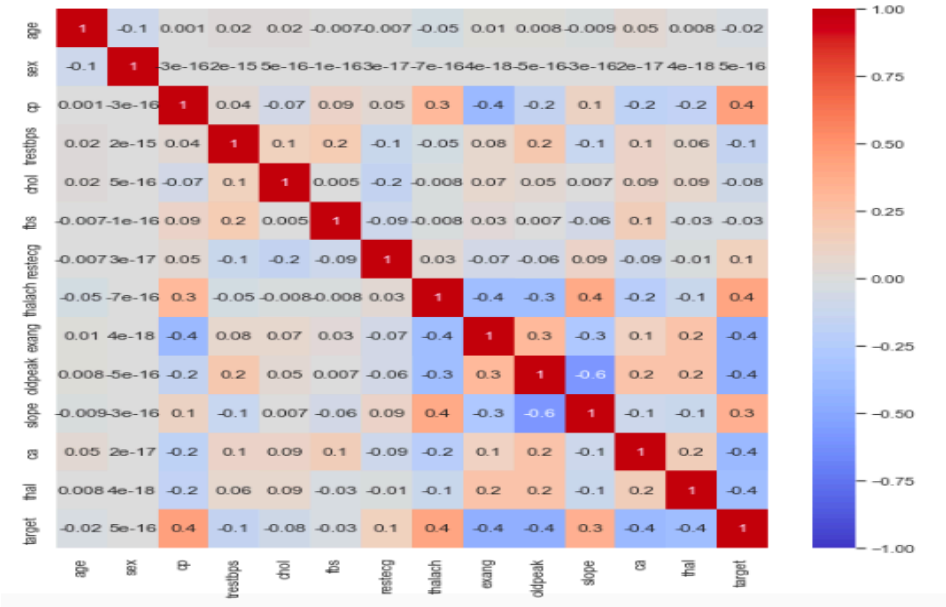
The slope of the peak exercise average outcome: 1.1631806249470742
The slope of the peak exercise average outcome: 1.596469104665826
Estimated ATE is: 0.43328847971875173

The Resting ECG average outcome: 0.4478787365568634
The Resting ECG average outcome: 0.5915650833683621
Estimated ATE is: 0.14368634681149872

I used ATE to assume which feature has the significant effect to hypertension. It looks like the maximum heart rate has a significant effect across two groups split by the target.

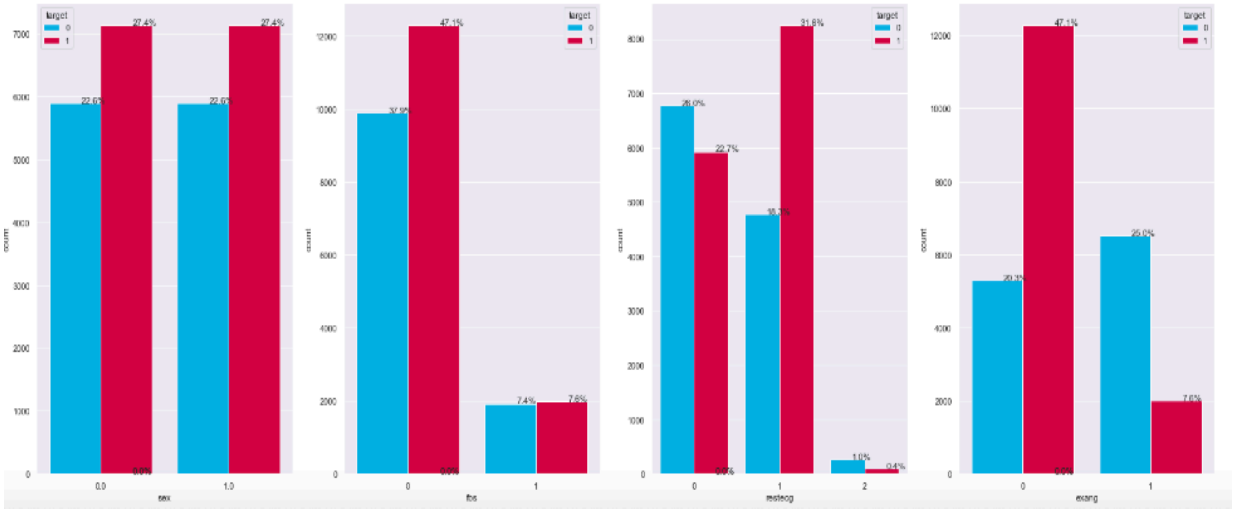
2. Correlation between any two features:

Findings from step 2:



The findings will help me to verify if any of these combination is good fit for training/validation. Since ‘cp’, ‘thalach’,and ‘slope’ are not higher/smaller than +- 0.3. Hence I cannot drop any features at this moment.

3. Check if binary/categorical value correlated to hypertension



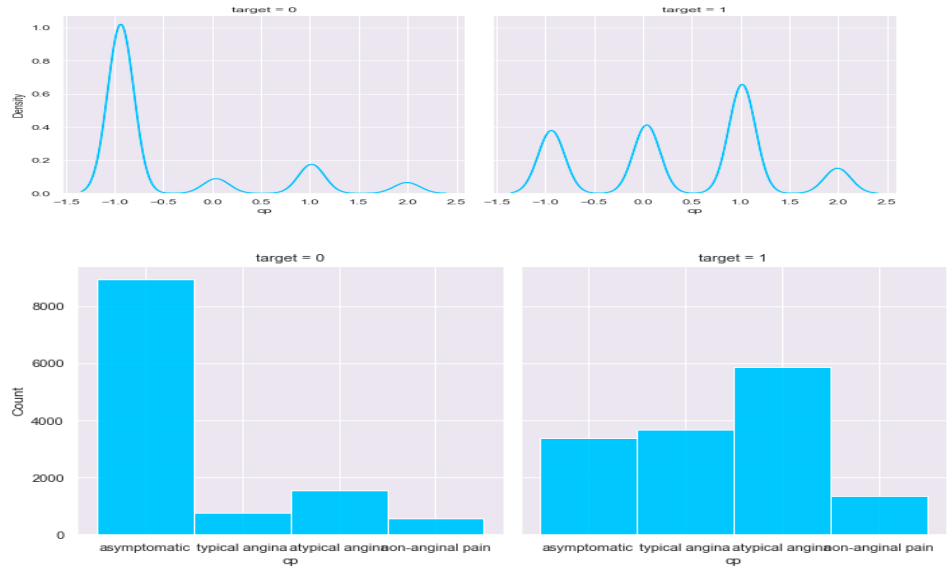
It seems sex has no effect on hypertension.
Patients who are not fasting blood sugar have a higher risk of hypertension.
Abnormal resting ECG has higher risk of hypertension
Looks like exercise induced angina has lower risk of hypertension

4. Distribution of correlated features to hypertension

Chest Pain:

target	cp	asymptomatic	typical angina	atypical angina	non-anginal pain
0		8930.0	776.0	1532.0	571.0
1		3384.0	3680.0	5860.0	1350.0

<Figure size 864x360 with 0 Axes>

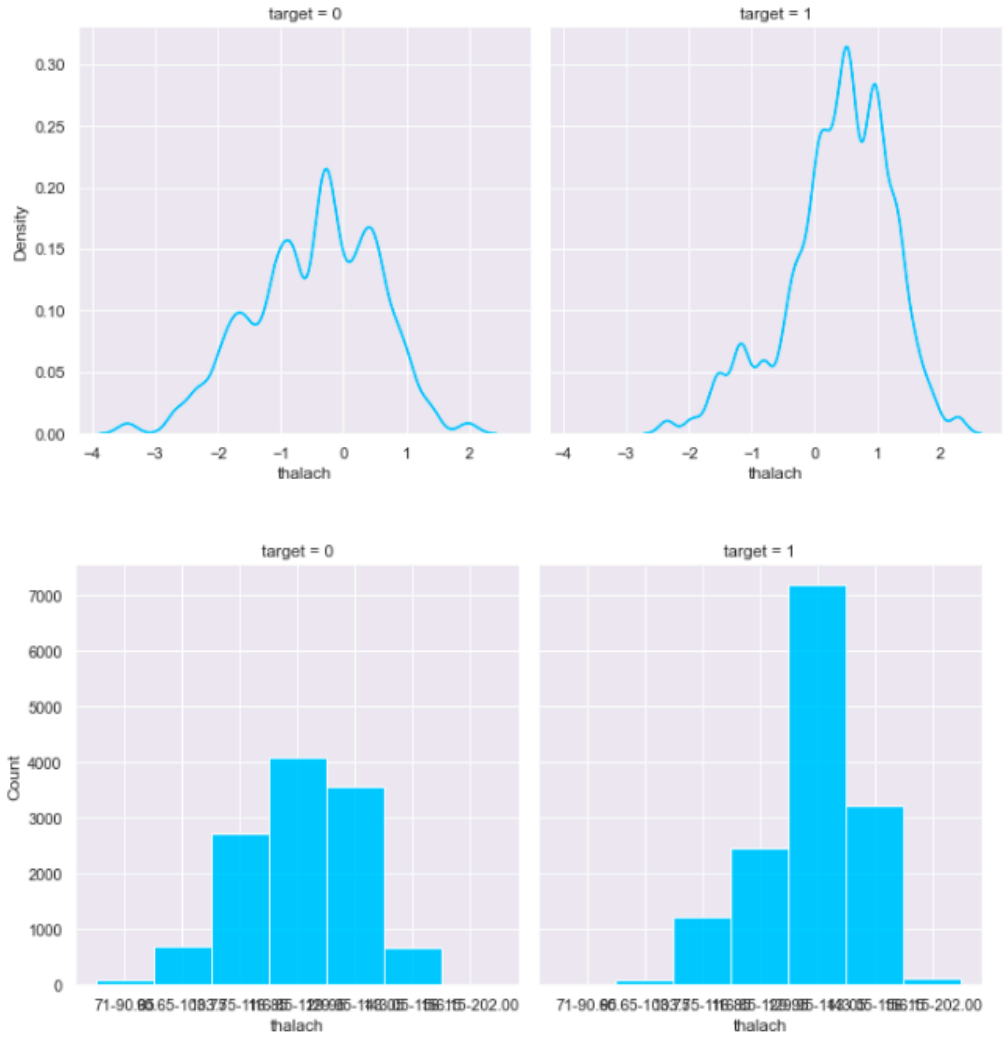


The distribution of the chest pain level of the target column values is different. This feature is an explanatory feature; People with high chest pain are more likely to have hypertension than those with normal values.

Maximum heart rate achieved:

thalach	71-90.65	90.65-103.75	103.75-116.85	116.85-129.95	129.95-143.05	143.05-156.15	156.15-202.00
target							
0	82.0	692.0	2730.0	4082.0	3569.0	654.0	0.0
1	0.0	84.0	1216.0	2448.0	7186.0	3232.0	108.0

```
<Figure size 864x360 with 0 Axes>
```

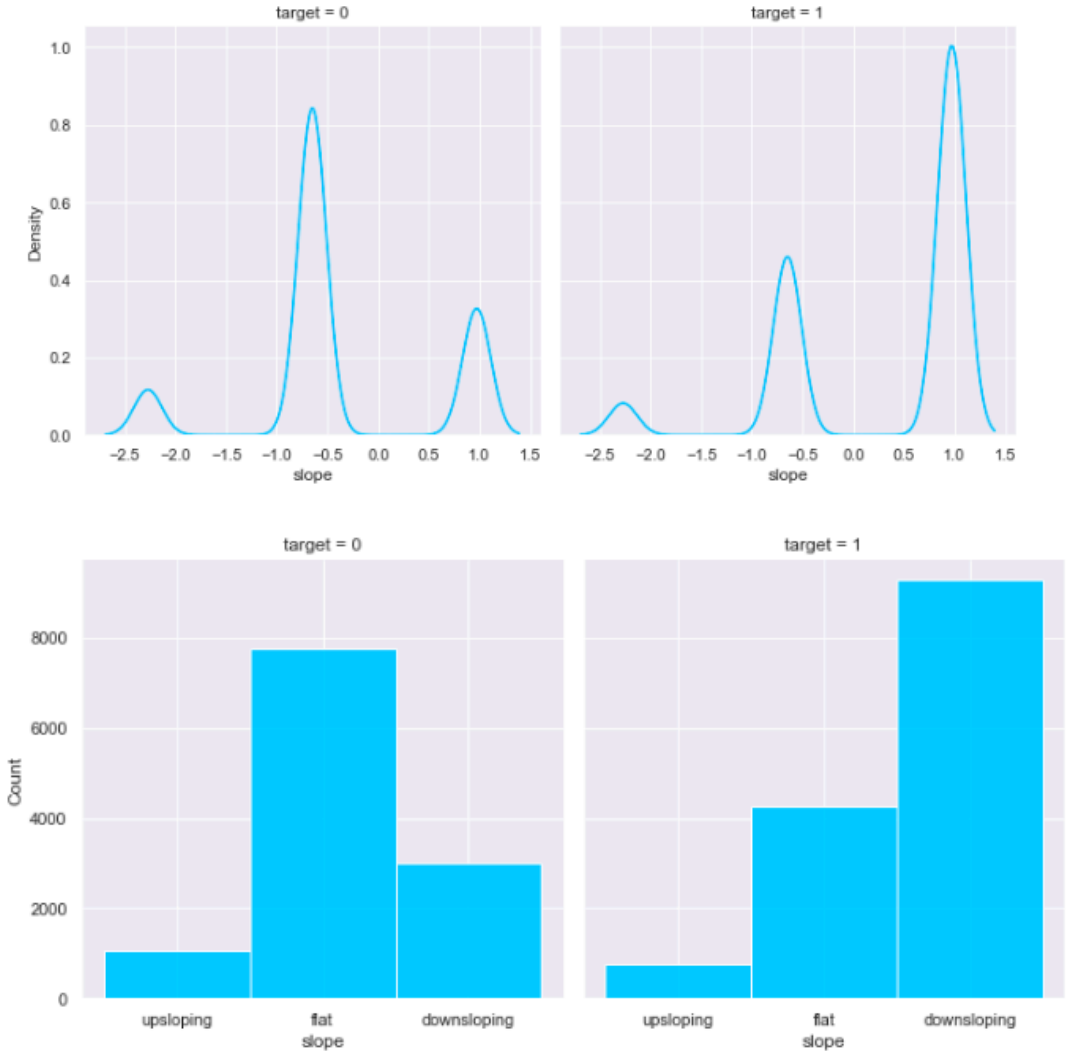


The distribution of the maximum heart rate of the target column values is different. This feature is an explanatory feature; People with high maximum heart rate are more likely to have hypertension than those with normal values.

The slope of the peak exercise ST segment:

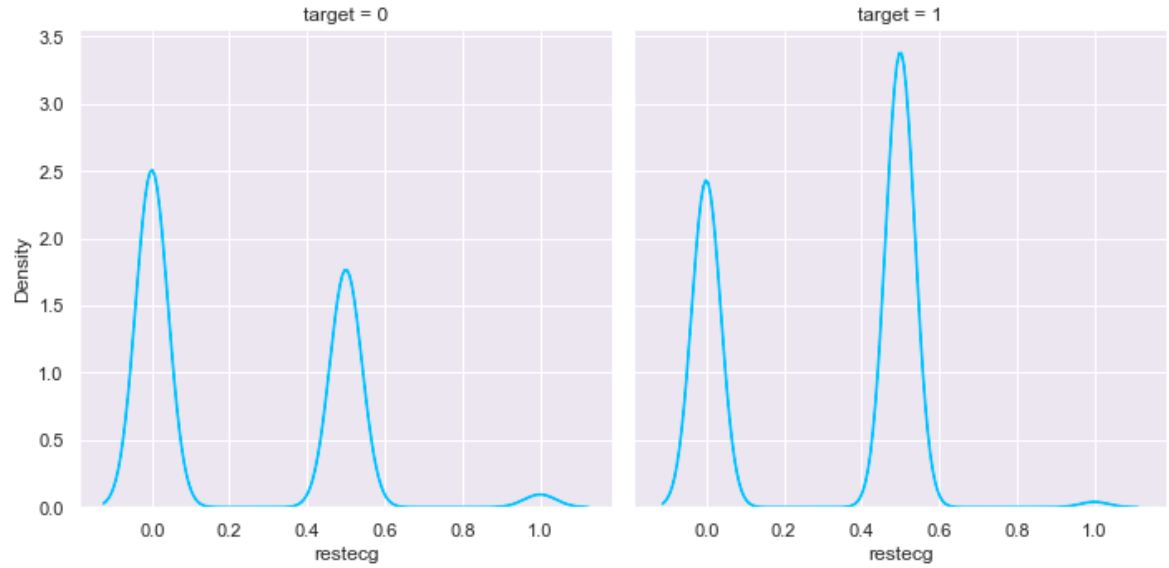
slope	upsloping	flat	downsloping
target			
0	1070.0	7742.0	2997.0
1	756.0	4248.0	9270.0

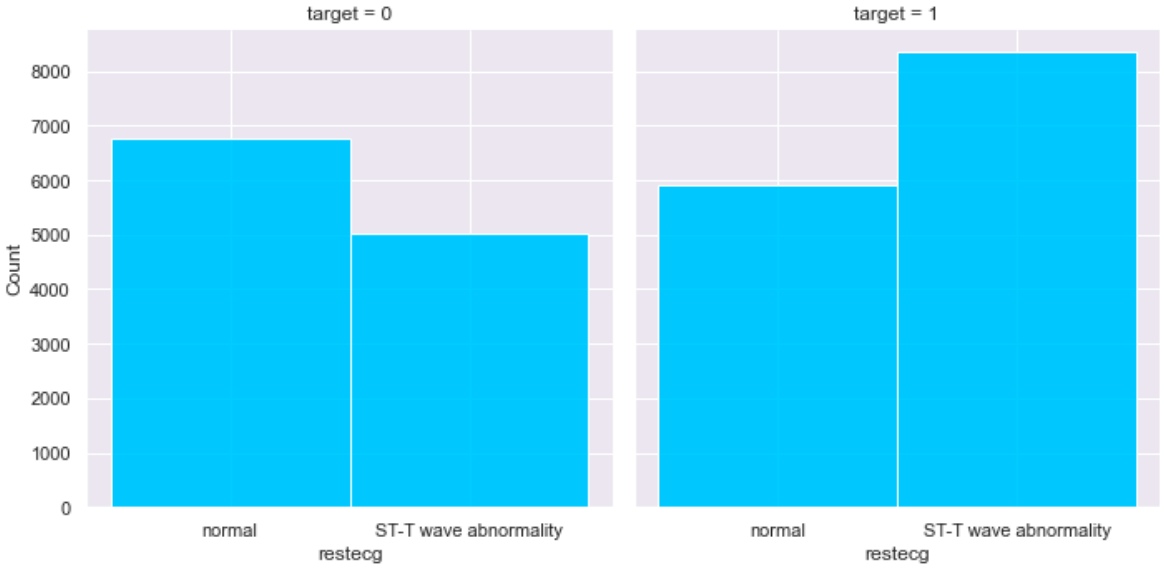
```
<Figure size 864x360 with 0 Axes>
```



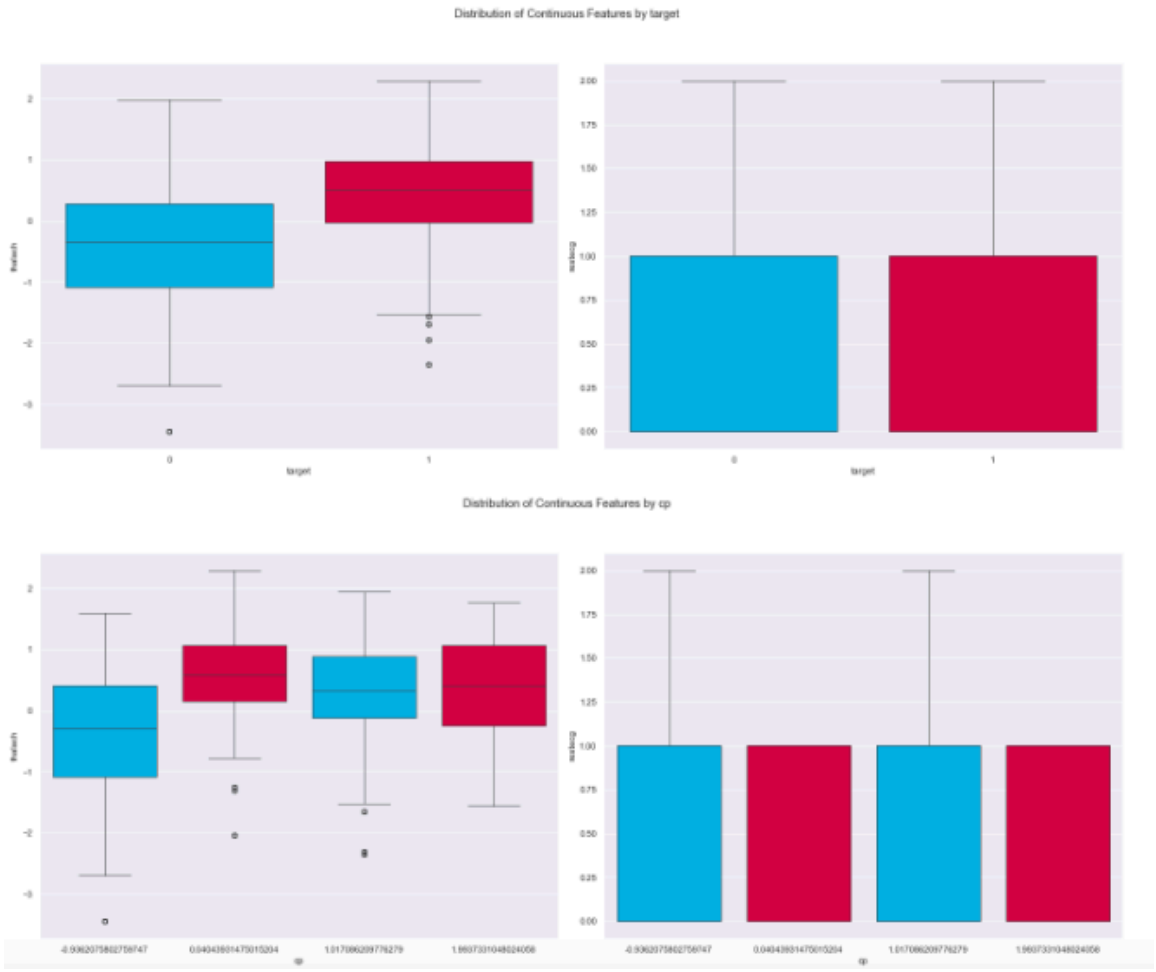
The distribution of the slope of the peak exercise of the target column values is different. This feature is an explanatory feature; People with downsloping are more likely to have hypertension than those with flat slope.

The resting ecg:





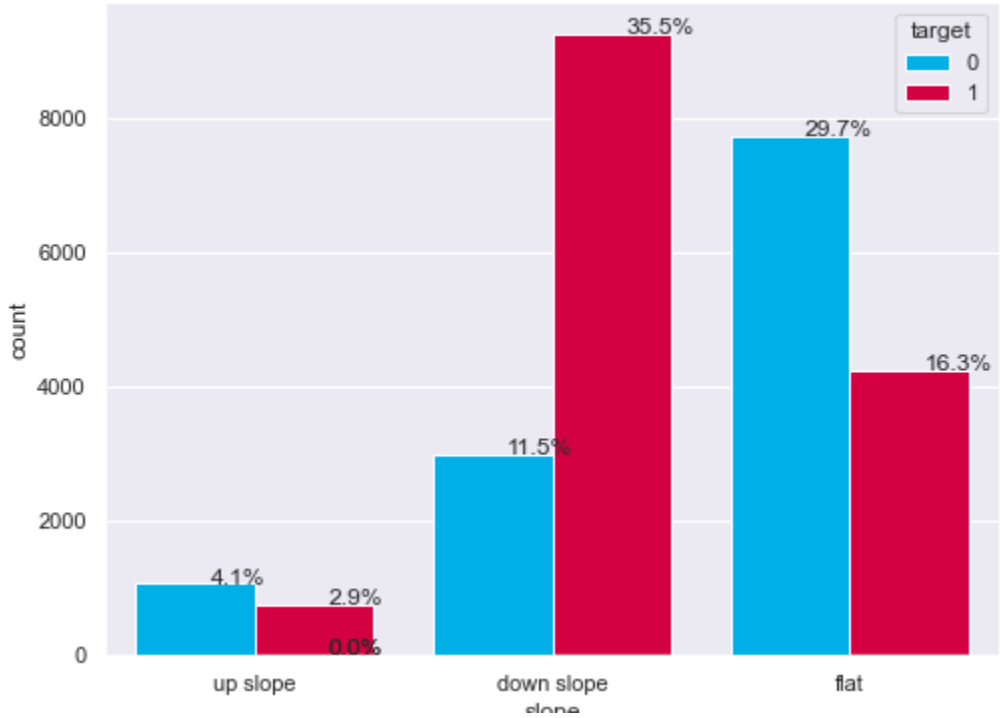
5. Distribution of numerical features by hypertension and correlated features



The median, range and values in the "thalach" feature is higher when there is chest pain. The median and values in the same feature is higher when there is hypertension.

6. Split the category feature ‘slope’ by values and apply it the dataset:

The postive correlated feature selection ‘slope’ needs to split by category:



All category has has hypertension, I will keep all selection even the up slope is at low percentage (4.1% / 2.9%)

The “slope" feature is of the Nominal Variable type. Therefore One Hot Encoder must be activated here.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	thal	target	slope_down slope	slope_normal	slope_up slope
0	57.0	1.0	3	145	233	1	0	150	0	2.3	0	1	1	0	0	1
1	64.0	0.0	2	130	250	0	1	187	0	3.5	0	2	1	0	0	1
2	52.0	1.0	1	130	204	0	0	172	0	1.4	0	2	1	1	0	0
3	56.0	0.0	1	120	236	0	1	178	0	0.8	0	2	1	1	0	0
4	66.0	0.0	0	120	354	0	1	163	1	0.6	0	2	1	1	0	0

7. Skewness Checking

	skew	too_skewed
cp	0.494495	False
thalach	-0.521341	False
restecg	0.172989	False
slope_down slope	0.118992	False
slope_normal	0.161790	False
slope_up slope	3.370580	True

There are 1 row with high Skewness. Therefore, we will normalize them using the QuantileTransformer method.

After I tried the quantile transform. The up slope still shows skewed. Hence I have to drop the 'up slope' feature for ML prediction.

8. Data Scaling

The StandardScaler method was chosen because of the shape of the distribution of these features.

	cp	thalach	restecg	slope_down slope	slope_normal
count	2.608300e+04	2.608300e+04	2.608300e+04	2.608300e+04	2.608300e+04
mean	8.662830e-17	3.440614e-16	-3.098732e-17	-9.316628e-17	-2.274674e-17
std	1.000019e+00	1.000019e+00	1.000019e+00	1.000019e+00	1.000019e+00
min	-9.362076e-01	-3.441078e+00	-1.001675e+00	-9.422758e-01	-9.223757e-01
25%	-9.362076e-01	-7.286405e-01	-1.001675e+00	-9.422758e-01	-9.223757e-01
50%	4.043931e-02	1.463393e-01	9.008002e-01	-9.422758e-01	-9.223757e-01
75%	1.017086e+00	7.150762e-01	9.008002e-01	1.061260e+00	1.084157e+00
max	1.993733e+00	2.290040e+00	2.803276e+00	1.061260e+00	1.084157e+00

9. Data Splitting

```
1    0.546609
0    0.453391
Name: target, dtype: float64
1    0.548559
0    0.451441
Name: target, dtype: float64
```

10. ML train and prediction - Method 1:

Method 1 is included all valid features (no skewness and positively correlated)

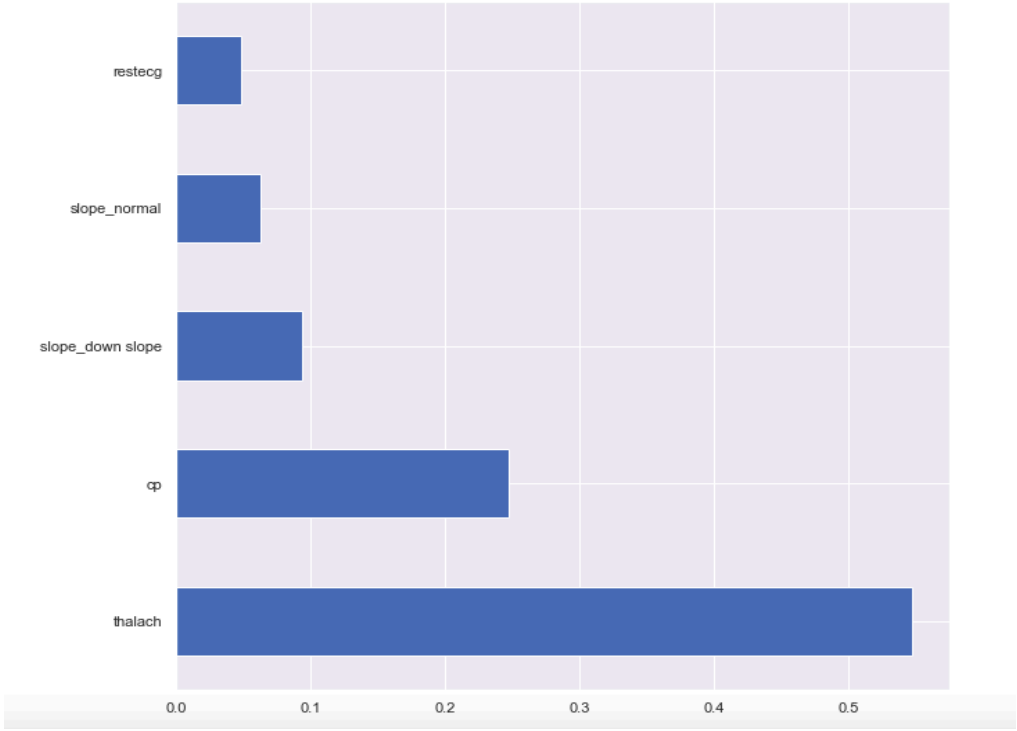
	accuracy	f1_score	precision	recall	balanced_accuracy	auc
K Nearest Neighbors - Method 1	0.954693	0.958660	0.959677	0.957645	0.954376	0.954376
Logistic Regression - Method 1	0.770795	0.802245	0.761560	0.847522	0.762541	0.762541
Naive Bayes - Method 1	0.749419	0.775850	0.761681	0.790555	0.744994	0.744994
Support Vector Machine - Method 1	0.872909	0.879886	0.913589	0.848581	0.875526	0.875526
Decision Trees - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Random Forest - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Extra Trees - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Gradient Boosting - Method 1	0.957249	0.960388	0.976576	0.944727	0.958596	0.958596
Ada Boost - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Stacking Voting - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007

We see the majority of ML methods are predicted with excellent accuracy (above 95% with no overfitting)

11. ML train and prediction - Method 2:

Method 2 will select top 3 features based on three different feature selection models.
Use inbuilt class feature_importances of tree based classifiers.

ExtraTreesClassifier:



SelectKBest:

	Specs	Score
1	slope_down slope	2150.936239
2	slope_normal	1801.772365
0	cp	1789.566335
3	thalach	232.239305

LogisticRegression:

Best accuracy score: 0.77
Best subset (indices): (0, 1, 3, 4)
Best subset (corresponding names): ('cp', 'slope_down slope', 'thalach', 'restecg')

It's more likely using top 3 ranked features('cp', 'down slope' and 'thalach')

After applied these selected features:

K Nearest Neighbors - Method 2	0.921534	0.928737	0.928306	0.929169	0.920680	0.920680
Logistic Regression - Method 2	0.753994	0.799792	0.724242	0.892940	0.738456	0.738456
Naive Bayes - Method 2	0.752332	0.788011	0.744830	0.836507	0.742920	0.742920
Support Vector Machine - Method 2	0.779681	0.799768	0.799954	0.799582	0.777455	0.777455
Decision Trees - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Random Forest - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Extra Trees - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Gradient Boosting - Method 2	0.928562	0.934643	0.941135	0.928240	0.928598	0.928598
Ada Boost - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Stacking Voting - Method 2	0.927796	0.933615	0.944828	0.922666	0.928369	0.928369

The result is slightly worse than method1. This could be caused by feature selection bias. We should add a normal slope feature for better performance.

12. Summary

	accuracy	f1_score	precision	recall	balanced_accuracy	auc
Gradient Boosting - Method 1	0.957249	0.960388	0.976576	0.944727	0.958596	0.958596
Ada Boost - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Stacking Voting - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Decision Trees - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Random Forest - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
Extra Trees - Method 1	0.955158	0.958637	0.970282	0.947268	0.956007	0.956007
K Nearest Neighbors - Method 1	0.953764	0.957542	0.964746	0.950445	0.954121	0.954121
Gradient Boosting - Method 2	0.928562	0.934643	0.941135	0.928240	0.928598	0.928598
Stacking Voting - Method 2	0.927796	0.933615	0.944828	0.922666	0.928369	0.928369
Decision Trees - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Random Forest - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Extra Trees - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
Ada Boost - Method 2	0.926773	0.933719	0.930168	0.937297	0.925596	0.925596
K Nearest Neighbors - Method 2	0.921534	0.928737	0.928306	0.929169	0.920680	0.920680
Support Vector Machine - Method 1	0.872909	0.879886	0.913589	0.848581	0.875526	0.875526
Support Vector Machine - Method 2	0.779681	0.799768	0.799954	0.799582	0.777455	0.777455
Logistic Regression - Method 1	0.769865	0.802394	0.758439	0.851758	0.761056	0.761056
Logistic Regression - Method 2	0.753994	0.799792	0.724242	0.892940	0.738456	0.738456
Naive Bayes - Method 2	0.752332	0.788011	0.744830	0.836507	0.742920	0.742920
Naive Bayes - Method 1	0.728857	0.748762	0.761384	0.736552	0.728029	0.728029

- No overfitting/underfitting for all high accuracy ML methods.
- Feature selection causes bias, therefore method 1 is the best method to predict hypertension.
- Unfortunately I cannot achieve 100% accuracy. Suggest to add more relevant health features to improve ML performance.
- SVM, Logistic Regression, and Naive Bayes are not good models for any method.
- High accuracy performance are very similar to each method from the majority of ML models. I also observed method 1 has slightly better performance than method 2

Health features are more likely to affect the goal column - adding such features can improve models in the medical field.

According to the results in this study, using one of the models that yielded about 96% accuracy can serve as an excellent tool for predicting stroke.

Test the models with new data can help validate the models.