

# Capstone Project-3 Cardiovascular Risk Prediction

**By-Prabir Debnath** 



#### **Points of Discussion**

- 1. The Problem Statement
- 2. Concept of ML Model
- 3. Summary of the Experience Set
- 4. Exploration and Pre-processing of Data
- 5. Building and Evaluation of Model-1

- 6. Building and Evaluation of Model-2
- 7. Building and Evaluation of Model-3
- 8. Final Conclusion



#### **The Problem Statement**

We are provided with a labeled dataset on the details of patients with or without cardiovascular disease. The task is to explore and analyze the data and to build a classification model for 10 Years Coronary Heart Disease prediction.



#### **Concept of ML Model**

The performance of a machine learning model depends on three factors:

#### i. Quality of Data

(cleaner experiences for better learning)

#### ii. Quantity of Data

(more experiences for better learning)

#### iii. Quality of Model

(right model and right hyperparameters for better learning)



# **Summary of the Experience Set**

Here, the dataset has 3,390 rows, which means 3,390 experiences about patients with or without cardiovascular disease and

It has 17 columns, which means each experience is observed along 17 features or dimensions.



# Let's Decode the Experiences!

49.61	21.58	30.88	33.93	42.75	30.76	75.22	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02		35-25	76.68	67.06	52.25	41.20	35.31
45.73	19.58	22.53	32.58	35.98	28.27	31.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	35/25	14 E 134-	A 33	57.04	40.26	27.52	15.30
46.63	19.02	29.80	31.12		TV	30.95	63.01	48.00	34.47	28.92
53.76	27.32	41.23	37.99	ME 014		111	71.62	56.08	40.62	36.05
66.67	41.95	50.25	62.06	60 80	ALC	46	79.07	68.02	52.81	49.12
65.41	42.22	50.85	61.54	7700	UE	33.75	77.70	66.60	51.60	48.43
56.33		43.05	41.31	A TY	10 mg	75.65		59.13	42.93	39.85
	29.39	52.56	53.94	50,83	1		74.16			50.54
59.66	52.52	140000000000000000000000000000000000000	0.24	7	2775	87.66	64.60	62.80	56.05	- CONTROL OF THE
56.27	31.05	40.87	200	200		Park Street	72.93	59.22	43.56	39.45
54.32	29.70	39.30	19	49,03		July 2	70.45	57.22	41.40	37.53
49,61	21.58	30.86	241.73	E.75	10.70	18.22	66.61	50.29	36.73	30.76
52.25	28.31	38.77	177 PM	45.57		75.1	06	52.25	41.20	35.31
45.73	19.58	23,53	13/98	35.96		7	.95	47.39	30.16	28.27
40.51	12.74	14.24	A 47	36.48	150	67.3	₩ 04	40.26	27.52	15.30
46.63	19.02	29.907	12	19 M	21 2	70.9	5 .01	48.00	34.47	28.92
53.76	27.32	44.37	61/30		J.	6	7 62	56.08	40.62	36.05
66.67	41.95	90.25	82.04V	$\mathcal{I}$		1	7 07	68.02	52.81	49.12
65.41	42.22	60.65	61.5	1000	X		(1.1. Pa	66.60	51.60	48.43
56.33	29.39	村,09	41.317	11/	Date	# 5 TA	NV	59.13	42.93	39.85
59.66	52.52	52 E	53,91	7 2	1 No.	97.E		62.80	56.05	50.54
56.27	31.05	46.87	39 87	22.00	4	AT THE	11/1	59.22	43.56	39.45
54.32	29.70	39.36	38,15	49 93	本概	5.33	1.4/4	57.22	41.40	37.53
49.61	21.58	30.88	33.93	42.75	11/1/	# 1/22 )	est 1	NO.29	36.73	30.76
52.25	28.31	38.70	37.92	(地 5万	16.30	76,68	67 V	N2.25	41.20	35.31
45.73	19.58	22.53	32.50	35 6	Deck	71.11	53.56	47.39	30.16	28.27
40.51	12.74	19.24	48,87	3498	2 15 SO	87.00	57.04	40.26	27.52	15.30
46.63	19.02	29.80	\$1.12	48.00	28.92	10,05	163.01	48.00	34.47	28.92
53.76	27.32	41.23	57,95	48961	18.05	13/01	71.62	56.08	40.62	36.05
66.67	41.95	50.25	62.06	/ HO 34	9.12	85.46	79.07	68.02	52.81	49.12
65.41	42.22	50.65	61.55	24.30	48.43	83.75	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	39.85	75.65	74.16	59.13	42.93	39.85
59.66	52.52	52.56	53.91	57.83	50.54	67.66	64.60	62.80	56.05	50.54
56.27	31.05	40.87	39.87	51.06	39.45	77.79	72.93	59.22	43.56	39.45



## **Exploration and Pre-processing of Data**

We have done the exploration and pre-processing in seven steps to transform raw data into quality data for our ml model.

- Connection with the Data
- First Feelings of the Data
- 3. Deeper Understanding of the Data
- 4. Cleaning the Data
- 5. Treating Anomalies in the Data
- 6. Final Feature Selection from the Data
- 7. Preparation of Input and Output Data



#### **Cleaning the Data**

- > We have handled all null values in 'cigsPerDay', 'totChol', 'BMI', 'heartRate', 'glucose', 'education', and 'BPMeds' columns with imputation. Thus there is no loss of data.
- > We have encoded 'sex' column with two categories: F: 0, M: 1
- > We have encoded 'is\_smoking' column with two categories: NO:0, YES:1

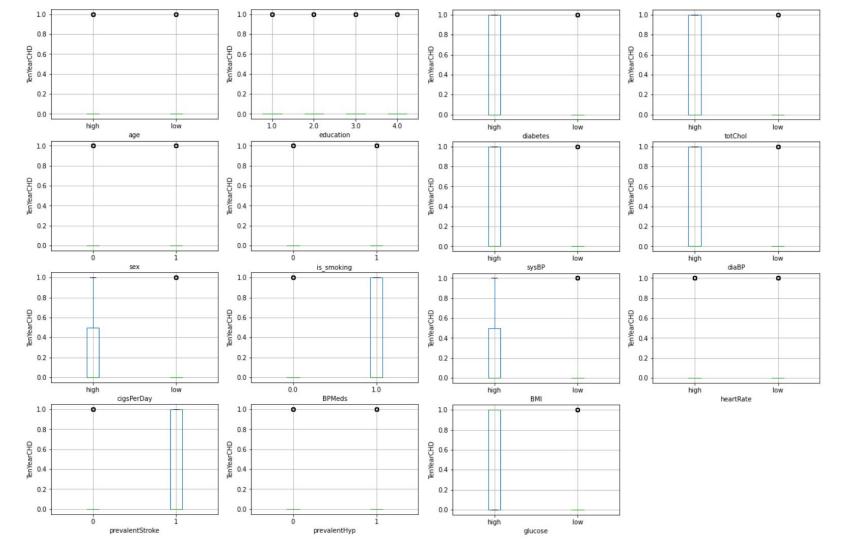


# **Treating Anomalies in the Data**

In our dataset, for most of the features class 1 targets are outliers. Thus we need more experience with class 1 targets to bring a balance in prediction.

In simple words, with the available experience set, our model will be expert in predicting the features for which there will be no heart disease.

Let's check the boxplot!



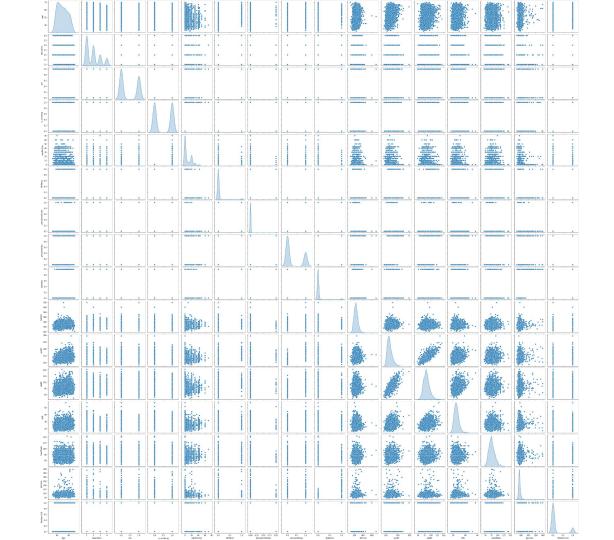




#### **Overall Feature Understanding**

Here, the distribution of 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate' and 'glucose' are positively skewed. Thus we have done log transformation on these features to normalize their distribution.

Let's check the pairplot!

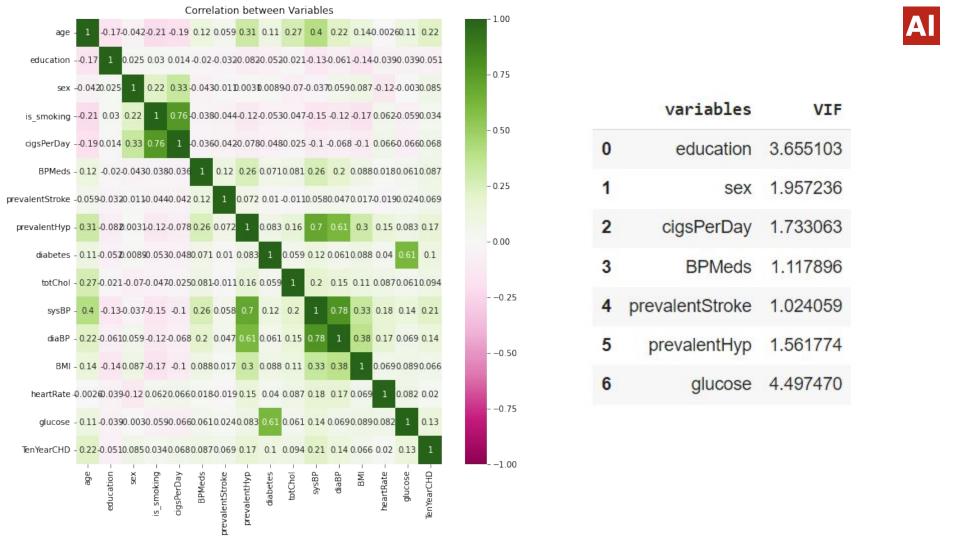




# **Looking for Truly Independent Features**

We have removed 'sysBP', 'diaBP', 'BMI', 'Age', 'heartRate', 'totChol', 'is\_smoking' and 'Diabetes' in sequence from our dataset to bring all the VIF values below 10. Thus all our input variables became truly independent.

Let's check the heatmap!





# Let's Train the Models!

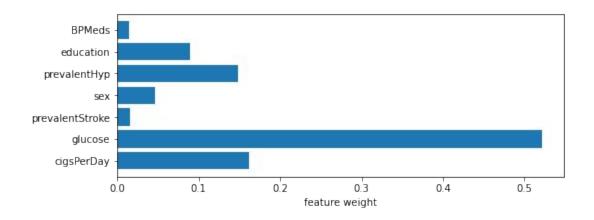




### **Building and Evaluation of Model-1**

#### **Final Random Forest Model:**

After cross validation and hyperparameter tuning, the best parameters are {'max\_depth': 25, 'max\_features': 'auto', 'min\_samples\_leaf': 5, 'min\_samples\_split': 15, 'n\_estimators': 100} (test accuracy is 84% and variance in prediction is 2%)

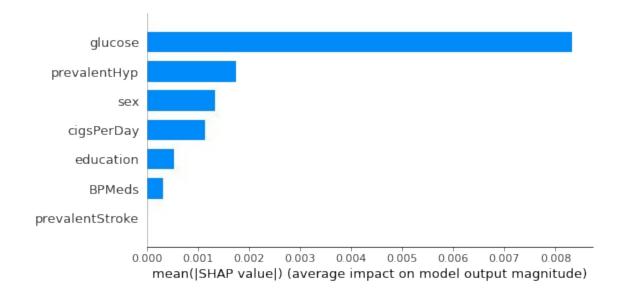




# **Building and Evaluation of Model-2**

#### **Final KNN Model:**

After cross validation and hyperparameter tuning, the best parameters are {'leaf\_size': 30, 'n\_neighbors': 19} (test accuracy is 84% and variance in prediction is 1%)

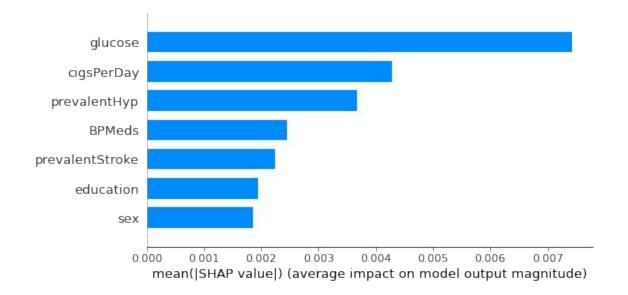




# **Building and Evaluation of Model-3**

#### **Final SVC Model:**

After cross validation and hyperparameter tuning, the best parameters are {'C': 6, 'gamma': 0.1} (test accuracy is 84% and variance in prediction is 2%)

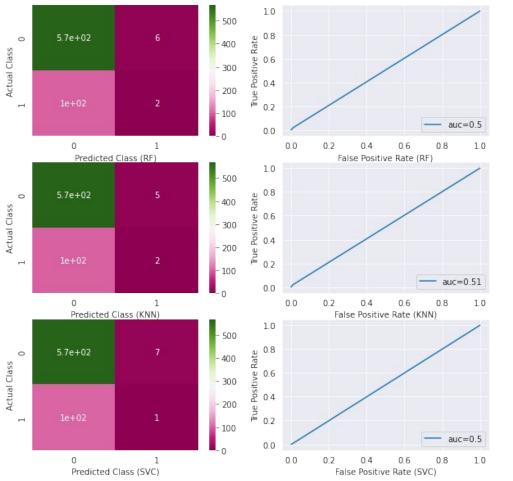


RF- Train\_Accuracy: 0.86, Test\_Accuracy: 0.84, Test\_F1: 0.47

KNN- Train\_Accuracy: 0.85, Test\_Accuracy: 0.84, Test\_F1: 0.48

SVC- Train\_Accuracy: 0.86, Test\_Accuracy: 0.84, Test\_F1: 0.46





#### **Final Conclusion**

On the basis of the performance study of our three models, we are selecting KNN classifier (the best warrior) for predicting 10 Years Coronary heart disease, as it has low variance in prediction, good f1\_score and good ROC\_AUC score among all three models

