SAMSUNG

Samsung Innovation Campus

Artificial Intelligence Course

Together for Tomorrow!
Enabling People
Education for Future Generations

SAMSUNG

Causes of Heart Disease

Supervisor: Dr. Doaa Mahmoud Abdel-Aty
Facilitator: Eng. Ola Nagy

Agenda

1 Introduction

- Problem Definition
- Project Goals
- About the Dataset

2 Exploration & Insights

- General Insights
- Demographic & Personal
- Routine Related
- Substance Related
- Other Diseases & Heart Disease
- Special Circumstances
- Other Insights

3 Cleaning & Modeling

- Missing data
- Outliers
- Encoding and Scaling
- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- SVM
- AdaBoost
- XGBoost
- CatBoost
- Voting
- Comparison

Introduction

Heart Disease in 2020

80%
Percentage of preventable cases of heart disease

#1
Leading cause of death in the U.S

20%
Of heart attacks are silent

647K
Deaths every year in the U.S

Source: https://www.healthcentral.com/condition/heart-disease

Introduction

Heart Disease Prevention

Source: https://heartfoundation-prod.azurewebsites.net/bundles/your-heart/are-you-at-risk-of-heart-disease

Introduction

Project Goals

1. Draw statistical insights about the causes of heart disease.
   a. Raise Awareness about the common causes of heart disease

2. Design a detection system for heart disease.
   a. Reduce costs of heart disease prevention.
   b. Detect potential cases early

Introduction

About the Dataset

Key Indicators of Heart Disease dataset has 17 indicators of heart disease from 319,795 surveyed individuals in the U.S.

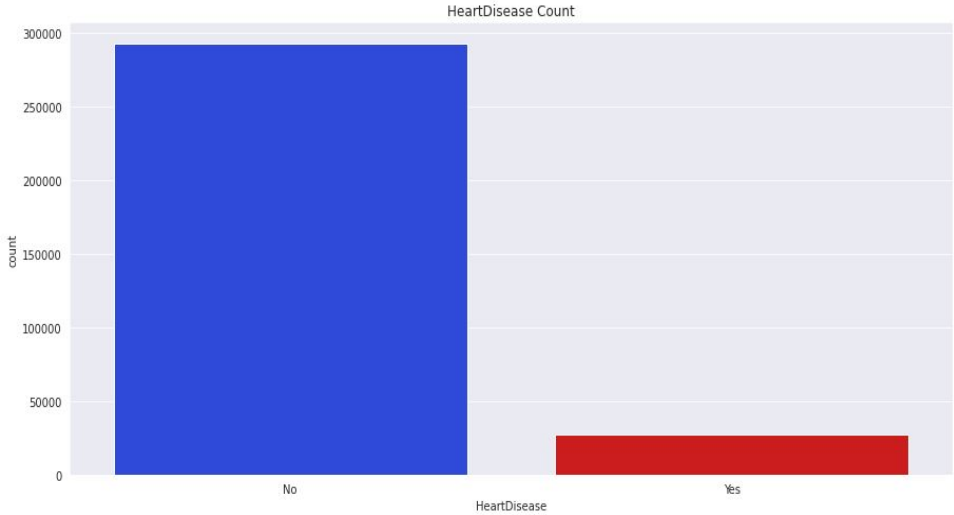
#	Feature	Description
1	HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2	BMI	Body Mass Index (BMI)
3	Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
4	AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
5	Stroke	(Ever told) (you had) a stroke?
6	PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30
7	MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?
8	DiffWalking	Do you have serious difficulty walking or climbing stairs?
9	Sex	Are you male or female?

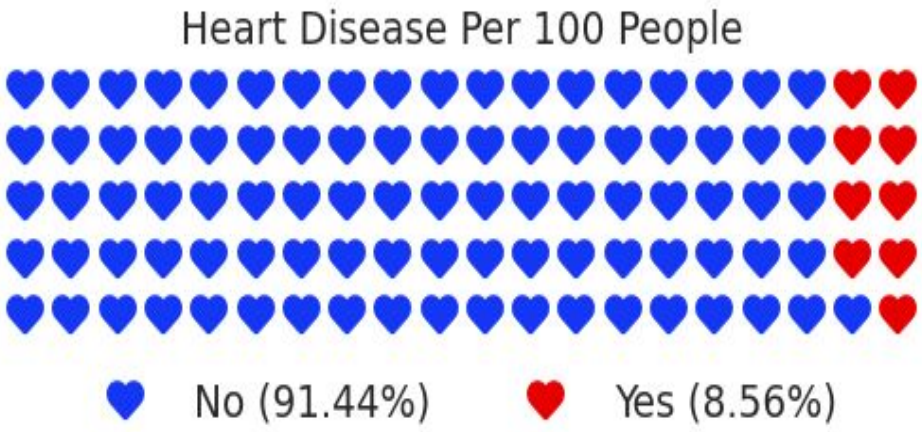
Introduction

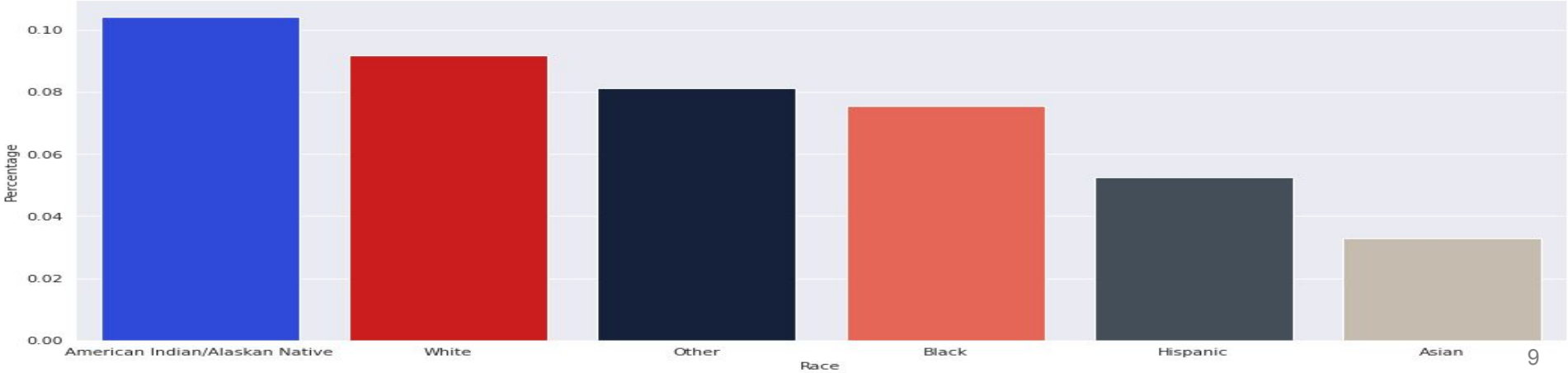
About the Dataset Continued

#	Feature	Description
10	AgeCategory	Fourteen-level age category
11	Race	Imputed race/ethnicity value
12	Diabetic	(Ever told) (you had) diabetes?
13	PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
14	GenHealth	Would you say that in general your health is...
15	SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
16	Asthma	(Ever told) (you had) asthma?
17	KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18	SkinCancer	(Ever told) (you had) skin cancer?

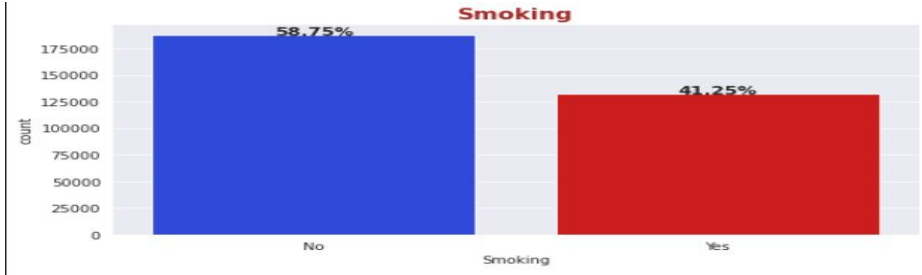
General Questions

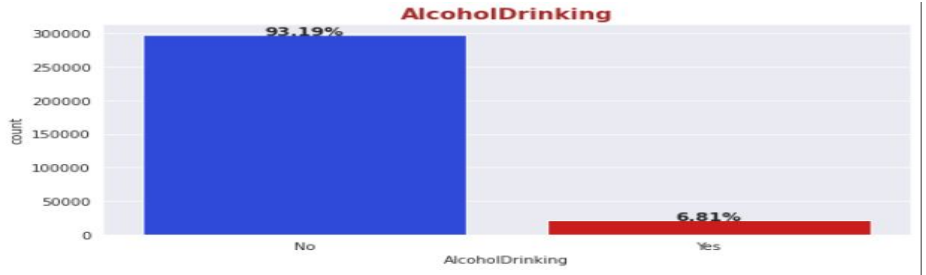


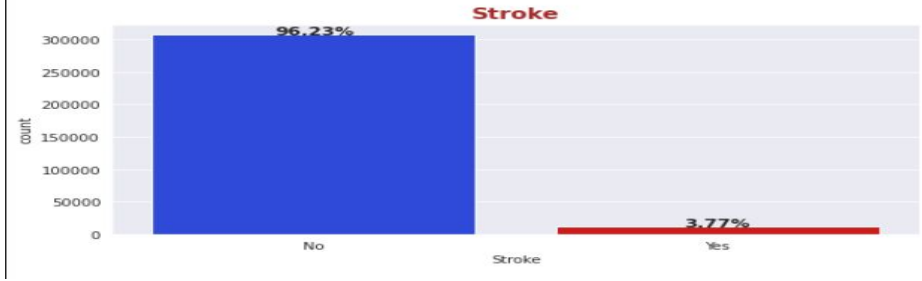


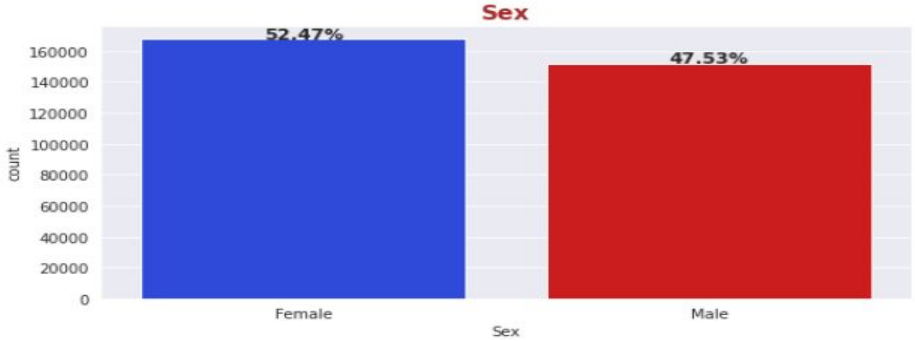


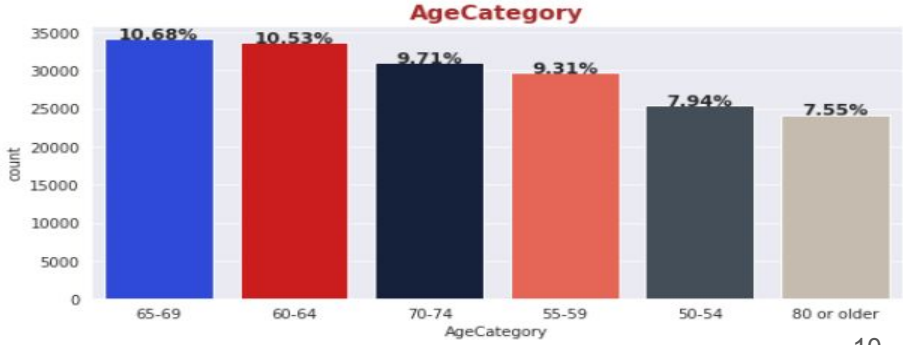
What are the distributions of our features?

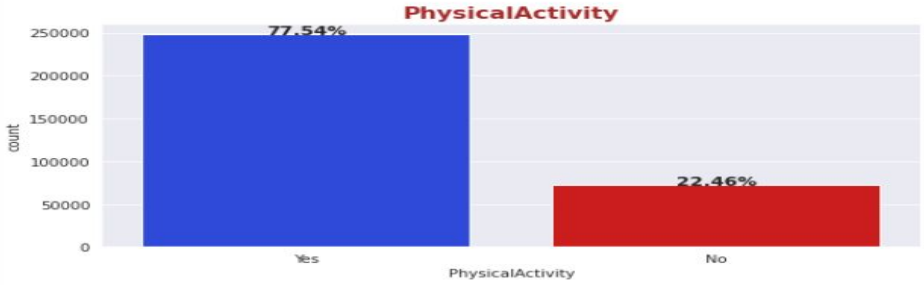


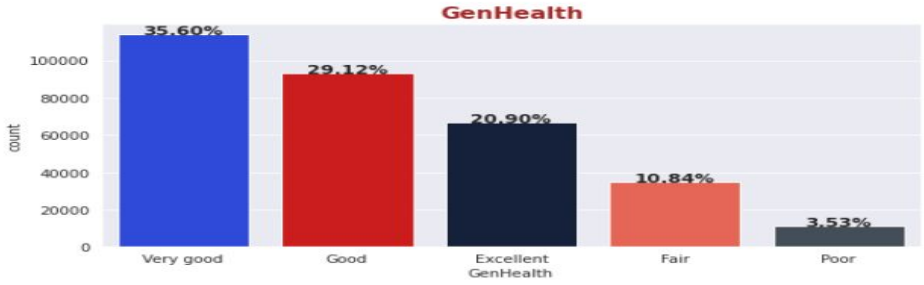


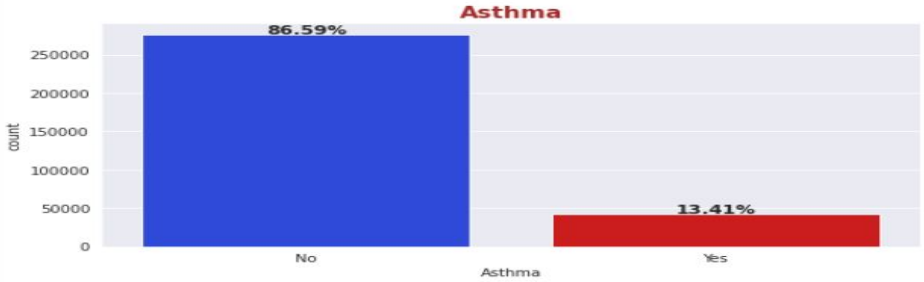


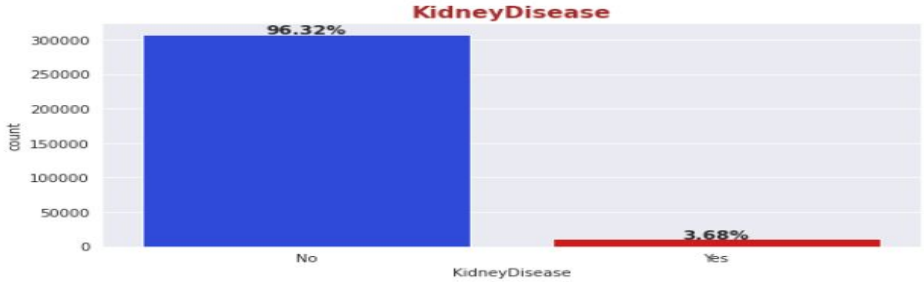


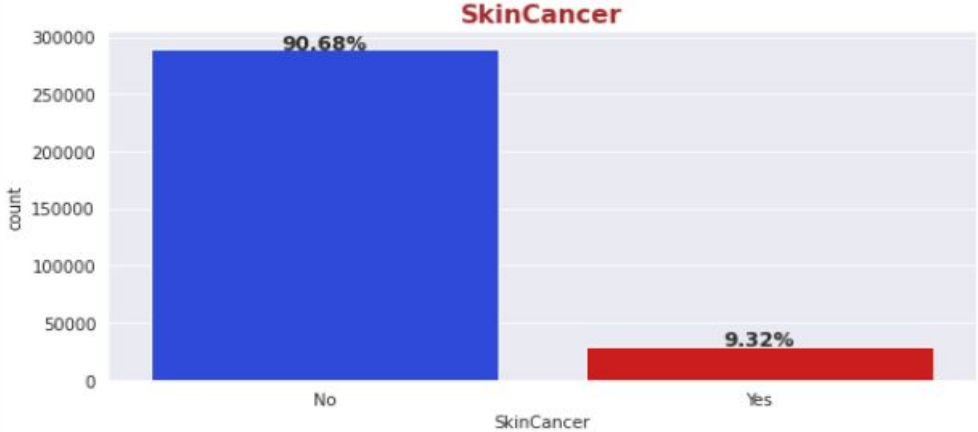



PhysicalActivity
77.54%
22.46%
count
250000
200000
150000
100000
50000
0
Yes
No
PhysicalActivity

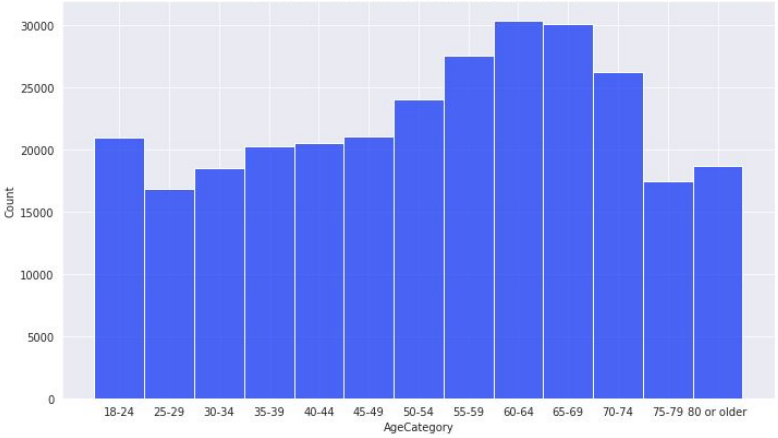

GenHealth
35.60%
29.12%
20.90%
10.84%
3.53%
count
100000
80000
60000
40000
20000
0
Very good
Good
Excellent
Fair
Poor
GenHealth

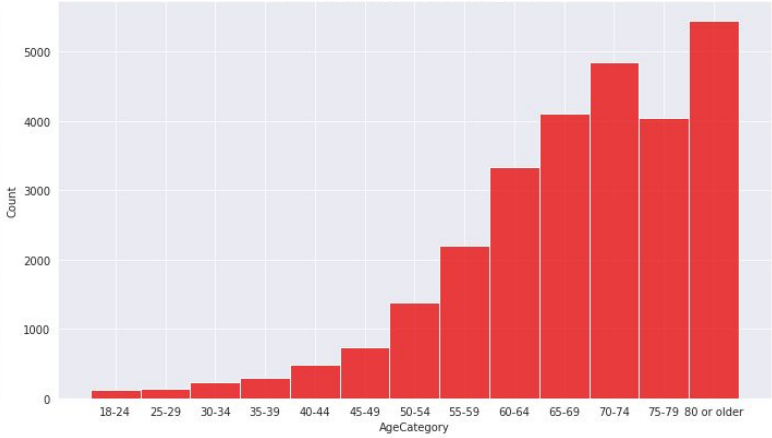

Asthma
86.59%
13.41%
count
250000
200000
150000
100000
50000
0
No
Yes
Asthma


KidneyDisease
96.32%
3.68%
count
300000
250000
200000
150000
100000
50000
0
No
Yes
KidneyDisease

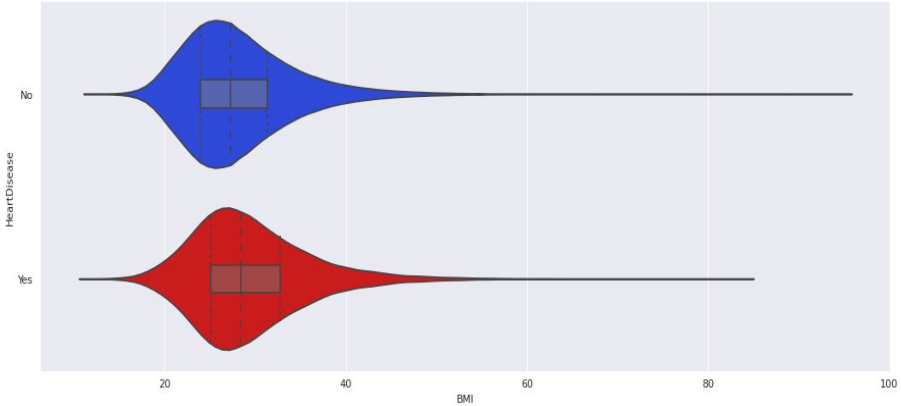

SkinCancer
90.68%
9.32%
count
300000
250000
200000
150000
100000
50000
0
No
Yes
SkinCancer

Are older individuals more susceptible to heart disease?

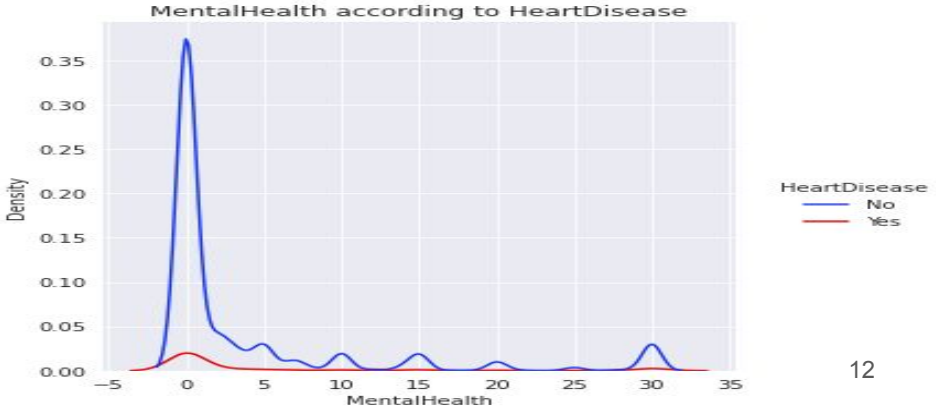




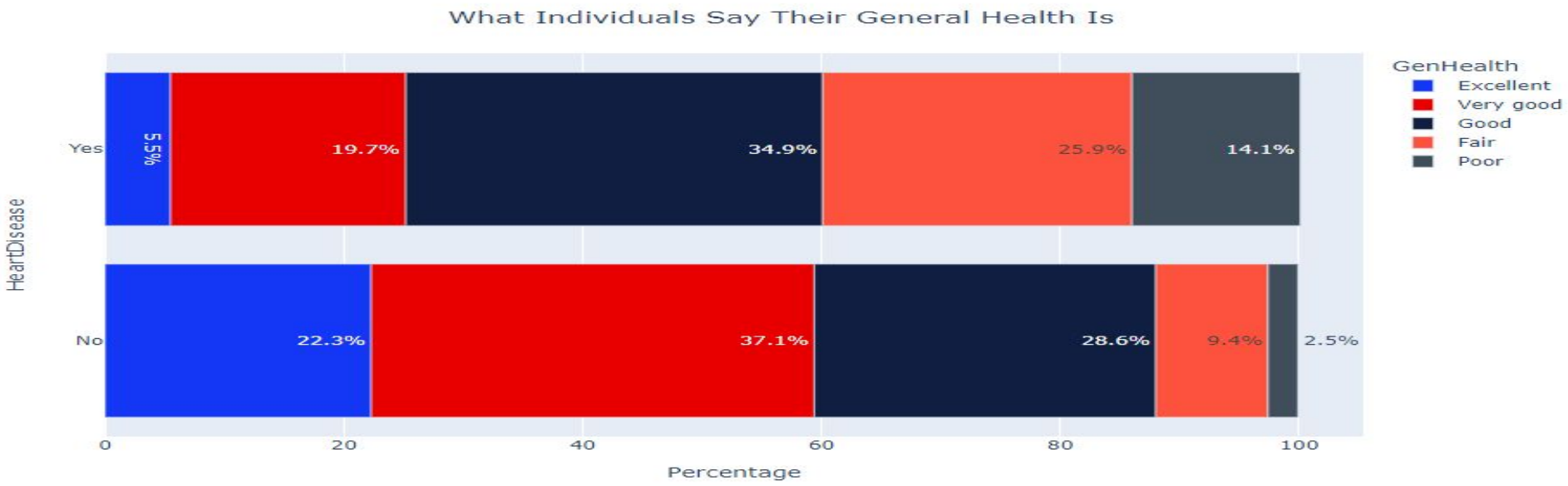
Is the BMI of heart disease patients different?



Are heart disease patients more mentally unwell?

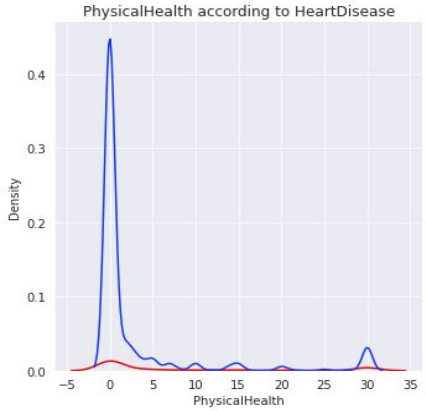


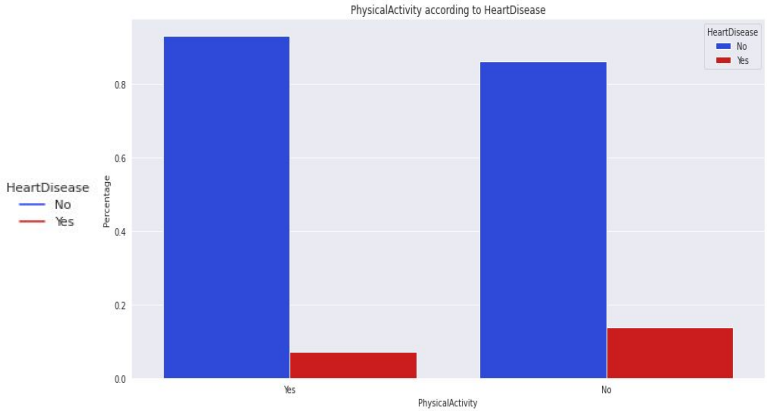
What do people who suffer from heart disease perceive their general health?

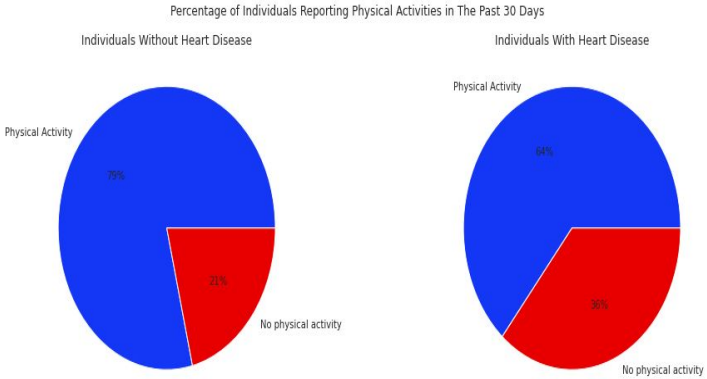


Routine Related Questions

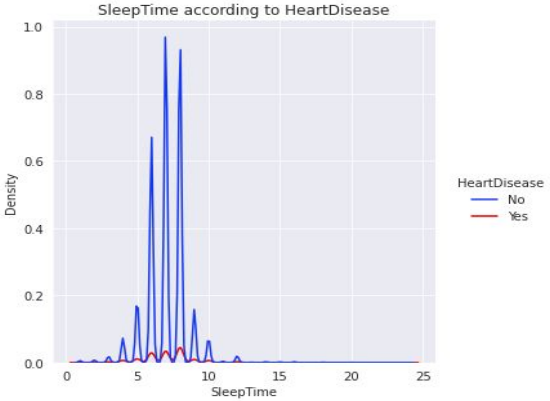
Are heart disease patients less healthy or active?

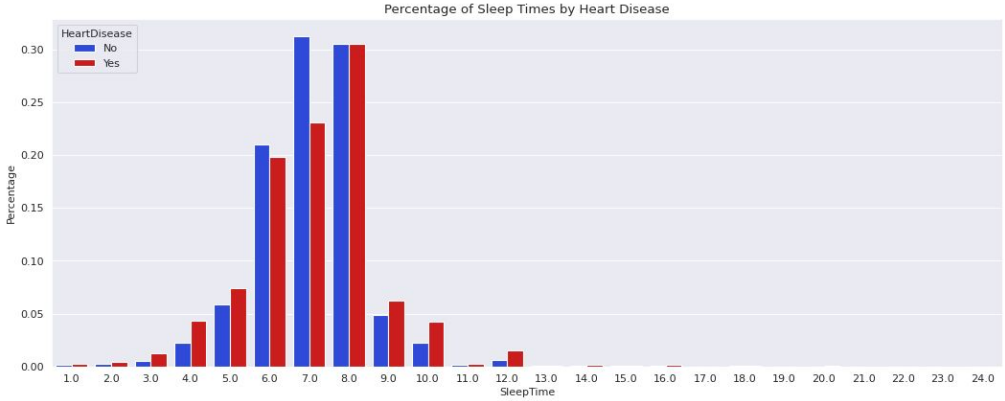






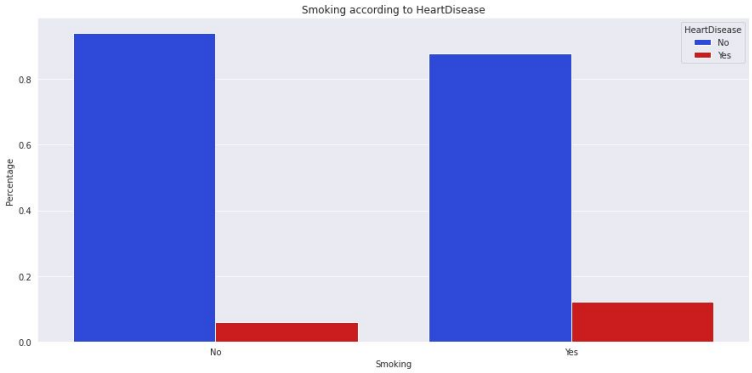
Is the distribution of sleep time among heart disease patients different?



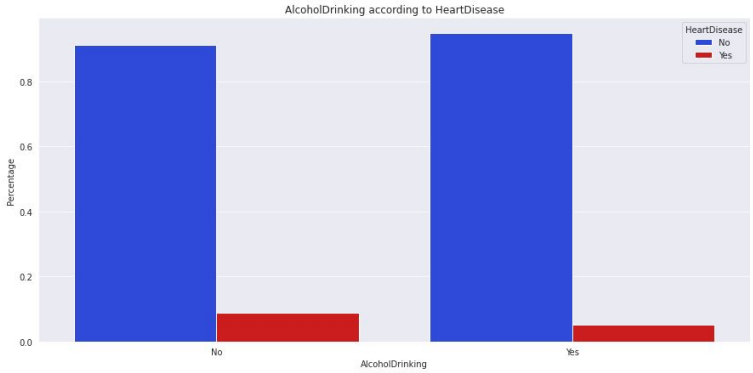


Substance Related Visualizations:

Do people with heart disease smoke more?

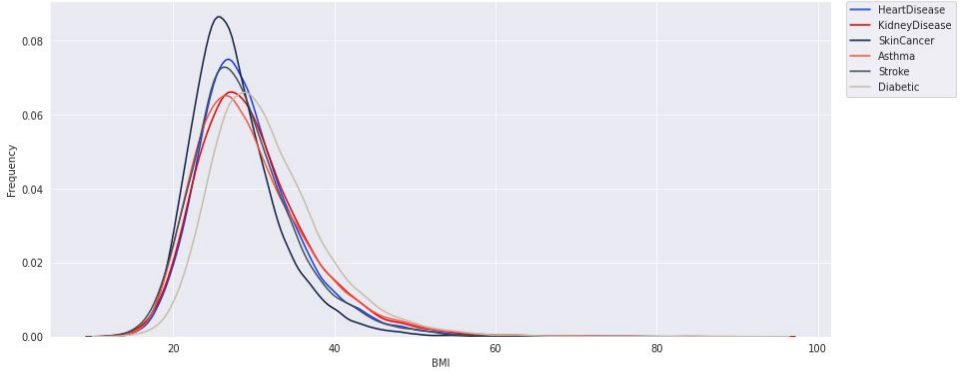


Do people with heart disease consume more alcohol?

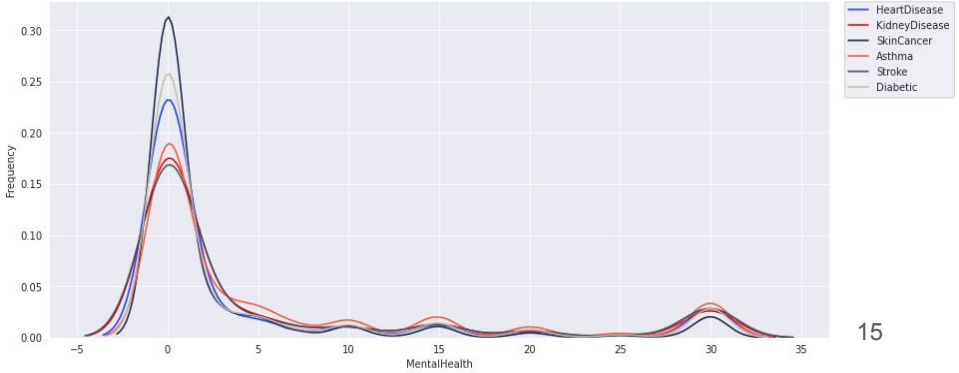


Other Questions:

Does BMI differ across diseases?

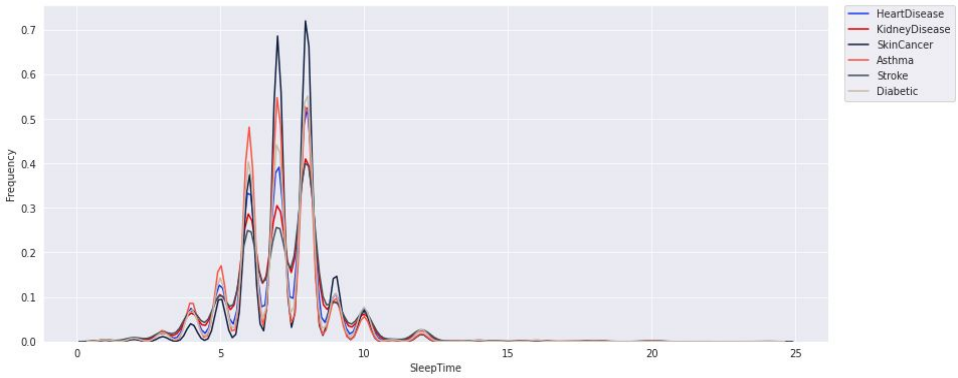


Do different diseases impact mental health differently?

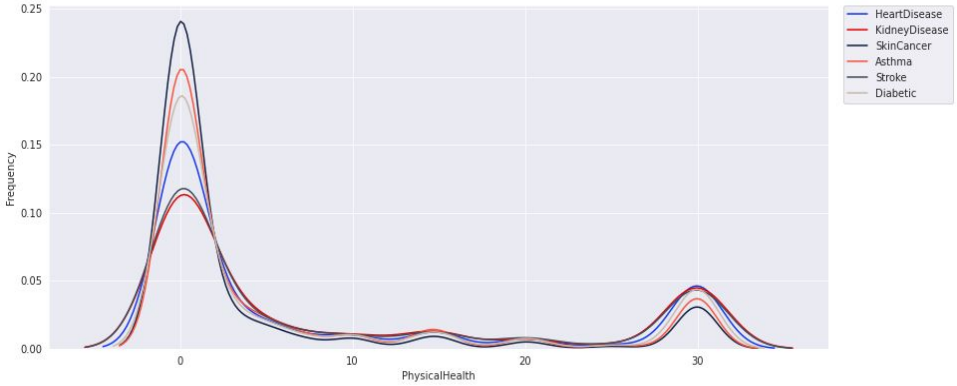


Other Questions Continued

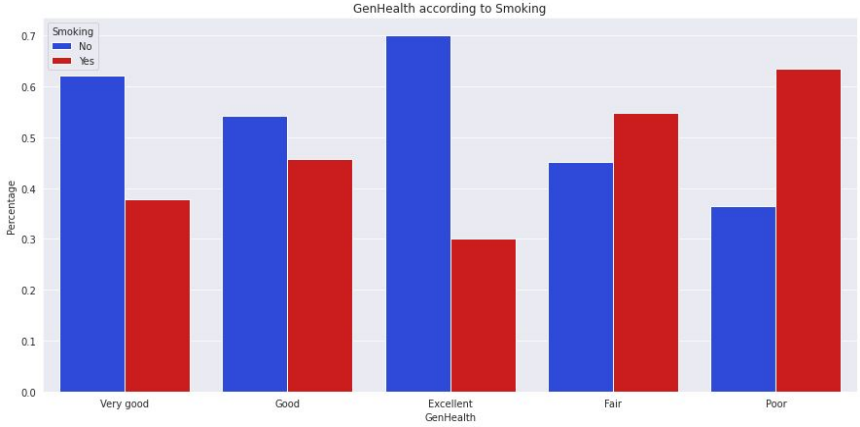
What is the effect of different diseases on sleep times?



How different is the physical health across different diseases?

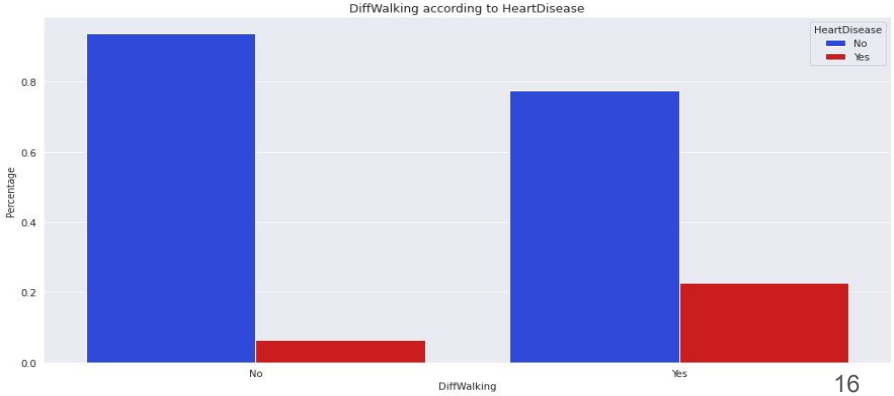


Are smokers satisfied with their health?

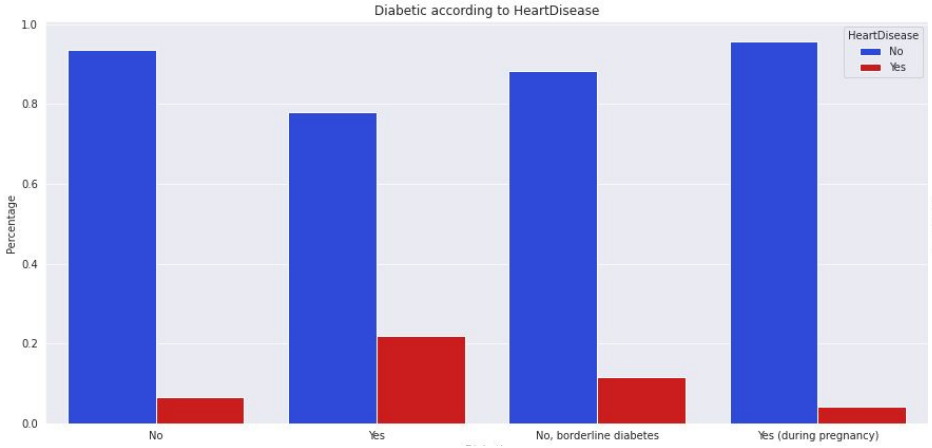


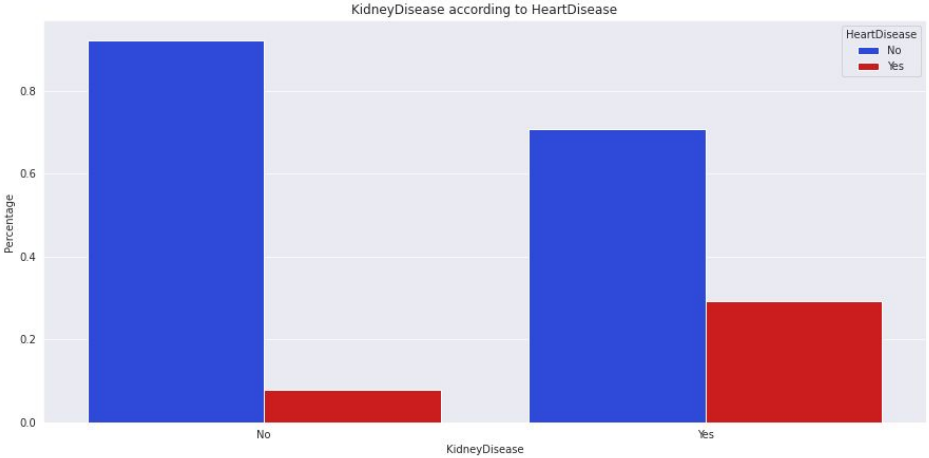
Special Circumstances and Heart Disease

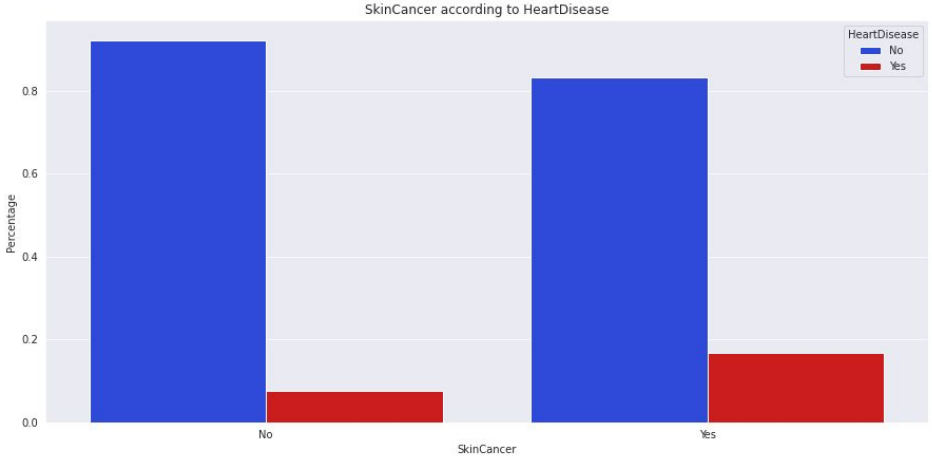
Does having difficulty walking affect heart disease?

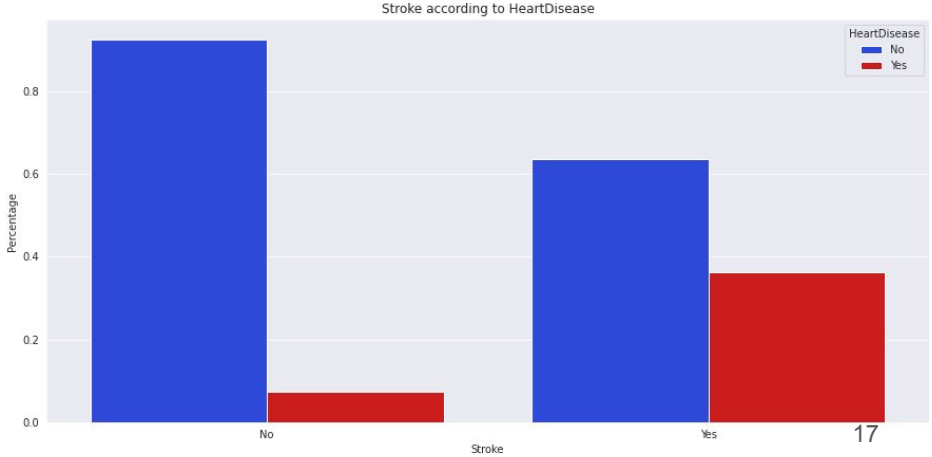


Other Diseases Effect on Heart Disease









Other Diseases Effect on Heart Disease

1. People who have had a stroke before have a heart disease percentage of ~38%. On the other hand, people who did not suffer a stroke had a significantly lower percentage of heart disease (~8%).

2. Diabetic people are at higher risk of heart disease (~25%).

3. Asthmatic people are at a slightly higher risk of heart disease.

4. Those who have suffered from kidney disease are at a significantly higher risk of heart disease. With a percentage of ~30% compared to ~9% in healthy people.

5. People who suffered from skin cancer are at a moderately higher risk of heart disease (~18% vs ~9%).

Missing Data

1. The data does not contain missing values.
2. Nor does it contain unusual values (???, etc).

Missing values per column:

HeartDisease	0
BMI	0
Smoking	0
AlcoholDrinking	0
Stroke	0
PhysicalHealth	0
MentalHealth	0
DiffWalking	0
Sex	0
AgeCategory	0
Race	0
Diabetic	0
PhysicalActivity	0
GenHealth	0
SleepTime	0
Asthma	0
KidneyDisease	0
SkinCancer	0

Unique values for categorical columns:

- HeartDisease: ['No' 'Yes']

- Smoking: ['Yes' 'No']

- AlcoholDrinking: ['No' 'Yes']

- Stroke: ['No' 'Yes']

- DiffWalking: ['No' 'Yes']

- Sex: ['Female' 'Male']

- AgeCategory: ['55-59' '80 or older' '65-69' '75-79' '40-44' '70-74' '60-64' '50-54' ''45-49' '18-24' '35-39' '30-34' '25-29']

- Race: ['White' 'Black' 'Asian' 'American Indian/Alaskan Native' 'Other' 'Hispanic']

- Diabetic: ['Yes' 'No' 'No, borderline diabetes' 'Yes (during pregnancy)']

- PhysicalActivity: ['Yes' 'No']

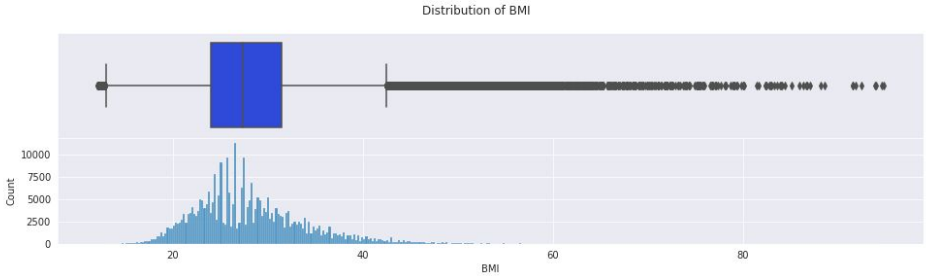
- GenHealth: ['Very good' 'Fair' 'Good' 'Poor' 'Excellent']

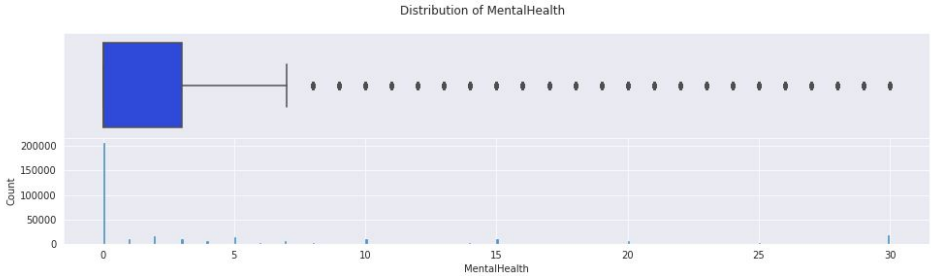
- Asthma: ['Yes' 'No']

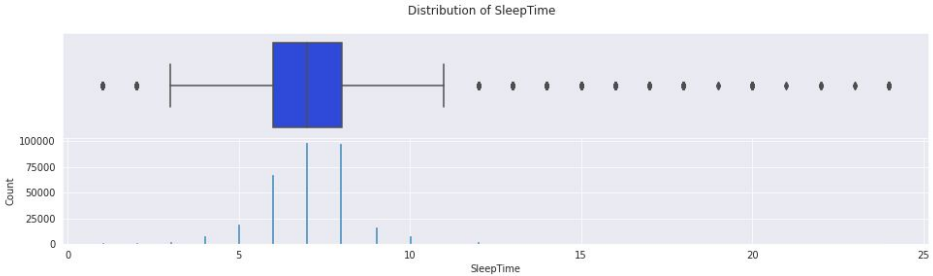
- KidneyDisease: ['No' 'Yes']

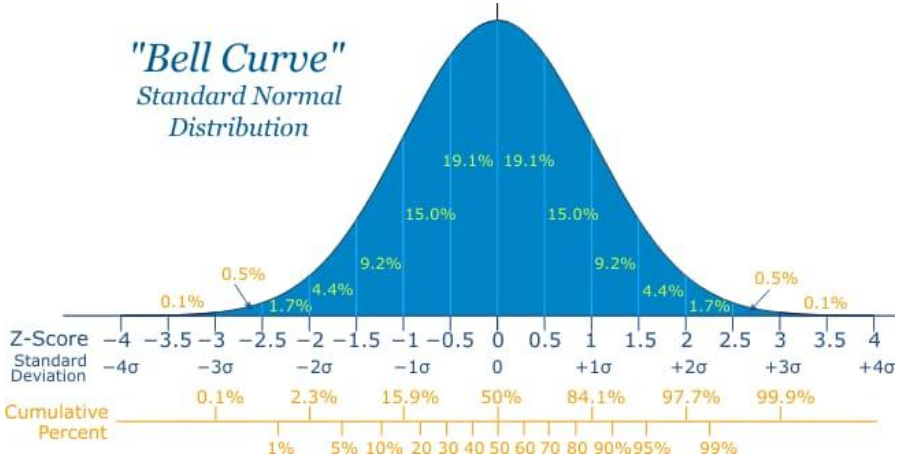
- SkinCancer: ['Yes' 'No']

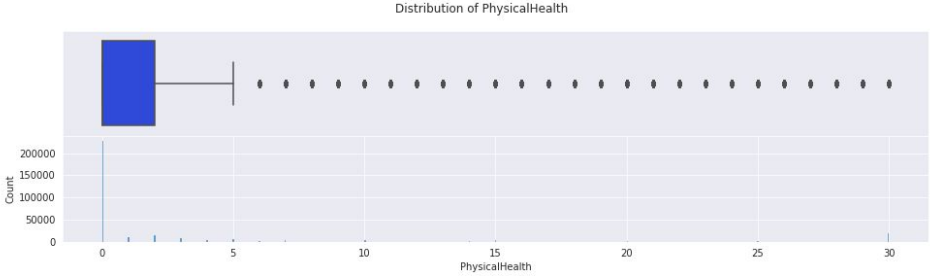
Outliers


Distribution of BMI
Count
BMI


Distribution of MentalHealth
Count
MentalHealth


Distribution of SleepTime
Count
SleepTime


"Bell Curve"
Standard Normal Distribution
19.1% 19.1%
15.0% 15.0%
9.2% 9.2%
4.4% 4.4%
1.7% 1.7%
0.5% 0.5%
0.1% 0.1%
Z-Score −4 −3.5 −3 −2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5 3 3.5 4
Standard Deviation −4σ −3σ −2σ −1σ 0 +1σ +2σ +3σ +4σ
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
Cumulative Percent
1% 5% 10% 20 30 40 50 60 70 80 90% 95% 99%

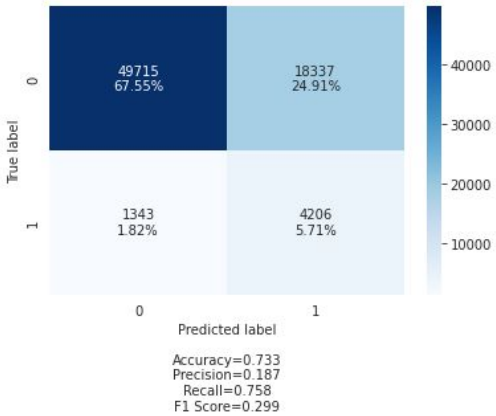

Distribution of PhysicalHealth
Count
PhysicalHealth

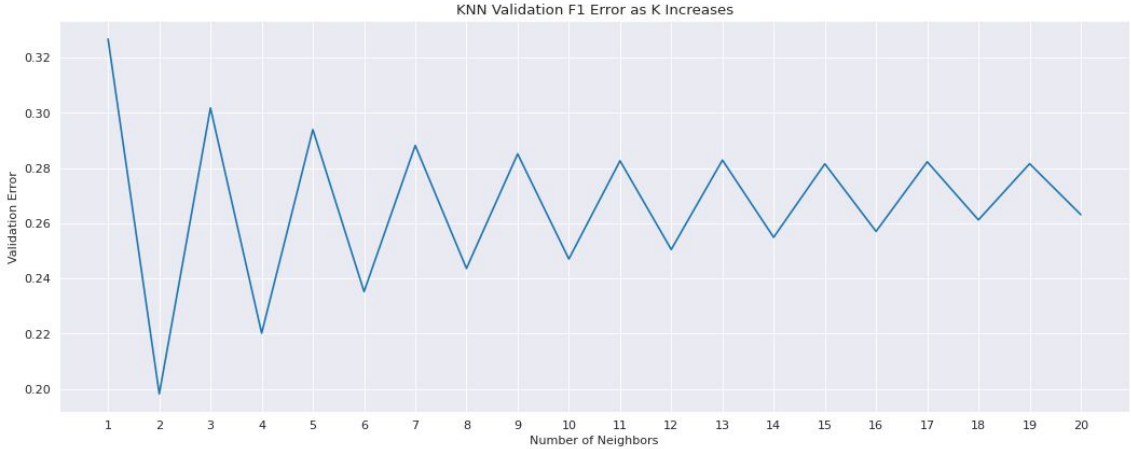
Evaluation \ Algorithm	Logistic Regression		Decision Tree		Random Forest	
Train Accuracy	72.41%		80.49%		81.11%	
Test Accuracy	72.61%		80.16%		80.83%	
Precision	98% No	19% Yes	96% No	22% Yes	96% No	22% Yes
Recall	72% No	78% Yes	82% No	63% Yes	82% No	63% Yes
F1-Score	83% No	30% Yes	88% No	32% Yes	89% No	33% Yes

KNN with Random Undersampling

Training Accuracy : 0.734
Test Accuracy : 0.732

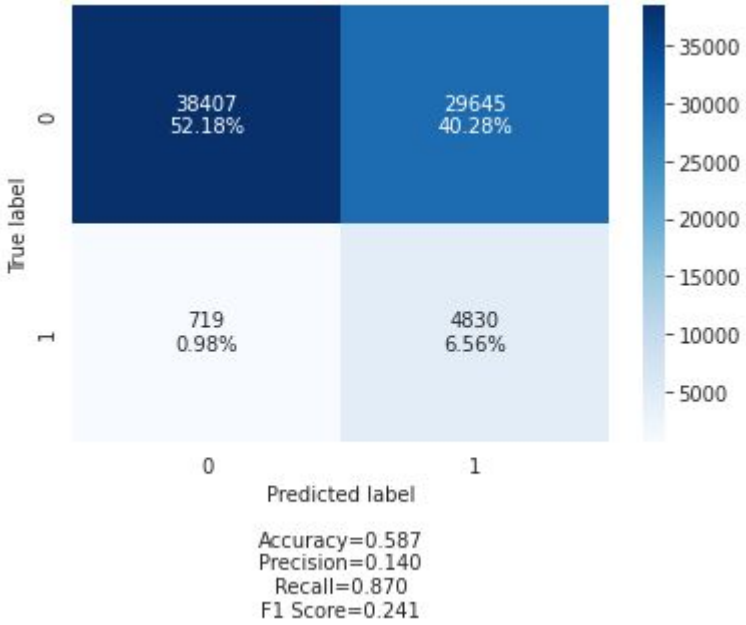
	precision	recall	f1-score	support
No	0.97	0.73	0.83	68052
Yes	0.19	0.76	0.30	5549
accuracy			0.73	73601
macro avg	0.58	0.74	0.57	73601
weighted avg	0.91	0.73	0.79	73601





SVM with Random Undersampling

	precision	recall	f1-score	support
No	0.98	0.56	0.72	68052
Yes	0.14	0.87	0.24	5549
accuracy			0.59	73601
macro avg	0.56	0.72	0.48	73601
weighted avg	0.92	0.59	0.68	73601



Training Accuracy : 0.586
Test Accuracy : 0.587

SAMSUNG

AdaBoosting

	precision	recall	f1-score	support
No	0.97	0.71	0.82	68052
Yes	0.16	0.68	0.26	5549
accuracy			0.71	73601
macro avg	0.56	0.70	0.54	73601
weighted avg	0.90	0.71	0.78	73601

Training Accuracy : 0.714
Test Accuracy :0.670

XGBoosting

	precision	recall	f1-score	support
No	0.98	0.73	0.84	68052
Yes	0.19	0.80	0.31	5549
accuracy			0.73	73601
macro avg	0.59	0.76	0.57	73601
weighted avg	0.92	0.73	0.80	73601

Training Accuracy : 0.918
Test Accuracy : 0.897

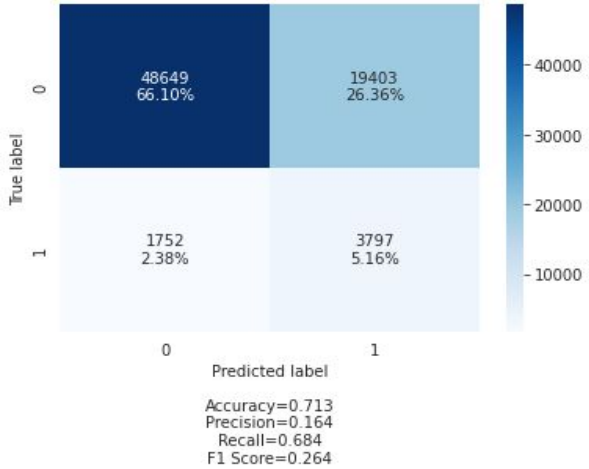
CatBoosting

	precision	recall	f1-score	support
No	0.94	0.96	0.95	68052
Yes	0.28	0.21	0.24	5549
accuracy			0.90	73601
macro avg	0.61	0.58	0.59	73601
weighted avg	0.89	0.90	0.89	73601

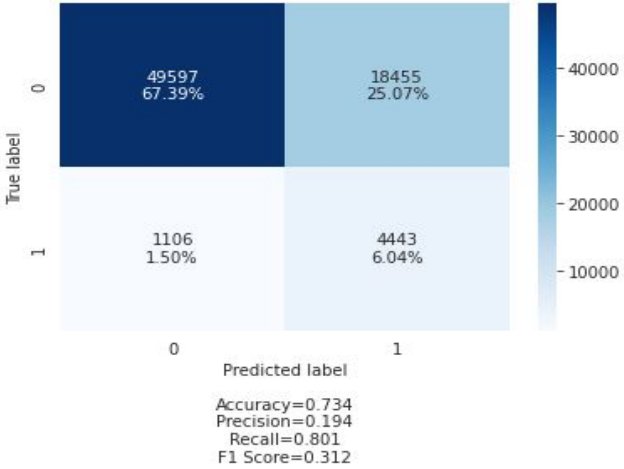
Training Accuracy : 0.908
Test Accuracy : 0.898

SAMSUNG

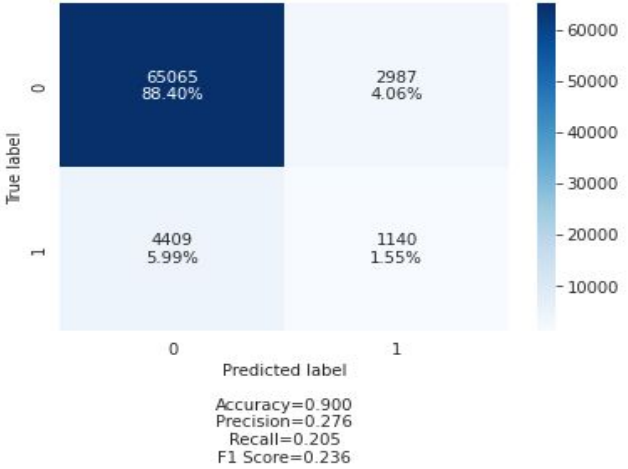
AdaBoosting



XGBoosting



CatBoosting

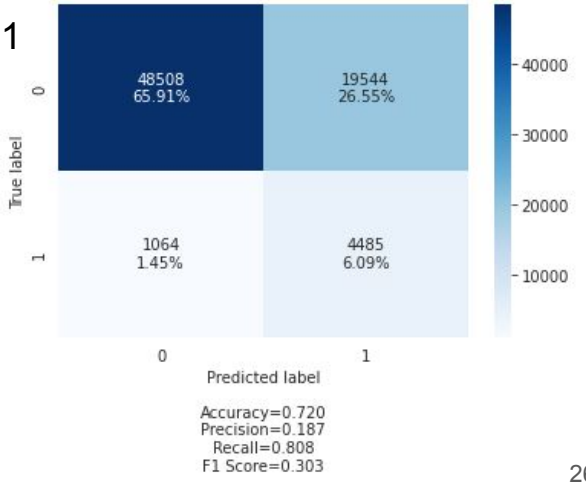


Voting Model with Random Undersampling

1. Logistic Regression
2. Decision Tree
3. KNN
4. SVM
5. XGBoost

	precision	recall	f1-score	support
No	0.98	0.71	0.82	68052
Yes	0.19	0.81	0.30	5549
accuracy			0.72	73601
macro avg	0.58	0.76	0.56	73601
weighted avg	0.92	0.72	0.79	73601

Training Accuracy : 0.721
Test Accuracy : 0.720



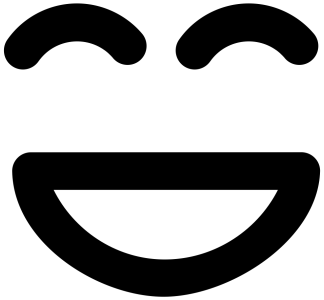
Comparison

Model	Sampling Method	Heart Disease Precision	No Heart Disease Precision	Heart Disease Recall	No Heart Disease Recall	Heart Disease F1	No Heart Disease F1	Accuracy
Logistic Regression	SMOTE	0.19	**0.98**	0.78	0.72	0.30	0.83	0.73
Decision Tree	SMOTE	0.22	0.96	0.63	0.82	0.32	0.88	0.80
Random Forest	SMOTE	0.22	0.96	0.63	0.82	**0.33**	0.89	0.81
KNN	Random Undersampling	0.19	0.97	0.76	0.73	0.30	0.83	0.73
SVM	Random Undersampling	0.14	**0.98**	**0.87**	0.56	0.24	0.72	0.59
AdaBoost	Random Undersampling	0.17	0.97	0.73	0.70	0.27	0.81	0.70
XGBoost	Random Undersampling	0.19	**0.98**	0.80	0.73	0.31	0.84	0.73
CatBoost	SMOTE	**0.28**	0.94	0.21	**0.96**	0.24	**0.95**	**0.90**
Voting Classifier	Random Undersampling	0.19	**0.98**	0.81	0.71	0.30	0.82	0.72

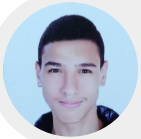
Conclusion

1. We investigated the Personal Key Indicators of Heart Disease which had 17 indicators of heart disease of 319,795 surveyed individuals in the U.S.
2. Age is a major factor in heart disease.
3. Heart disease is more prominent in smokers (~12%), kidney disease victims (~30%), stroke victims (~38%), skin cancer patients (~18%), people who have difficulty walking (~18%), and diabetics (~25%).
4. SVMs with random undersampling yield the best recall for the heart disease class (87%).
5. CatBoost with SMOTE yields the best recall for the no heart disease class (96%)
6. Decision tree/Random Forest had the best compromise.
7. An application of our model is to be used by medical experts in selecting the patients suspected of heart disease in order to conduct further testing on them.

Thank You!



Contact Details



Mohamed Salem

Habiba Shera

Ahmed Ashraf Mokhtar