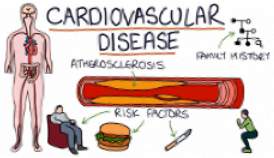
Cardiovascular Disease prediction



Problem Statement

More people are suffering from cardiovascular disease, so I want to make a model that can help people predict their risk of having cardiovascular disease and get examination earlier if they have a high risk.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	-0.436071	2	0.443399	-0.847978	-0.122181	-0.088254	1	1	0	0	1	0
1	0.307835	1	-1.018016	0.749564	0.072577	-0.035205	3	1	0	0	1	1
2	-0.248017	1	0.078045	-0.709062	0.007658	-0.141303	3	1	0	0	0	1
3	-0.748143	2	0.565184	0.541160	0.137496	0.017844	1	1	0	0	1	1
4	-0.808532	1	-1.018016	-1.264725	-0.187100	-0.194362	1	1	0	0	0	0

Data processes

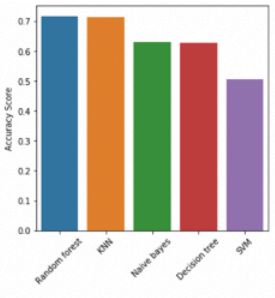
The dataset I choose consists of 70 000 records of patient data, 11 features + target. To improve the accuracy of my model, I clear the duplicated and missing data. Then, I standardize the first five data to ensure they are at the same level as the rest of six ones. I split the data, 80% to train my data and 20% for validation.

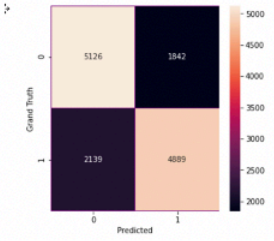
Linear SVC

The first model I used is Linear SVC, but the data set has too many features and the model is not working well. The accuracy is around 60%, merely higher than the bassline 50%.

Random Forest

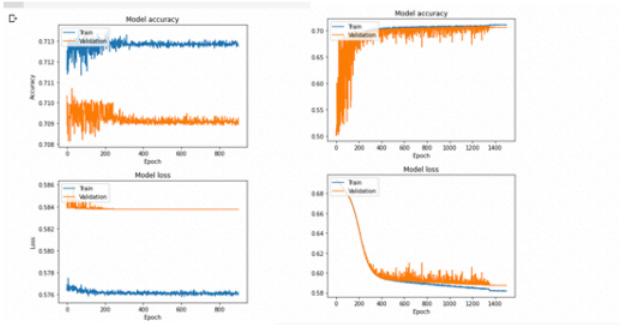
The second model I used is Random Forest. Compared to the first model, it improved a lot. At first, it has an accuracy of 70%. Then, from visualization of the model, I found that the model is overfitting. Thus, I limited the depth of the tree and raise the accuracy to 71.5%.





Model Comparison

Then I tried different models and compared their accuracies. The random forest is still the best score I can achieved. Picture 3 is its grade truth table.



ANN model

After assignment 4 learning the CNN, I decided to add the ANN model for my project. I learned how to visualize my model and found that the accuracy between the train set and the validation set has a clear gap. Then I adjusted the model to addressee the overfitting problem. The final accuracy is 71.7%, a little higher than what I achieved in random forest

Conclusion

This project helps me to review almost all the models I learned during this semester, and it gives me a whole picture of how different models perform in this problem. The most complicated model is not always the best model for a problem, and different problems need their most suitable model to get the best result.

Reference

[1] data set from Kaggle https://www.kaggle.com/sulianova/cardiovascular-disease-dataset/kernels

[2] the model choices are inspired by Ben AKCA, Cardiovascular Disease Prediction https://www.kaggle.com/benanakca/comparison-of-classification-disease-prediction