STA160 Midterm Project

By Cheng Liu, Junxi Chen, Zhongwen Liang, Ziyi Ma, Xinmei Wang

Introduction:

In modern society, Cardiovascular disease has been the number one lethal disease in the U.S. and about 16 percent of all deaths worldwide. As individuals who are able to develop heart disease in the future, we want to find out which variable or variables have greater effects on developing heart disease so that we could try to prevent it from happening and minimize the chance of it occurring. In order to explore the major factors associated with Cardiovascular disease, we are doing extensive data exploration by using machine learning tools to predict the number of people who may face a higher risk of cardiovascular disease combined with several essential variables. Some methods that we will be using are logistic regression, the KNN neighbor classification, and the SVM as well.

Background:

The dataset we are going to use is "Heart Disease Health Indicators Dataset" from Kaggle. This dataset contains 253,680 survey responses from cleaned BRFSS 2015 to be used primarily for the binary classification of heart disease, and this data includes lots of variables that are related to this heart disease. Some examples would be "Highbp, highChol, BMI, Smoker, and AnyHealthcare. We will use all of the variables and not drop any of them because there is no response that is null. Our data was collected by both questions directly asked of participants and calculated variables based on individual participant responses. We will not draw graphs for all of the variables listed but we will choose to display graphs that are related to the variables that we think would have an effect on whether an individual would have a high risk of getting the disease. Hence, we will try to discover some relationship between those variables, and to see what we can get.

Data Visualization

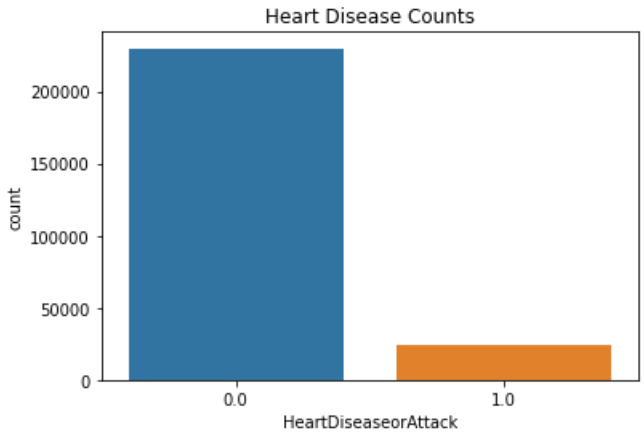
Correlation Heat Map

As we can see in the right lower corner of our correlation heat map. There are several orangeish blocks, which show a comparable higher correlation with other blueish blocks. These are *PhysHlth and GenHlth*, *DiffWalk and GenHlth*, *Income and Education, DiffWalk and PhyHlth.*

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
HeartDiseaseorAttack	1.000000	0.209361	0.180765	0.044206	0.052904	0.114441	0.203002	0.180272	−0.087299	−0.019790	−0.039167	−0.028991	0.018734	0.031000	0.258383	0.064821	0.181698	0.212709	0.086096	0.221618	−0.099600	−0.141011
HighBP	0.209361	1.000000	0.298199	0.098508	0.213748	0.096991	0.129575	0.271596	−0.125267	−0.040555	−0.061266	−0.003972	0.038425	0.017358	0.300530	0.056456	0.161212	0.223618	0.052207	0.344452	−0.141358	−0.171235
HighChol	0.180765	0.298199	1.000000	0.085642	0.106722	0.091299	0.092620	0.209085	−0.078046	−0.040859	−0.039874	−0.011543	0.042230	0.013310	0.208426	0.062069	0.121751	0.144672	0.031205	0.272318	−0.070802	−0.085459
CholCheck	0.044206	0.098508	0.085642	1.000000	0.034495	−0.009929	0.024158	0.067546	0.004190	0.023849	0.006121	−0.023730	0.117626	−0.058255	0.046589	−0.008366	0.031775	0.040585	−0.022115	0.090321	0.001510	0.014259
BMI	0.052904	0.213748	0.106722	0.034495	1.000000	0.013804	0.020153	0.224379	−0.147294	−0.087518	−0.062275	−0.048736	−0.018471	0.058206	0.239185	0.085310	0.121141	0.197078	0.042950	−0.036618	−0.103932	−0.100069
Smoker	0.114441	0.096991	0.091299	−0.009929	0.013804	1.000000	0.061173	0.062914	−0.087401	−0.077666	−0.030678	0.101619	−0.023251	0.048946	0.163143	0.092196	0.116460	0.122463	0.093662	0.120641	−0.161955	−0.123937
Stroke	0.203002	0.129575	0.092620	0.024158	0.020153	0.061173	1.000000	0.107179	−0.069151	−0.013389	−0.041124	−0.016950	0.008776	0.034804	0.177942	0.070172	0.148944	0.176567	0.002978	0.126974	−0.076009	−0.128599
Diabetes	0.180272	0.271596	0.209085	0.067546	0.224379	0.062914	0.107179	1.000000	−0.121947	−0.042192	−0.058972	−0.057882	0.015410	0.035436	0.302587	0.073507	0.176287	0.224239	0.031040	0.185026	−0.130517	−0.171483
PhysActivity	−0.087299	−0.125267	−0.078046	0.004190	−0.147294	−0.087401	−0.069151	−0.121947	1.000000	0.142756	0.153150	0.012392	0.035505	−0.061638	−0.266186	−0.125587	−0.219230	−0.253174	0.032482	−0.092511	0.199658	0.198539
Fruits	−0.019790	−0.040555	−0.040859	0.023849	−0.087518	−0.077666	−0.013389	−0.042192	0.142756	1.000000	0.254342	−0.035288	0.031544	−0.044243	−0.103854	−0.068217	−0.044633	−0.048352	−0.091175	0.064547	0.110187	0.079929
Veggies	−0.039167	−0.061266	−0.039874	0.006121	−0.062275	−0.030678	−0.041124	−0.058972	0.153150	0.254342	1.000000	0.021064	0.029584	−0.032232	−0.123066	−0.058884	−0.064290	−0.080506	−0.064765	−0.009771	0.154329	0.151087
HvyAlcoholConsump	−0.028991	−0.003972	−0.011543	−0.023730	−0.048736	0.101619	−0.016950	−0.057882	0.012392	−0.035288	0.021064	1.000000	−0.010488	0.004684	−0.036724	0.024716	−0.026415	−0.037668	0.005740	−0.034578	0.023997	0.053619
AnyHealthcare	0.018734	0.038425	0.042230	0.117626	−0.018471	−0.023251	0.008776	0.015410	0.035505	0.031544	0.029584	−0.010488	1.000000	−0.232532	−0.040817	−0.052707	−0.008276	0.007074	−0.019405	0.138046	0.122514	0.157999
NoDocbcCost	0.031000	0.017358	0.013310	−0.058255	0.058206	0.048946	0.034804	0.035436	−0.061638	−0.044243	−0.032232	0.004684	−0.232532	1.000000	0.166397	0.192107	0.148998	0.118447	−0.044931	−0.119777	−0.100701	−0.203182
GenHlth	0.258383	0.300530	0.208426	0.046589	0.239185	0.163143	0.177942	0.302587	−0.266186	−0.103854	−0.123066	−0.036724	−0.040817	0.166397	1.000000	0.301674	0.524364	0.456920	−0.006091	0.152450	−0.284912	−0.370014
MentHlth	0.064621	0.056456	0.062069	−0.008366	0.085310	0.092196	0.070172	0.073507	−0.125587	−0.068217	−0.058884	0.024716	−0.052707	0.192107	0.301674	1.000000	0.353619	0.233688	−0.080705	−0.092068	−0.101830	−0.209806
PhysHlth	0.181698	0.161212	0.121751	0.031775	0.121141	0.116460	0.148944	0.176287	−0.219230	−0.044633	−0.064290	−0.026415	−0.008276	0.148998	0.524364	0.353619	1.000000	0.478417	−0.043137	0.099130	−0.155093	−0.266799
DiffWalk	0.212709	0.223618	0.144672	0.040585	0.197078	0.122463	0.176567	0.224239	−0.253174	−0.048352	−0.080506	−0.037668	0.007074	0.118447	0.456920	0.233688	0.478417	1.000000	−0.070299	0.204450	−0.192642	−0.320124
Sex	0.086096	0.052207	0.031205	−0.022115	0.042950	0.093662	0.002978	0.031040	0.032482	−0.091175	−0.064765	0.005740	−0.019405	−0.044931	−0.006091	−0.080705	−0.043137	−0.070299	1.000000	−0.027340	0.019480	0.127141
Age	0.221618	0.344452	0.272318	0.090321	−0.036618	0.120641	0.126974	0.185026	−0.092511	0.064547	−0.009771	−0.034578	0.138046	−0.119777	0.152450	−0.092068	0.099130	0.204450	−0.027340	1.000000	−0.101901	−0.127775
Education	−0.099600	−0.141358	−0.070802	0.001510	−0.103932	−0.161955	−0.076009	−0.130517	0.199658	0.110187	0.154329	0.023997	0.122514	−0.100701	−0.284912	−0.101830	−0.155093	−0.192642	0.019480	−0.101901	1.000000	0.449106
Income	−0.141011	−0.171235	−0.085459	0.014259	−0.100069	−0.123937	−0.128599	−0.171483	0.198539	0.079929	0.151087	0.053619	0.157999	−0.203182	−0.370014	−0.209806	−0.266799	−0.320124	0.127141	−0.127775	0.449106	1.000000

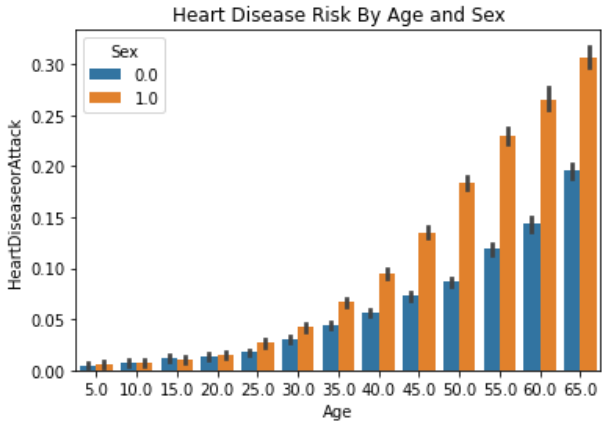
Heart Attack Counts

The count plot suggests to us that heart attack disease may not happen too frequently in all populations, and there might be some sort of lack of getting heart attack samples in the future fitted model.



Heart Disease Risk by Age and Gender

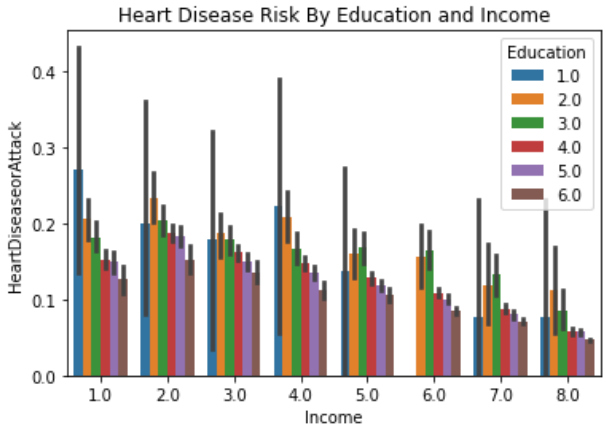
In the histogram below, we can see that as age increases, the heart disease risk difference between males and females is getting larger and larger. Men's diseases compared to females are about the same when the age is from 5 to 20. Age above 25, males tend to have more heart disease attacks than females of the same age.



Heart Disease Risk by Education and income

And for the heart disease risk by education and income, we can see from the graph that as the income increase, the heart disease risk decrease. This seems is similar to the education level. The more education they get, the less heart disease risk they have. This also appears in all different age levels. However, compared with the income and education at different levels, we

can conclude that even though we have a low income, if we have a high education level, we still have a low chance to get heart disease problems. In this way, the education level may be a more important factor compared with the income. Hence, for the income at 1 and education at 1, this seems to have the largest probability to have heart disease.

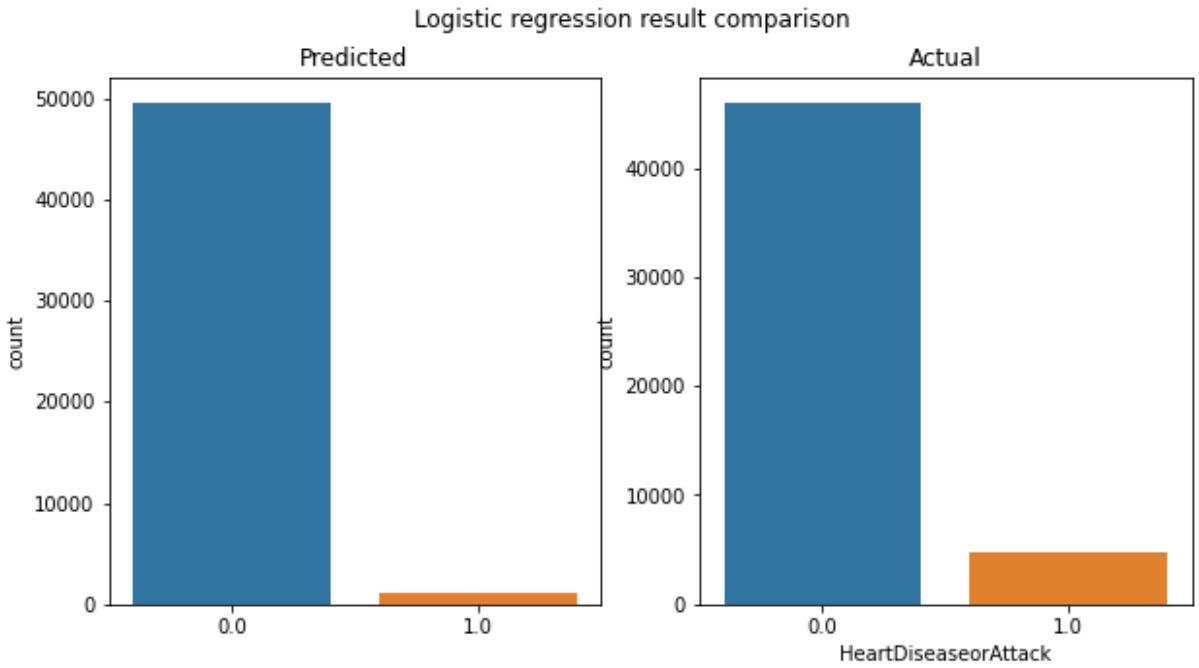


Methodology:

The potential tools that we use to predict the outcome of whether a person will get a heart attack or related disease are Logistic regression, Support vector machines(SVM), and k-nearest neighbors (KNN). In order to do the machine learning, we split data to two portions, one is the predictor and another is the outcome. Then we split each of them to train and test set by 80/20 ratio.

Logistic regression:

After that, we use the logistic regression model to predict a fit line for this training dataset with x and y. Besides, we also add up points along the fit line to create the regression model.

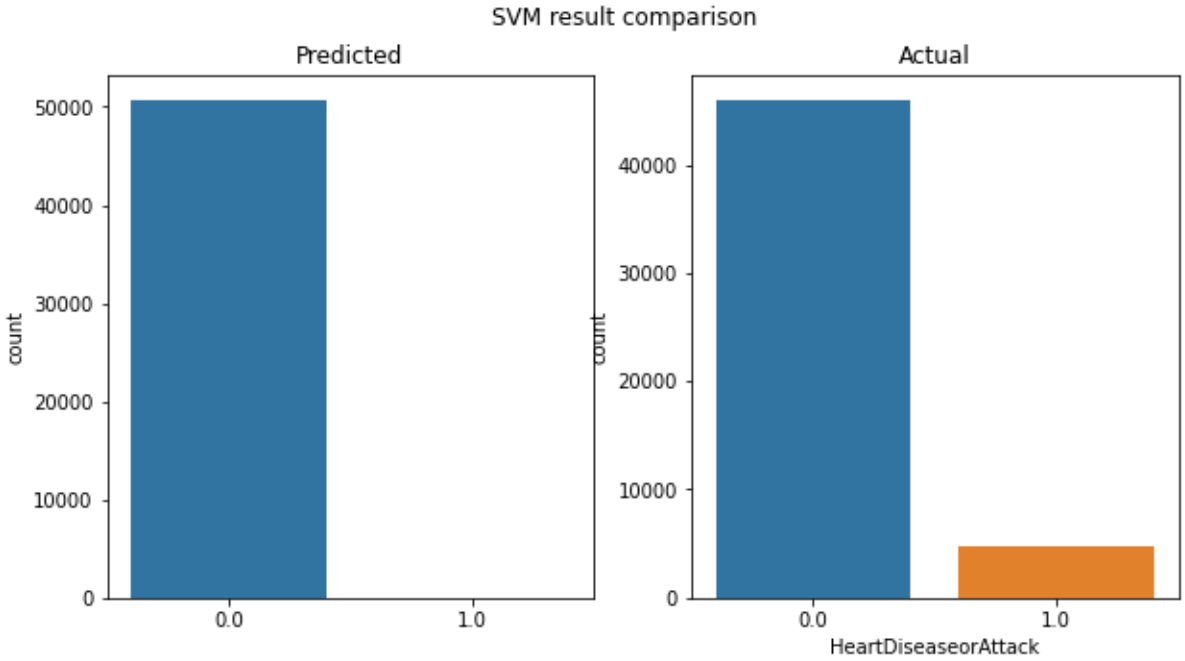


Also, as we can see from the graph, we can see the difference between before prediction and after prediction. The number of people who got heart attacks is relatively accurate. The predicted patients are around two thousand, and the actual number of patients is around four thousand.

SVM:

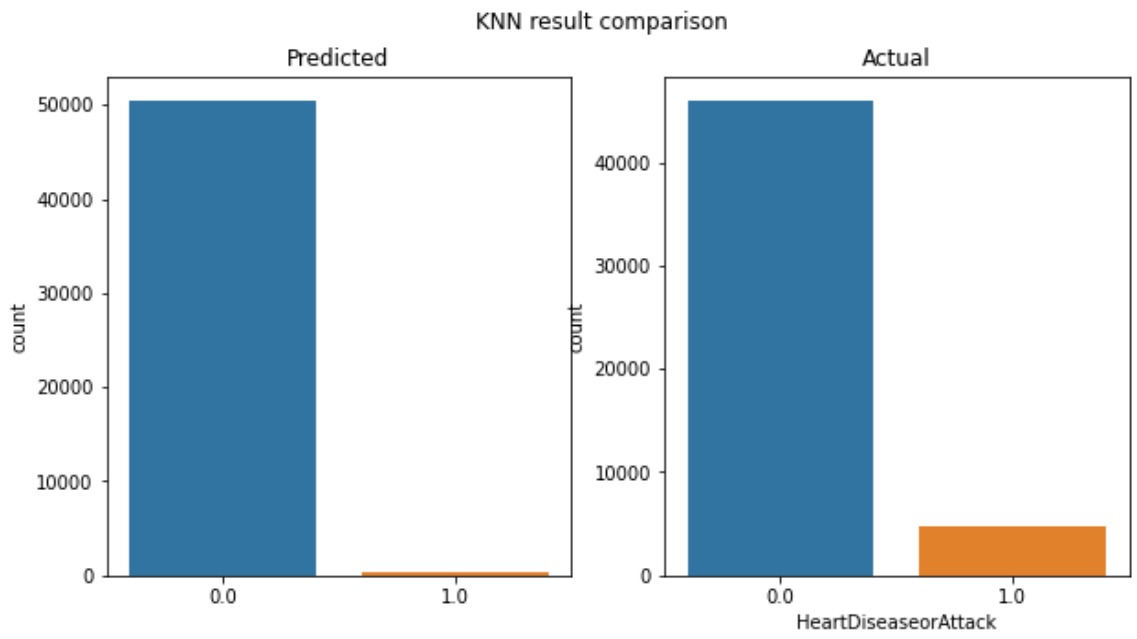
This method shows the distinction and relationship between those variables. In this way, not only can we see how those variables separate each other, but we also can see whether there is some wrong prediction or not. Besides, this method also can give us an accuracy score. Hence, this provides additional information to see the accuracy of the prediction. Therefore, this method assists with the preview method of the prediction of a heart attack. This method improves and proves the idea of the relationship between the other 21 variables and this heart attack variable. As we also can see from the graph, we before predict are close to zero, and the actual are close to 5000. Hence, this prediction is not precise when we predict, kind of different

from the actual.



KNN:

For this part, we use the K Neighbors Classification to find the close relationship between the heart attack variable to other variables. This method's idea is that we first calculate the distance from these target variables to others. Then, we find the closest one. Also, with this method, we also can find an accuracy score to see whether the prediction for the relationship of the target variable is good or not. We also can compare those multiply methods to determine which one can lead to the best prediction. This is similar to the method that we use in SVM for this method. before the predictions are precise compared with the actual data. Hence, after the prediction is less than 1000. And actual is also close to 5000.



Result:

From the comparison plots above, we know that every model is doing a fantastic job of predicting the number of non-CVD-related cases, however, only the logistic regression model predicts relatively closer numbers of heart attack cases. The accuracy score shows the same trends. The accuracy score for using SVM and KNN are 0.906 and 0.9056 respectively, and for logistic regression, the number is a little bit higher to 0.9095, which is not significant but distinct from all three models.

Conclusion:

We use some libraries provided by Python to implement this project. We checked the correlations between all predictors, plotted the heatmap, and made graphics to show the possible trends between predictors and outcomes. After the experiments, the Logistic regression gives us the best test accuracy. Though we get a good result of 90.95% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. There are some possible methods we could use to make a better model. Since logistic regression comes out as the best prediction, we may adjust the model by using the stepwise method to choose the most related variables to reduce the model. Then we may get fewer variables and better results.