Capstone Project – 3

Cardiovascular Disease Risk Prediction

by - Puneet Suthar

What are we talking about ?

Content :

- Introduction to Cardiovascular diseases
- Problem Statement
- General overview of the dataset
- Exploratory Data Analysis
- Model building
- Conclusion

Introduction

Cardiovascular diseases, also known as CVDs, are responsible for the highest number of fatalities worldwide, resulting in an estimated 17.9 million deaths annually. This group of disorders affects the heart and blood vessels, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other ailments. Heart attacks and strokes account for over 80% of all CVD-related deaths, and a significant portion of these fatalities occur prematurely in individuals under the age of 70. Unhealthy diet, physical inactivity, tobacco usage, and excessive alcohol consumption are the primary behavioral risk factors for heart disease and stroke. These risk factors may cause high blood pressure, high blood glucose, high blood lipids, overweight, and obesity in individuals.

Problem Statement

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD) or not. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

How our dataset Look like ?

General overview of the dataset

17 Columns

&

3390 Rows

General overview of the dataset

17 Columns

1. Sex : male or female ("M" or "F")
2. Education : 1,2,3,4 it could be the level of education , higher no. means higher edu.
3. Age : Age of the patient (Continuous Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
4. is_smoking : whether or not the patient is a current smoker.
5. Cigs Per Day : the number of cigarettes that the person smoked on average in one day. (Can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
6. BP Meds : whether or not the patient was on blood pressure medication (Nominal)

General overview of the dataset

17 Columns

7. Prevalent Stroke : whether or not the patient had previously had a stroke (Nominal)
8. Prevalent Hyp : whether or not the patient was hypertensive (Nominal)
9. Diabetes : whether or not the patient had diabetes (Nominal)
10. Tot Chol : total cholesterol level (Continuous)
11. Sys BP : systolic blood pressure (Continuous)
12. Dia BP : diastolic blood pressure (Continuous)
13. BMI : Body Mass Index (Continuous)
14. Heart Rate : heart rate (Continuous -In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of a large number of possible values.)

General overview of the dataset

17 Columns

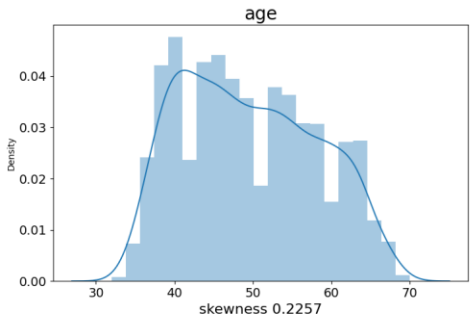
15. Glucose : glucose level (Continuous)
16. TenyearCHD :10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) – DV)
17. Id : unique id for every entry in the data

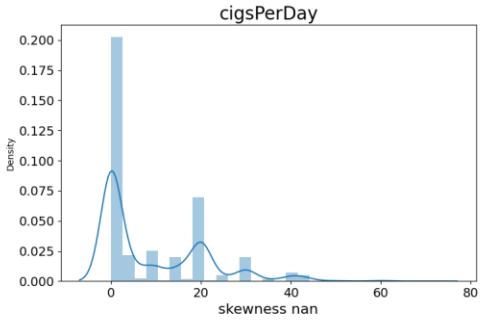
EDA
(Exploratory Data Analysis)

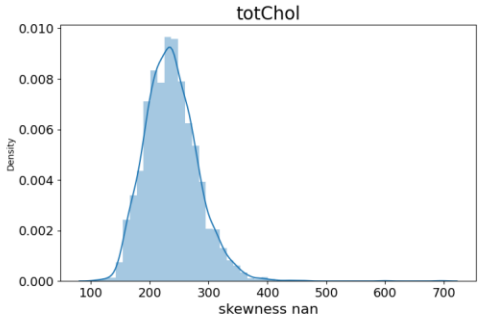
Distribution plot & skewness

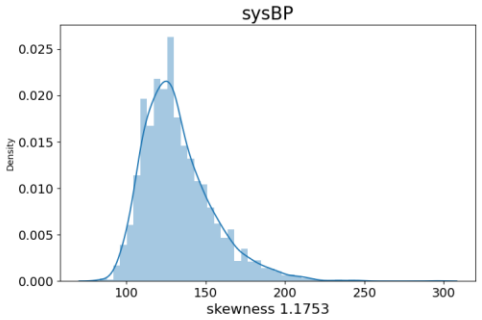
EDA
(Exploratory Data Analysis)

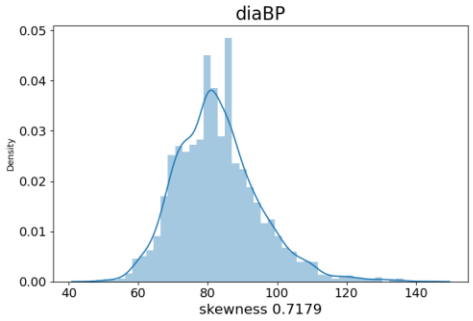
AI

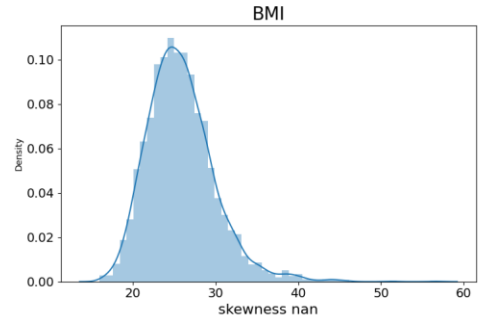


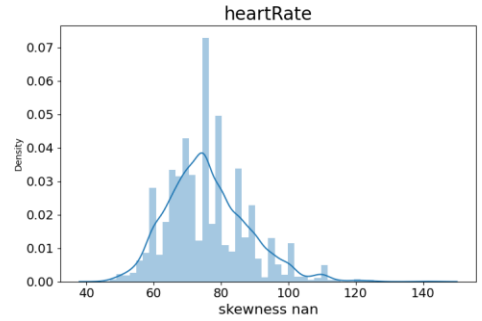


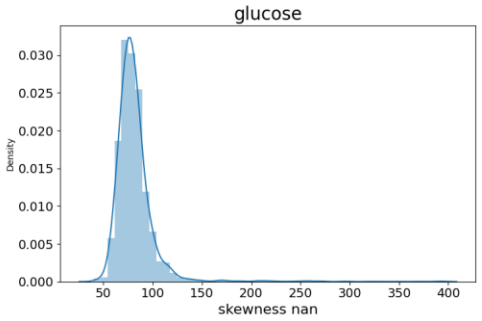












1. From above graph we can see that some of features are highly skewed which can effect our model.
2. First we check for outlier, if still skewness is not solved the we apply some transformation on them to normalize them.

Outliers

EDA
(Exploratory Data Analysis)

AI

```
num_col = ['age','cigsPerDay','totChol', 'sysBP','diaBP', 'BMI', 'heartRate', 'glucose']

wskr_df =  whiskers_df(df,num_col)
```

```
Total No. of rows in dataframe - 3390
Percentage of total outliers in data - 4.81%
```

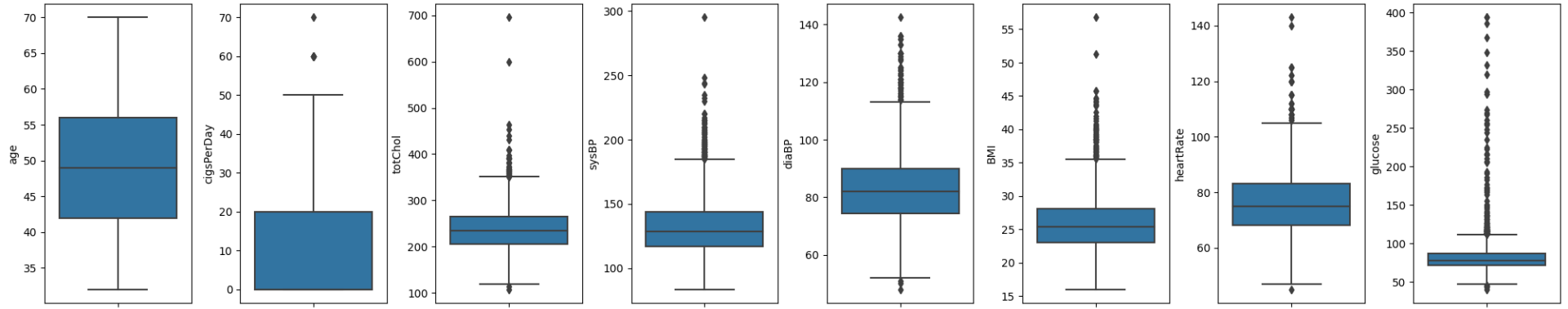
feature	upper_whisker	data_above_upper_whisker	indices_above_upper_whisker	lower_whisker	data_below_lower_whisker	indices_below_lower_whisker	outlier_%
age	77.00	0	[]	21.00	0	[]	0.00
cigsPerDay	NaN	0	[]	NaN	0	[]	0.00
totChol	NaN	0	[]	NaN	0	[]	0.00
sysBP	184.50	105	[6, 10, 37, 71, 163, 168, 190, 324, 359, 413, ...	76.50	0	[]	3.10
diaBP	113.25	55	[5, 6, 10, 14, 168, 171, 190, 324, 478, 684, 7...	51.25	3	[783, 1339, 3373]	1.71
BMI	NaN	0	[]	NaN	0	[]	0.00
heartRate	NaN	0	[]	NaN	0	[]	0.00
glucose	NaN	0	[]	NaN	0	[]	0.00

Box Plot

EDA
(Exploratory Data Analysis)

AI

Before outlier treatment

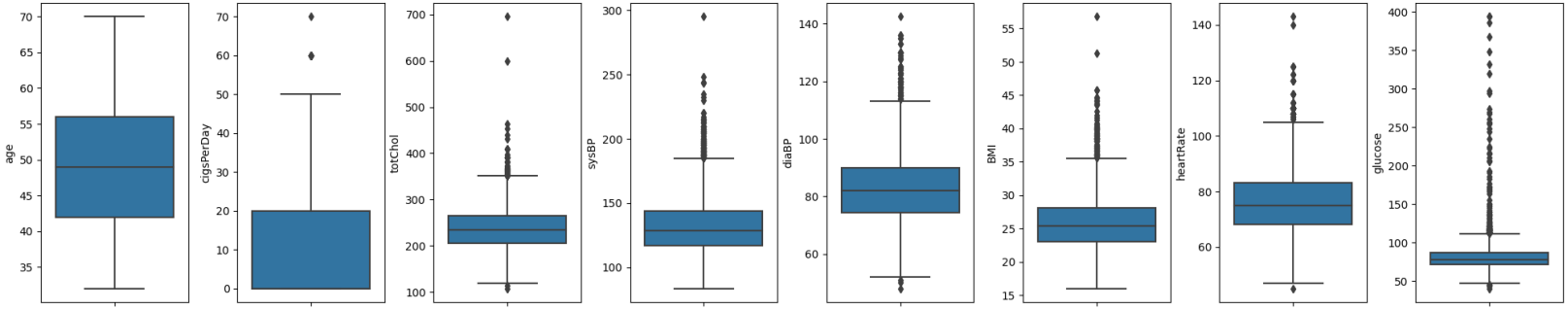


Box Plot

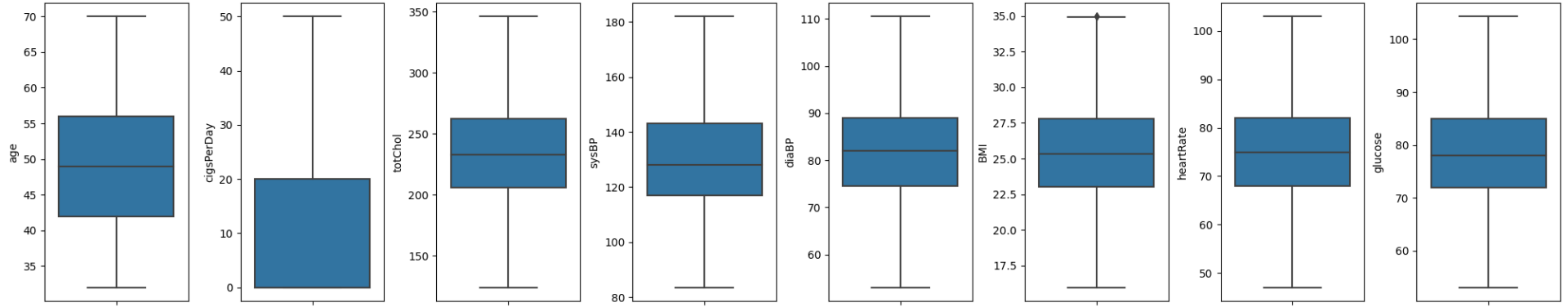
EDA
(Exploratory Data Analysis)

AI

Before outlier removal

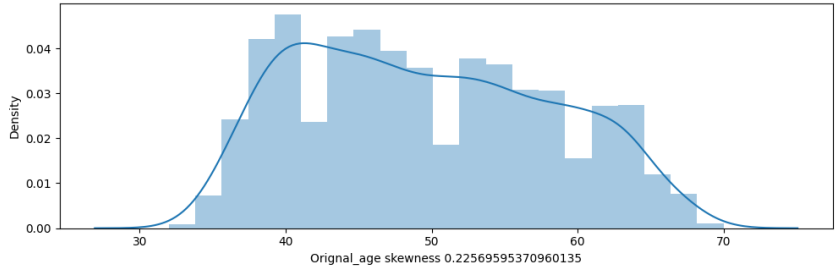


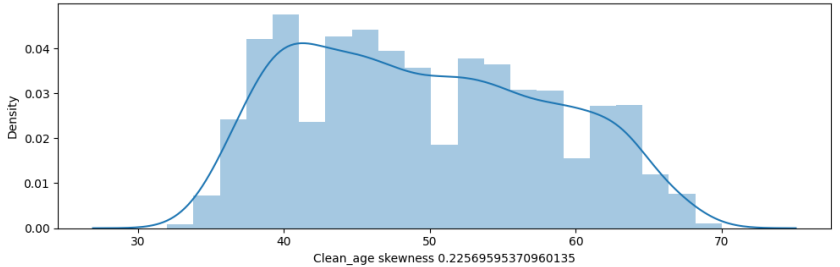
After outlier removal using KNN Imputation

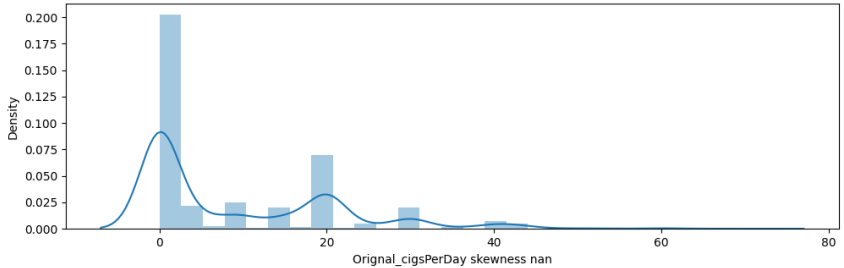


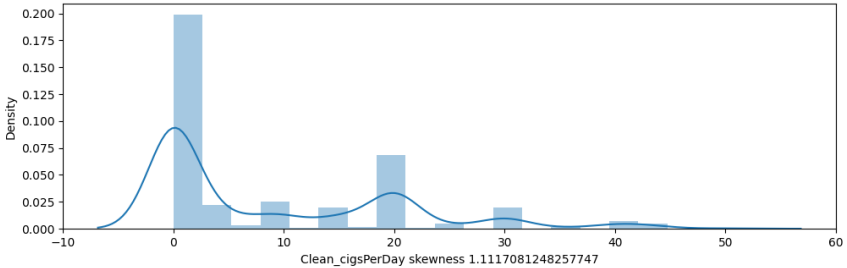
Change in skewness after outliers

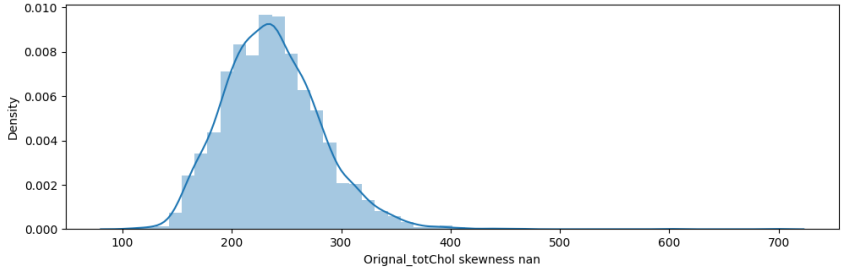
EDA
(Exploratory Data Analysis)

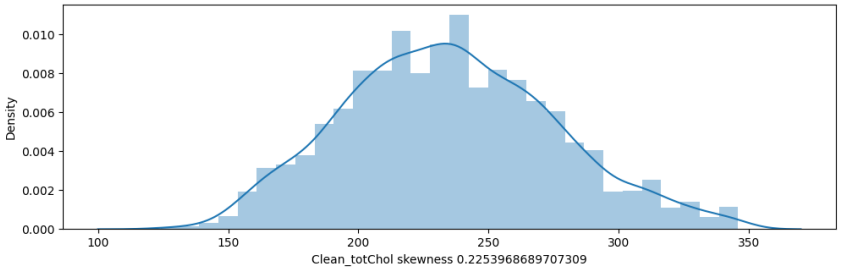






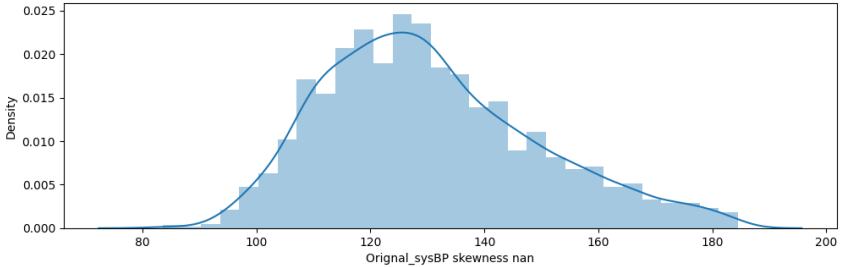


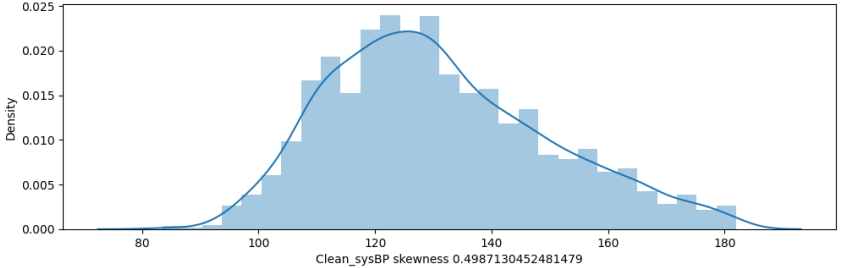


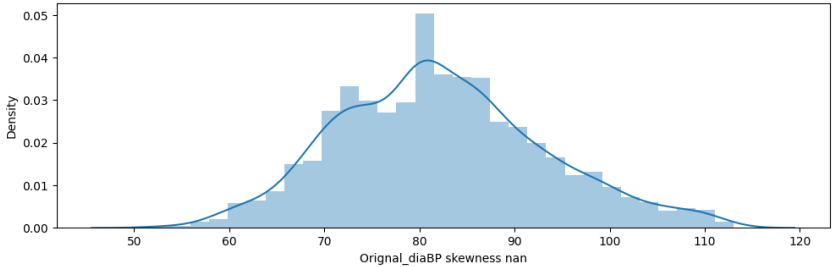


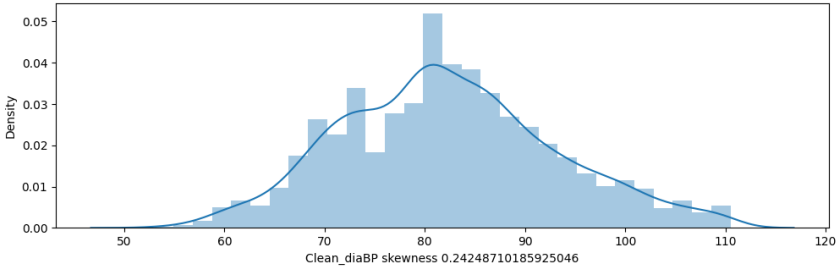
Change in skewness after outliers

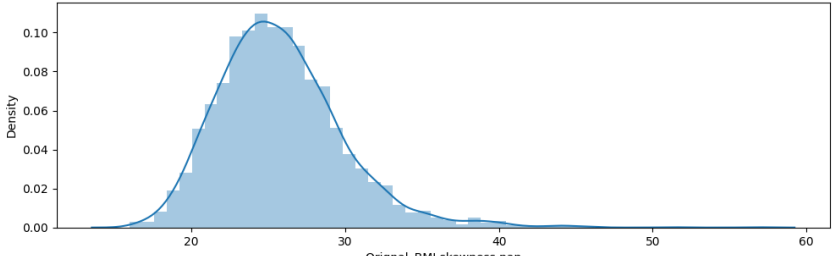
EDA
(Exploratory Data Analysis)

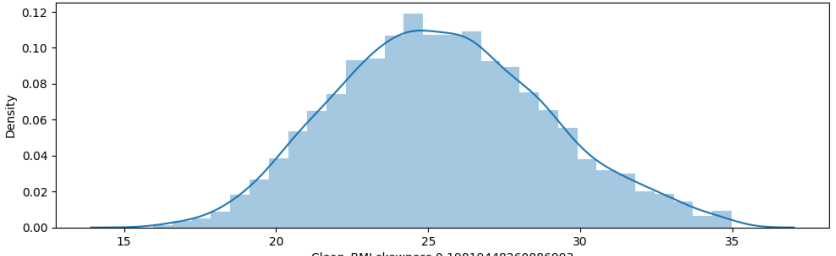






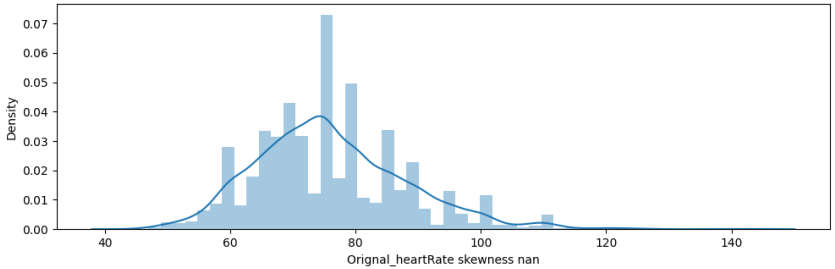


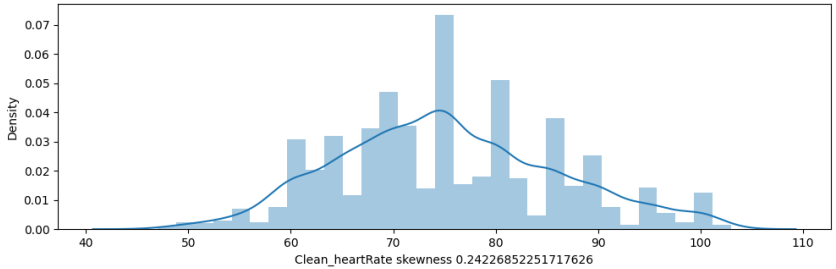


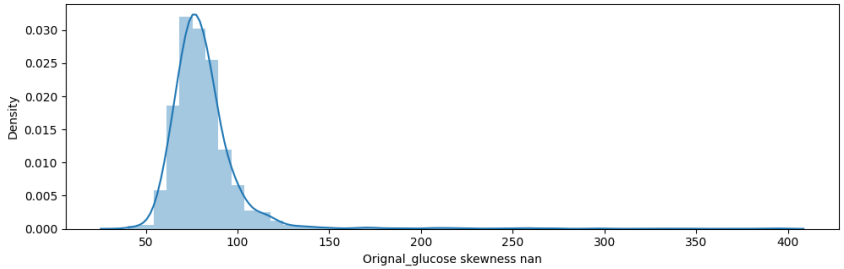


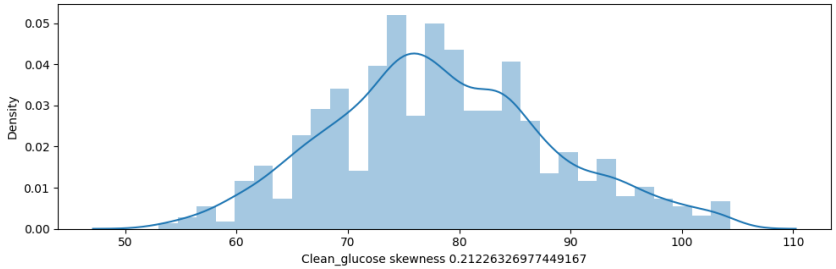
Change in skewness after outliers

EDA
(Exploratory Data Analysis)





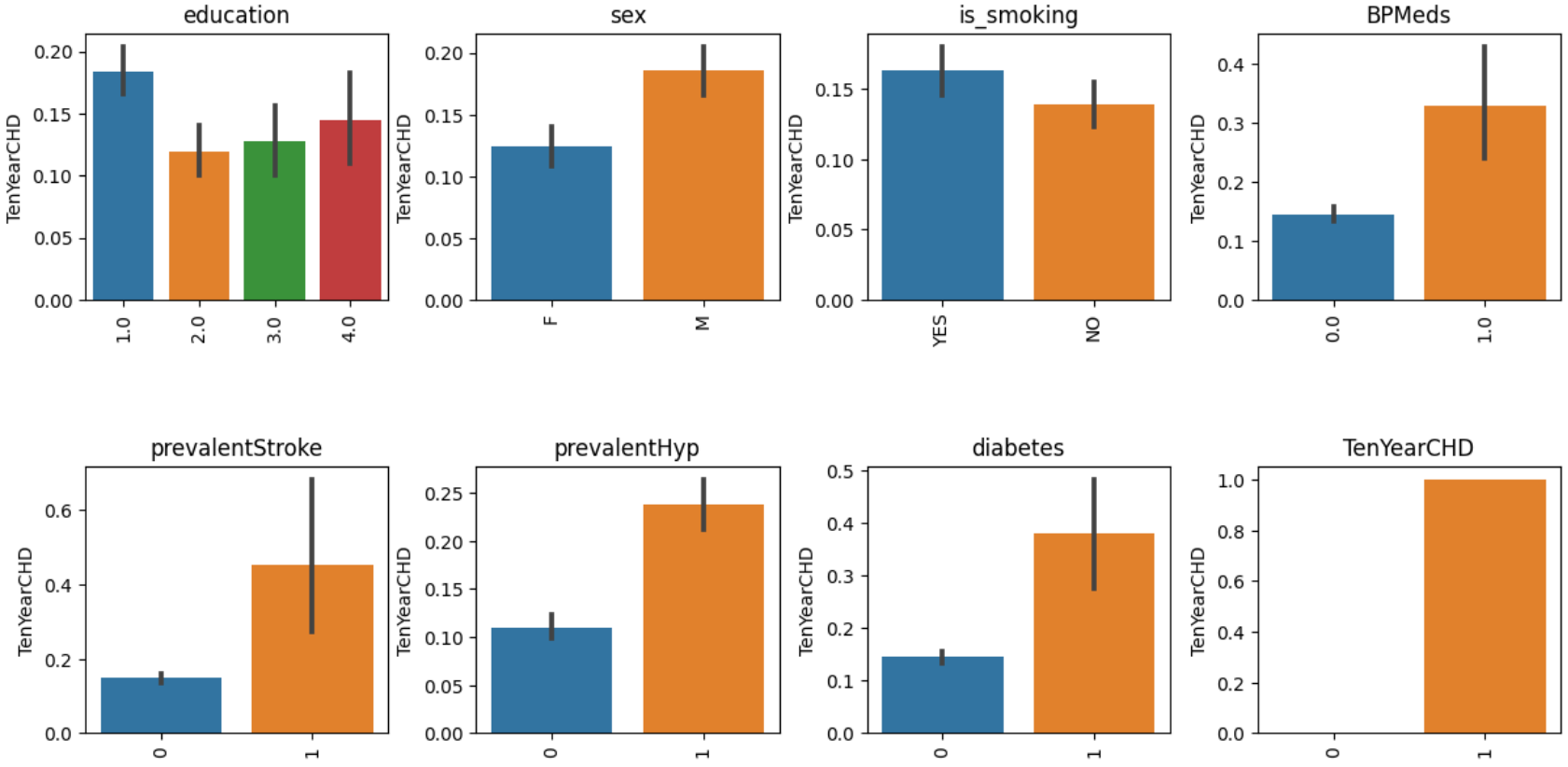




Bar Plot

EDA
(Exploratory Data Analysis)

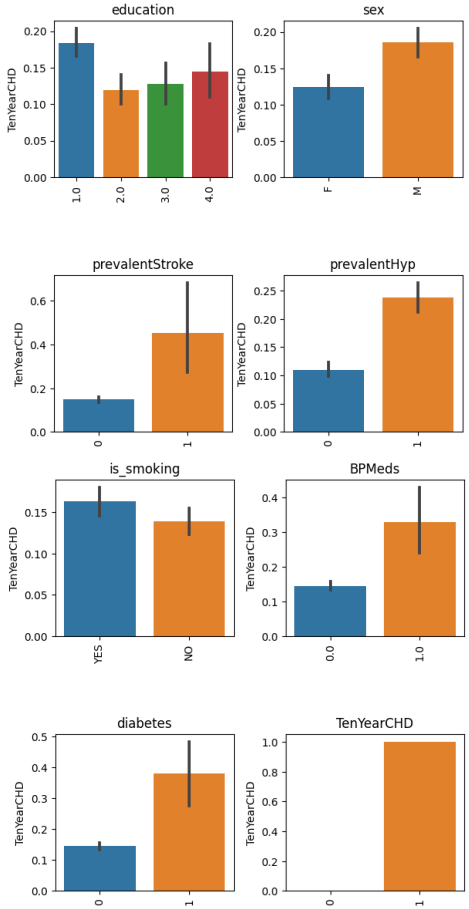
AI



Bar Plot

EDA
(Exploratory Data Analysis)

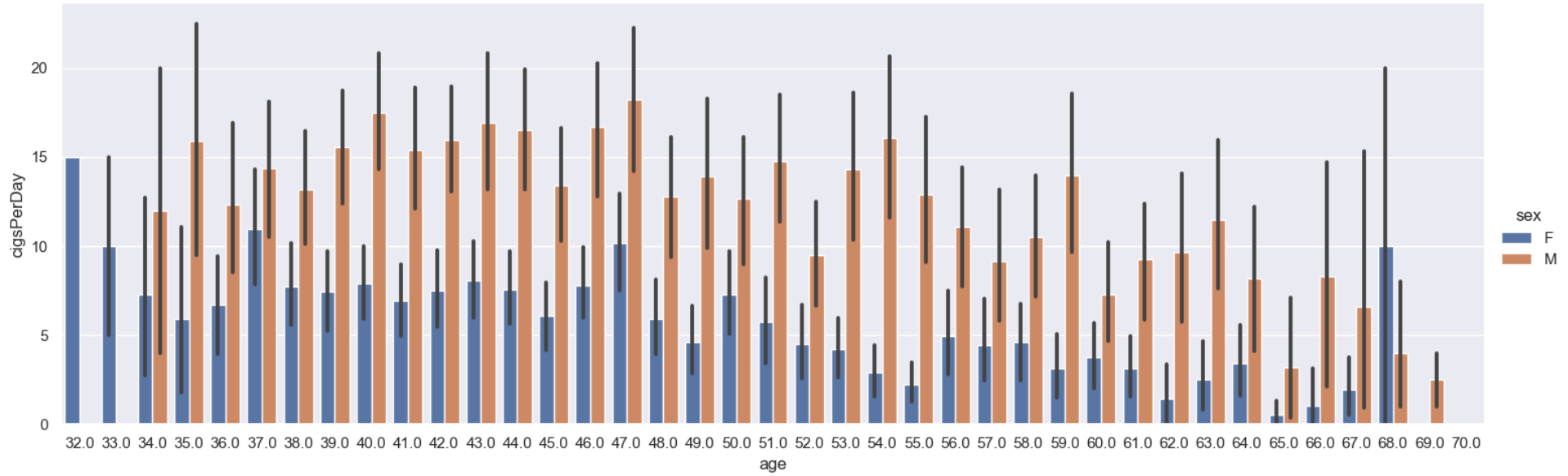
1. Less educated people have high percentage of CDH.
2. Male are more prone to CHD than female.
3. Smokers have high ratio of CHD than Non smokers.
4. Patients having BPmeds are more prone to CHD.
5. People have prevalant stroke have very high chances of CHD.
6. People have hypertension have very high chances of CHD.
7. People have diabetes have very high chances of CHD.



Bar Plot

EDA
(Exploratory Data Analysis)

AI

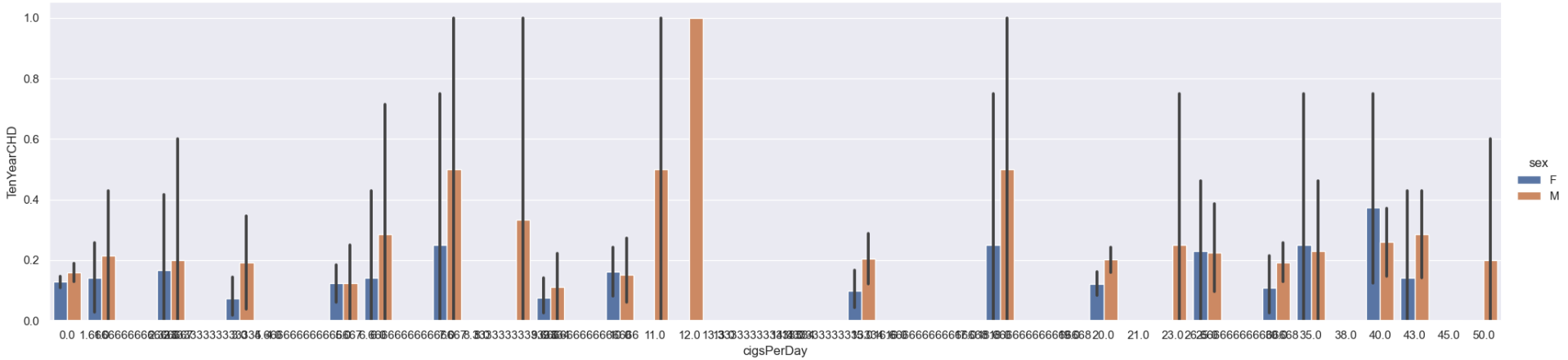


No. of cigrates per day is high amoung youngsters and it decreases as age increases but after 65 it increases again.

Bar Plot

EDA
(Exploratory Data Analysis)

AI

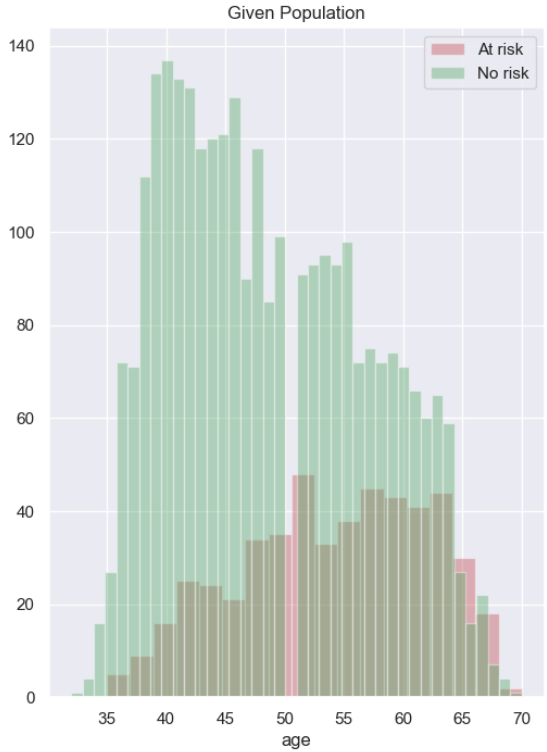


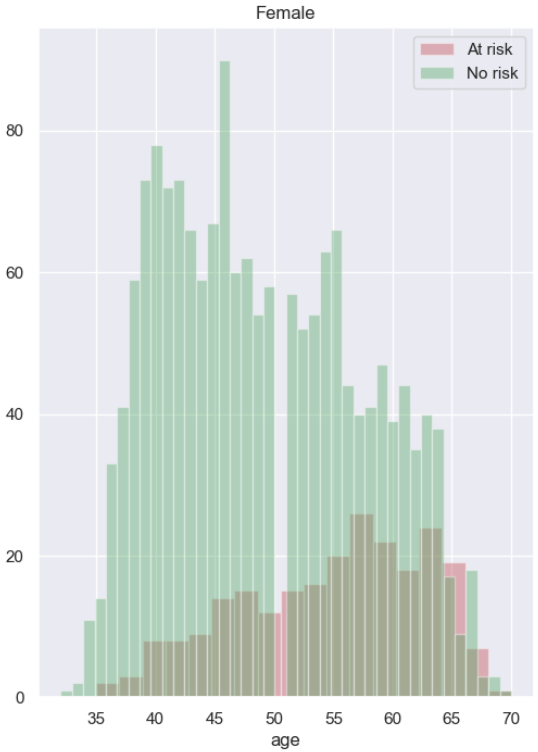
Person having highest risk who is consuming around 12 Cigs per day

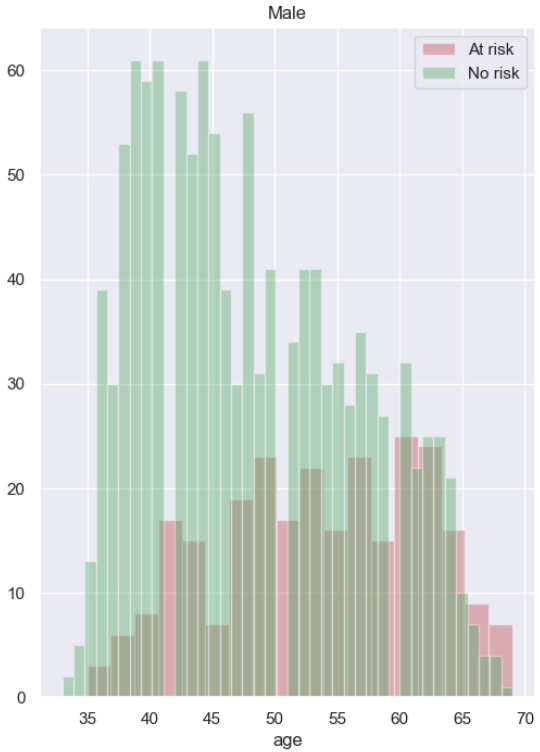
More Distribution plots

EDA
(Exploratory Data Analysis)

AI

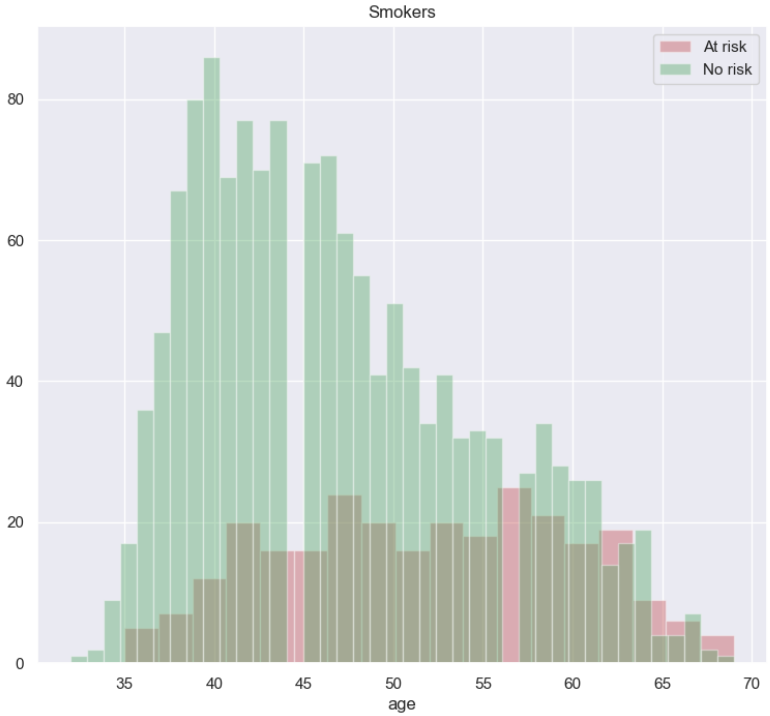


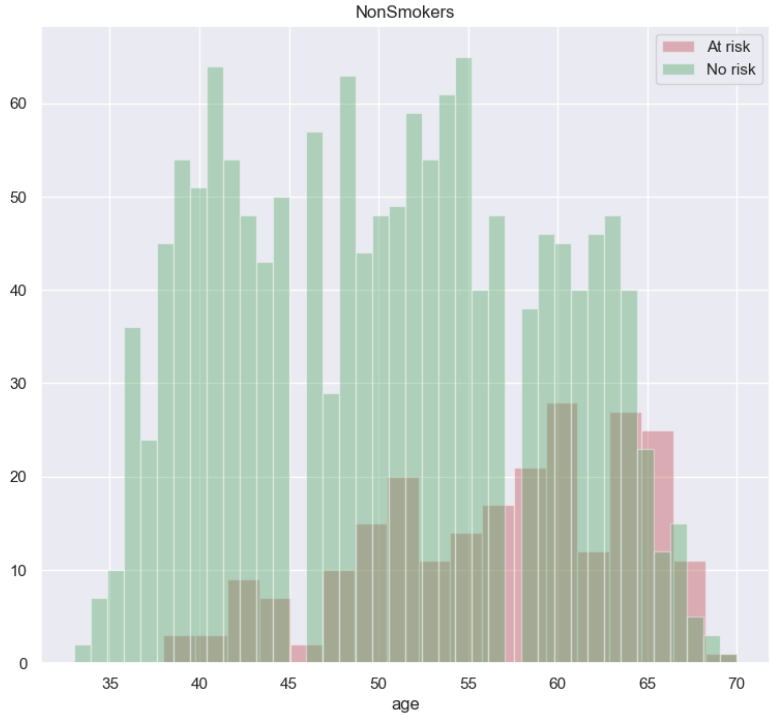




More Distribution plots

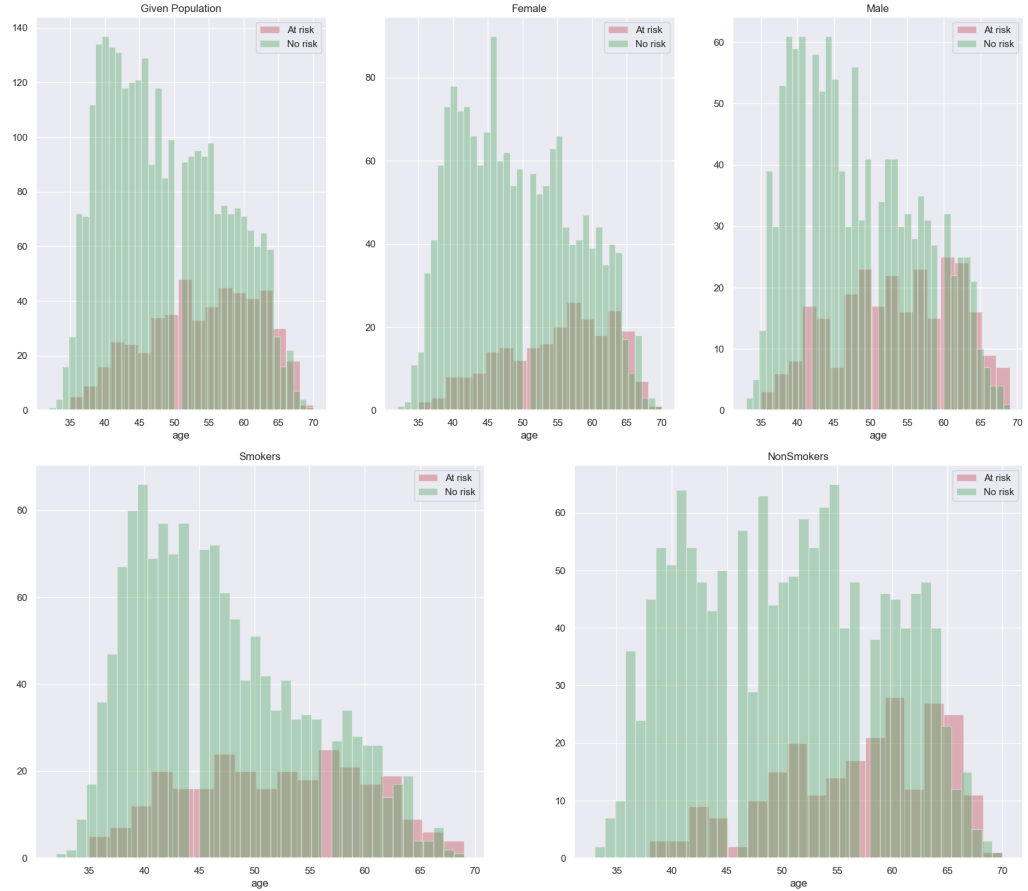
EDA
(Exploratory Data Analysis)





More Distribution plots

EDA
(Exploratory Data Analysis)



1. Both Women and Men lying in Age group of 50-52 have high risk of heart disease.

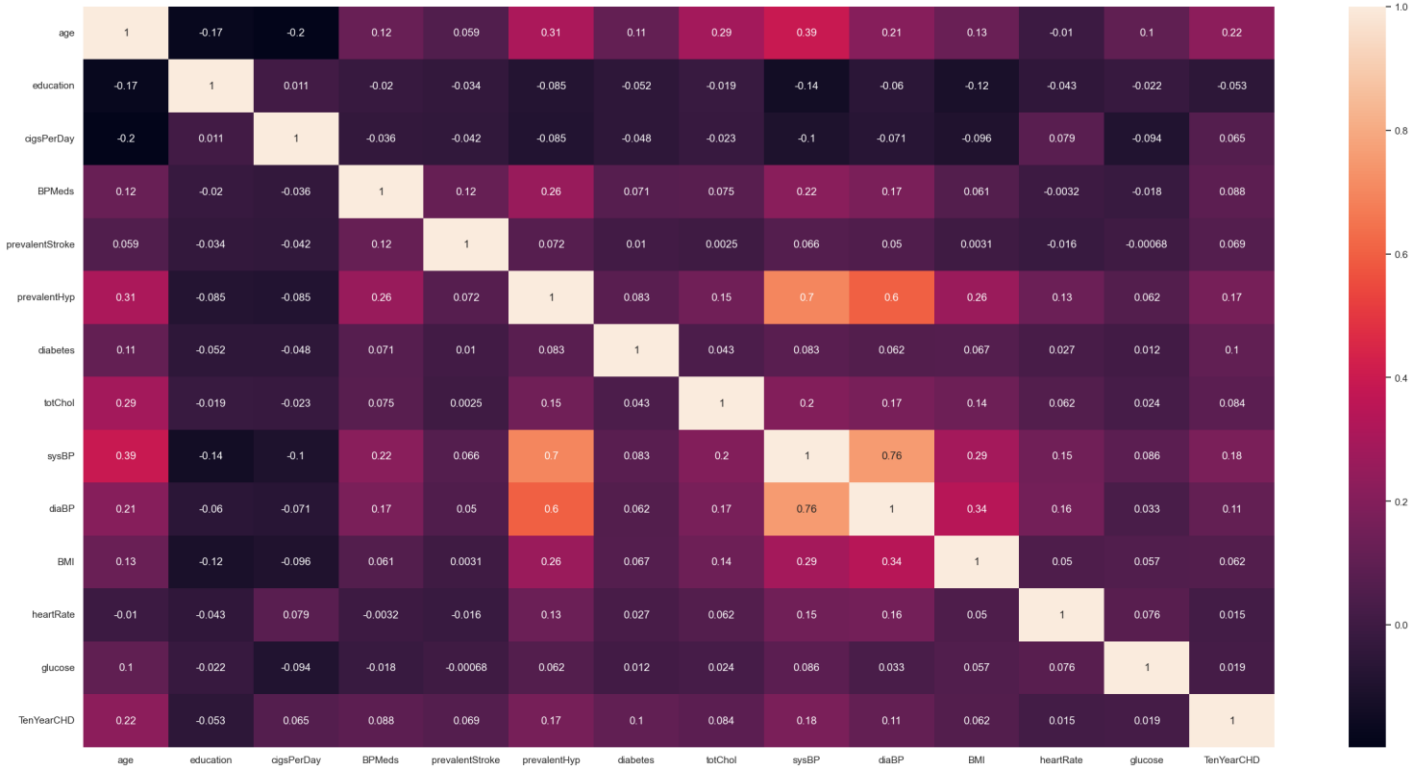
2. Men lying in age group 40-42 are at risk.

3. Men having age more than 65 are also at risk.

4. Risk is High in same age group despite they are Smokers or not.

Correlation heatmap

EDA
(Exploratory Data Analysis)



Sys BP and Dia BP is correlated to hypertension, and also correlated with each other

Model Building

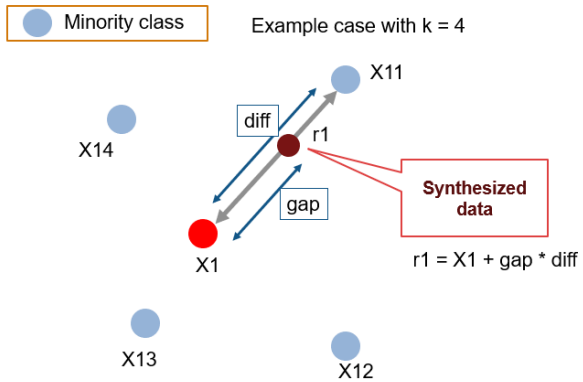
Fixing imbalanced data

0 are 1945
1 are 336

SMOTE

0 are 1945
1 are 1945

SMOTE: **Synthetic Minority Over-sampling Technique**



Hyperparameter tuning

```
[91]  para_df = ModelSelection(X_train, y_train, list_of_models, model_hyperparameters,10, 'precision')
                                                                                                  Python
```

Defined a function to perform hyperparameter tuning using grid search cv as a brain with takes in some arguments as shown and returns a data frame with best parameter for every model

	model used	highest precision score	best hyperparameters
0	LogisticRegression(max_iter=10000)	0.667935	{'C': 1}
1	SVC(probability=True)	0.819131	{'C': 50}
2	KNeighborsClassifier()	0.793518	{'n_neighbors': 3}
3	RandomForestClassifier()	0.975003	{'n_estimators': 180}
4	DecisionTreeClassifier()	0.867793	{'max_depth': 320}
5	XGBClassifier(base_score=None, booster=None, callbacks=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=None, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, feature_types=None, gamma=None, gpu_id=None, grow_policy=None, importance_type=None, interaction_constraints=None, learning_rate=None, max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None, max_delta_step=None, max_depth=None, max_leaves=None, min_child_weight=None, missing=nan, monotone_constraints=None, n_estimators=100, n_jobs=None, num_parallel_tree=None, predictor=None, random_state=None, ...)	0.927100	{'max_depth': 20, 'n_estimators': 100}

Model comparison

----------### Common_Confussion_Matrix ###

	Models	Act_Class_0	Act_Class_1
Pred_Class_0	LogisticRegression	535	53
	SVC	667	108
	KNeighborsClassifier	625	89
	RandomForestClassifier	801	144
	DecisionTreeClassifier	711	123
	XGBClassifier	772	125
Pred_Class_1	LogisticRegression	291	99
	SVC	159	44
	KNeighborsClassifier	201	63
	RandomForestClassifier	25	8
	DecisionTreeClassifier	115	29
	XGBClassifier	54	27

Type 2 error

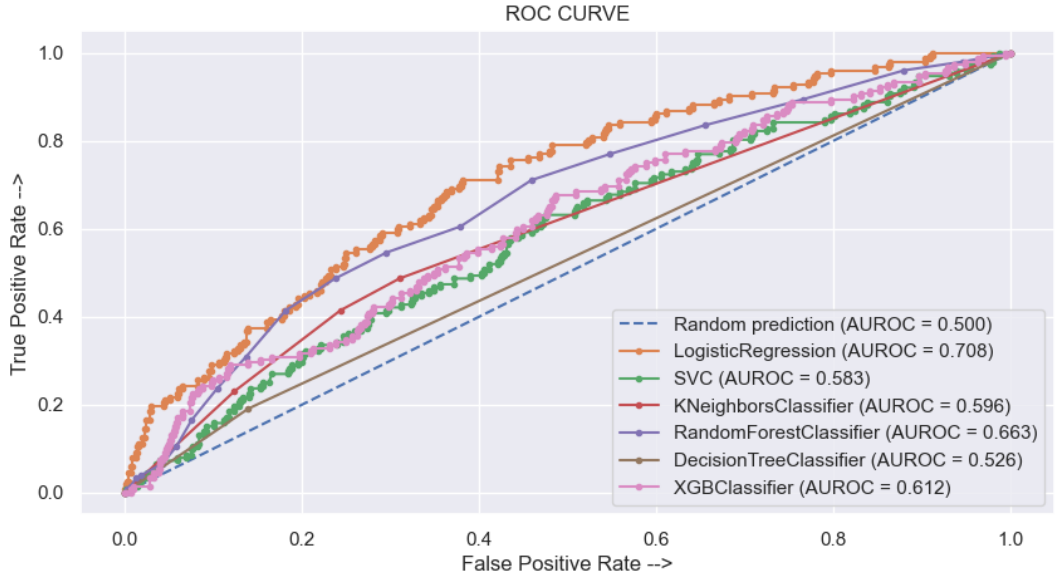
(pred are negative but actual is positive)

Type 1 error

(pred are positive but actual is negative)

Model comparison

For best recall for class (1)
we have logistic regression



----------### Common_Classifiaction_Report ###----------
Blue are highest
Red are Lowest

	Models	0	1	accuracy	macro avg	weighted avg
precision	LogisticRegression	0.909864	0.253846	-	0.581855	0.807906
	SVC	0.860645	0.216749	-	0.538697	0.760571
	KNeighborsClassifier	0.875350	0.238636	-	0.556993	0.776393
	RandomForestClassifier	0.847619	0.242424	-	0.545022	0.753560
	DecisionTreeClassifier	0.852518	0.201389	-	0.526953	0.751320
	XGBClassifier	0.860647	0.333333	-	0.596990	0.778692
recall	LogisticRegression	0.647700	0.651316	-	0.649508	0.648262
	SVC	0.807506	0.289474	-	0.548490	0.726994
	KNeighborsClassifier	0.756659	0.414474	-	0.585566	0.703476
	RandomForestClassifier	0.969734	0.052632	-	0.511183	0.827198
	DecisionTreeClassifier	0.860775	0.190789	-	0.525782	0.756646
	XGBClassifier	0.934625	0.177632	-	0.556128	0.816973
f1-score	LogisticRegression	0.756719	0.365314	0.648262	0.561016	0.695887
	SVC	0.833229	0.247887	0.726994	0.540558	0.742256
	KNeighborsClassifier	0.811688	0.302885	0.703476	0.557286	0.732610
	RandomForestClassifier	0.904574	0.086486	0.827198	0.495530	0.777427
	DecisionTreeClassifier	0.856627	0.195946	0.756646	0.526286	0.753944
	XGBClassifier	0.896111	0.231760	0.816973	0.563936	0.792858
support	LogisticRegression	826.000000	152.000000	-	978.000000	978.000000
	SVC	826.000000	152.000000	-	978.000000	978.000000
	KNeighborsClassifier	826.000000	152.000000	-	978.000000	978.000000
	RandomForestClassifier	826.000000	152.000000	-	978.000000	978.000000
	DecisionTreeClassifier	826.000000	152.000000	-	978.000000	978.000000
	XGBClassifier	826.000000	152.000000	-	978.000000	978.000000

Web application

10 Year Heart Disease Prediction

Enter your gender

Male

When's your birthday

2019/07/06

Highest academic qualification

High school diploma

Are you currently a smoker?

Yes

Number of daily cigarettes

0

Are you currently on BP medication?

Yes

Have you ever experienced a stroke?

Yes

Do you have hypertension?

Yes

Do you have diabetes?

Yes

Enter your cholesterol level

0.00

Enter your systolic blood pressure

0.00

Enter your diastolic blood pressure

0.00

Enter your BMI

0.00

Enter your resting heart rate

0.00

Enter your glucose level

0.00

Predict

You likely will NOT DEVELOP heart disease in 10 years.

Conclusion

Conclusion

1. The dataset has a high class imbalance, with significantly fewer cases of CHD than non-CHD cases.
2. The dataset contains null values, which were imputed using techniques like KNN imputation.
3. Men are more likely to suffer from CHD than women, and risk increases with age. Both men and women aged 50-52 have a higher risk, and men aged 40-42 and over 65 are also at risk.
4. Smoking does not appear to significantly affect CHD risk, but having diabetes or hypertension does increase the risk.
5. The percentage of people with CHD is almost equal between smokers and non-smokers.
6. The logistic regression model had the best performance in terms of minimizing false negatives (Type 2 errors), precision, and recall, with an AUROC of 0.708.

Conclusion

7. Based on these metrics, the logistic regression model would be the best choice for predicting CHD in this dataset.
8. Overall, our project has succeeded in developing and refining machine learning models that can make accurate predictions about cardiovascular heart disease risk prediction. These models can help doctors to get idea of heart disease for every individual patient.

Q & A