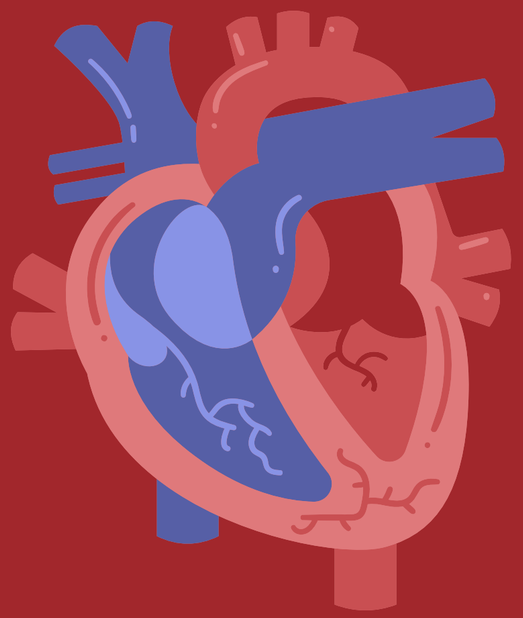
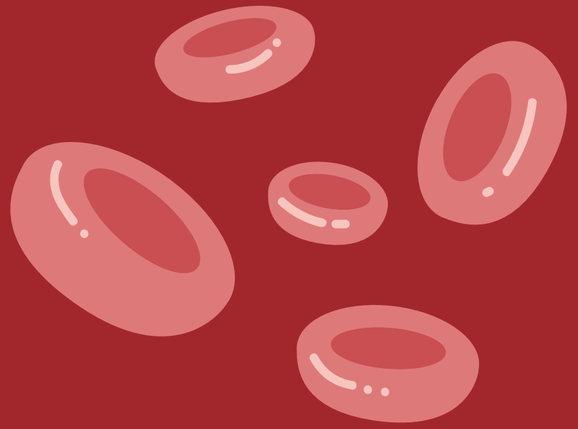


HEART DISEASE PREDICTION



ITMD 522 - Data Mining &
Machine Learning
Group No. 935

A20555648 : Patel Devarshi
A20551574-Mohammed Padghawala
A20548173 - Viraj Parmar
A20551543 - Raj Dedhia

BACKGROUND – DATASET

Source - https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/code?datasetId=1936563&searchQuery=imba

Total Rows : 319795
Duplicates : 18078
Null Values : 0
Total Features : 18
Target Feature : HeartDisease
Other Features : BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex , AgeCategory, Race Diabetic, PhysicalActivity, GenHealth, SleepTime , Asthma, KidneyDisease

BACKGROUND – DATASET

BMI: Body Mass Index

Smoking: ⩾100 cigarettes lifetime

Alcohol: Men >14, Women >7 drinks/week

Stroke: History of stroke

Physical Health: Illness/injury days in past 30

Mental Health: Days of poor mental health in past 30

Difficulty Walking: Serious difficulty with stairs or walking

Sex: Male or Female

BACKGROUND – DATASET

Age: 14 categories

Race: Ethnicity value

Diabetic: Diabetes history

Physical Activity: Exercise outside job in past 30 days

General Health: Self-rated health status

Sleep Time: Average sleep in 24 hours

Asthma: History of asthma

Kidney Disease: Excluding stones, bladder issues, etc.

Skin Cancer: History of skin cancer

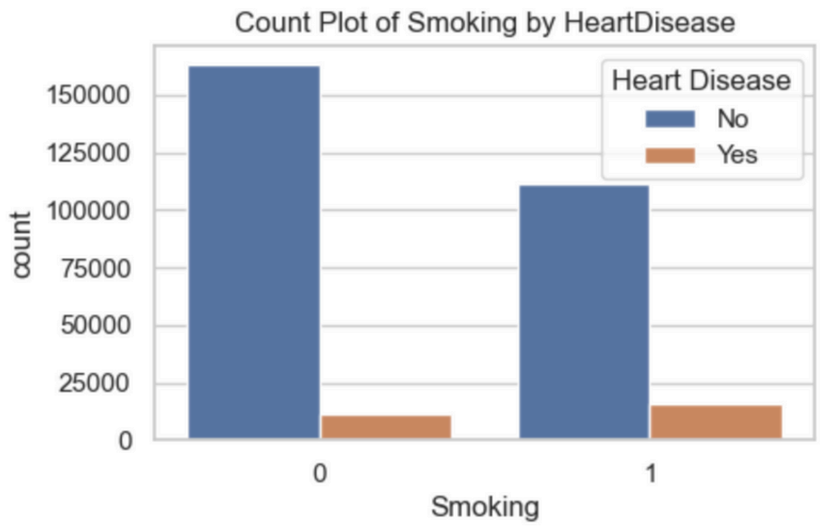
DATASET – BEFORE PRE PROCESSING:

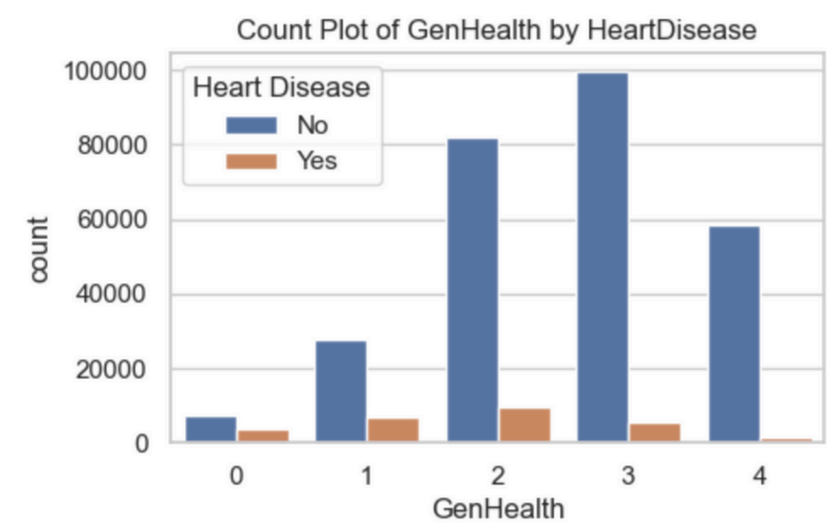
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes	Very good
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes	Very good
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes	Fair
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No	Good
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes	Very good
5	Yes	28.87	Yes	No	No	6.0	0.0	Yes	Female	75-79	Black	No	No	Fair
6	No	21.63	No	No	No	15.0	0.0	No	Female	70-74	White	No	Yes	Fair
7	No	31.64	Yes	No	No	5.0	0.0	Yes	Female	80 or older	White	Yes	No	Good
8	No	26.45	No	No	No	0.0	0.0	No	Female	80 or older	White	No, borderline diabetes	No	Fair
9	No	40.69	No	No	No	0.0	0.0	Yes	Male	65-69	White	No	Yes	Good

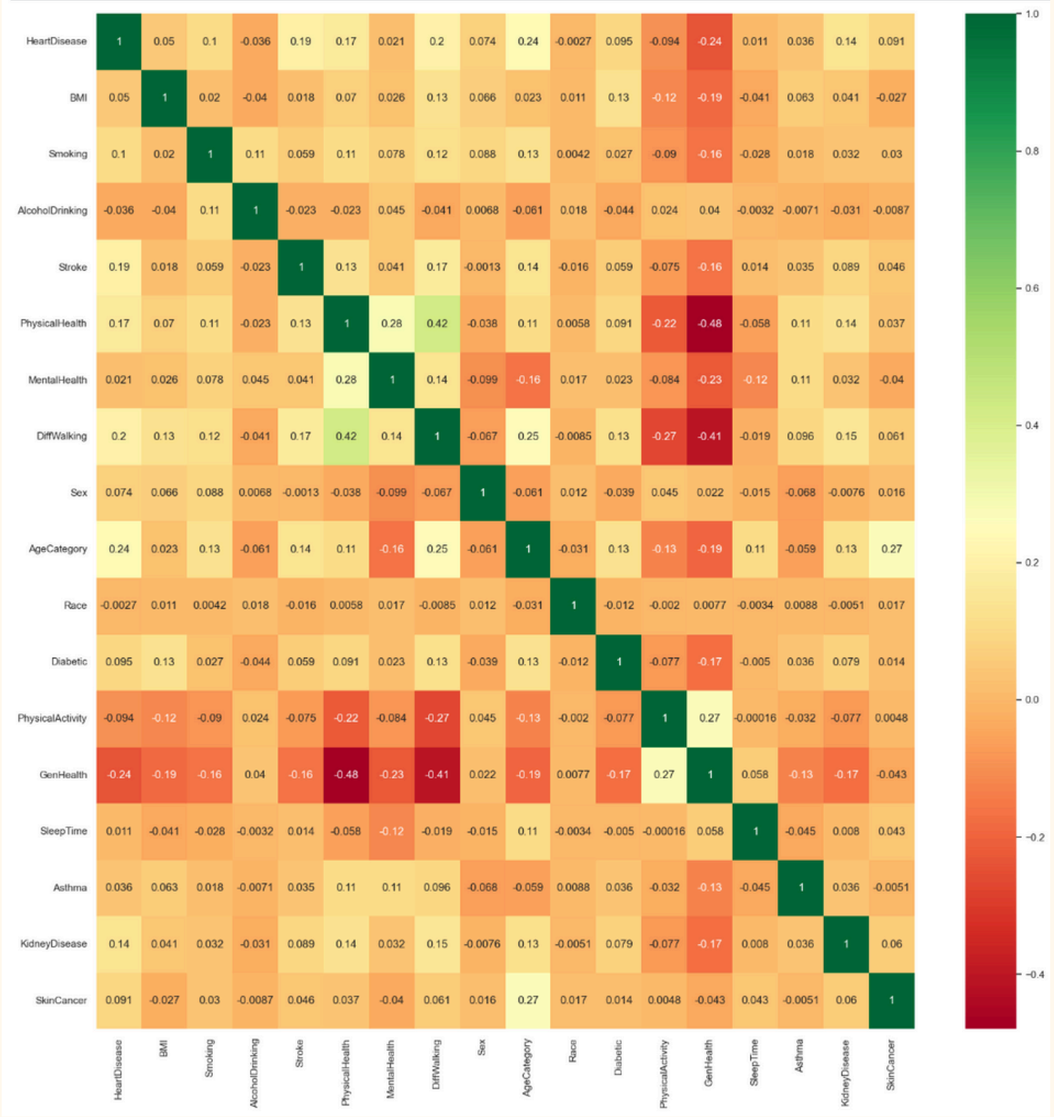
DATASET PRE PROCESSING:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth
0	0	0	1	0	0	3.0	30.0	0	0	7	3	1	1	3
1	0	1	0	0	1	0.0	0.0	0	0	12	3	0	1	3
2	0	2	1	0	0	20.0	30.0	0	1	9	3	1	1	1
3	0	1	0	0	0	0.0	0.0	0	0	11	3	0	0	2
4	0	1	0	0	0	28.0	0.0	1	0	4	3	0	1	3
5	1	2	1	0	0	6.0	0.0	1	0	11	2	0	0	1
6	0	1	0	0	0	15.0	0.0	0	0	10	3	0	1	1
7	0	3	1	0	0	5.0	0.0	1	0	12	3	1	0	2
8	0	2	0	0	0	0.0	0.0	0	0	12	3	3	0	1
9	0	3	0	0	0	0.0	0.0	1	1	9	3	0	1	2

IDENTIFICATION OF SIGNIFICANT RISK FACTORS:

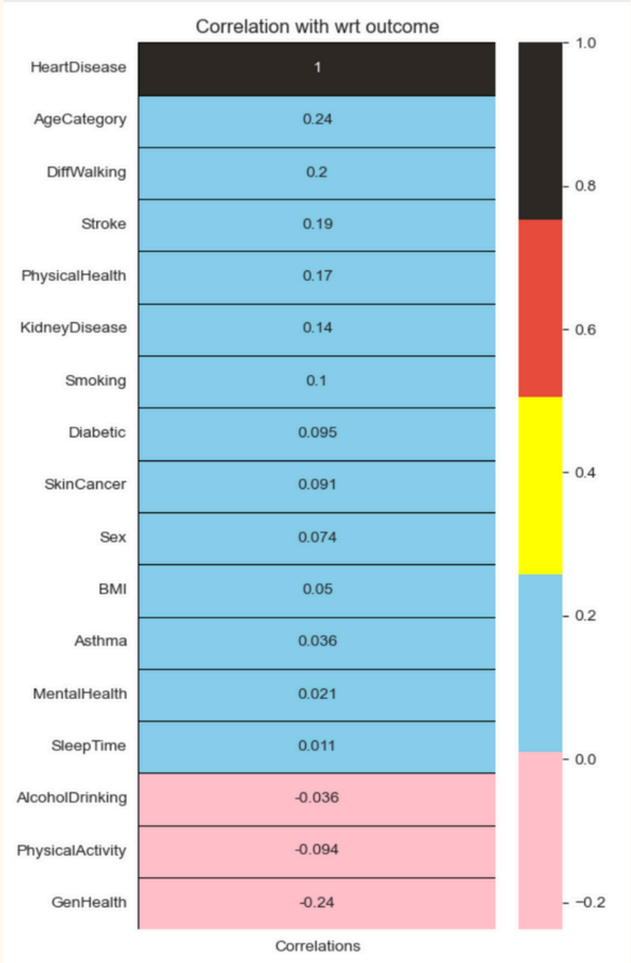




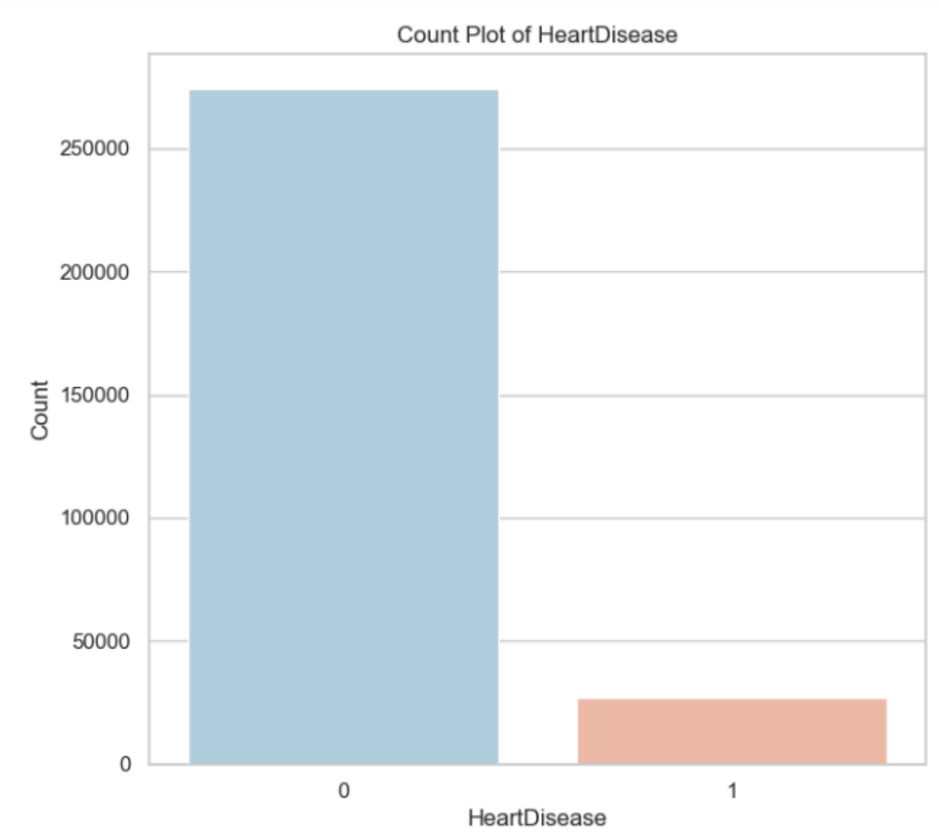


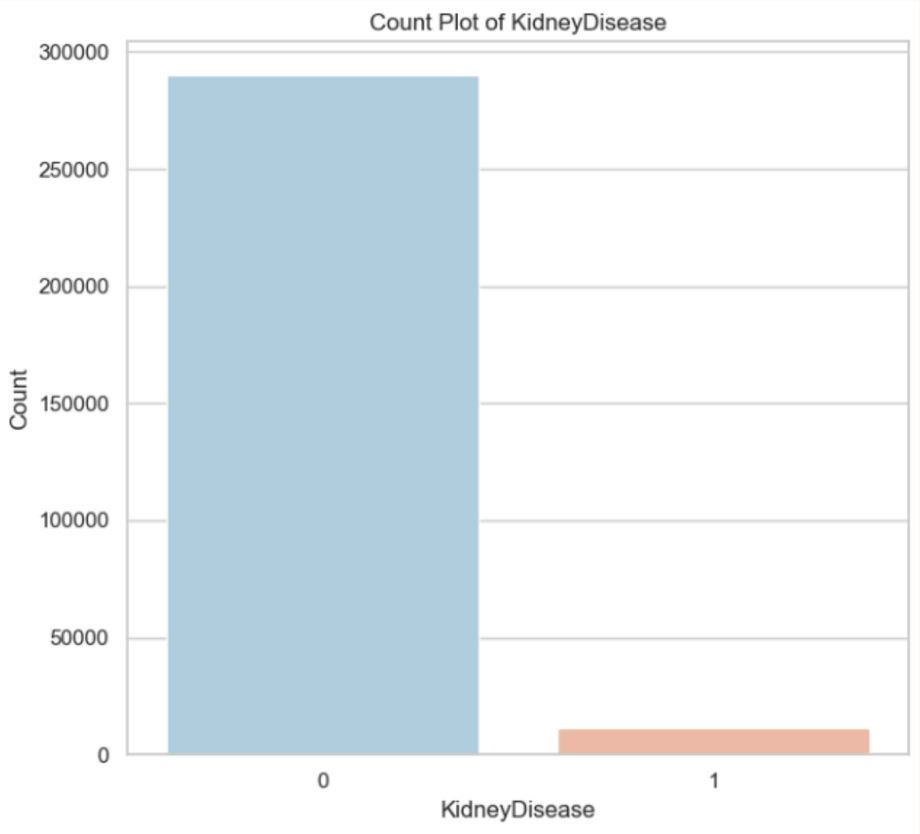
IDENTIFICATION OF SIGNIFICANT RISK FACTORS:

Age: Higher prevalence of heart disease in older age categories.
BMI: Increased heart disease cases in higher BMI classifications.
Kidney Disease: Lower heart disease presence among those without kidney disease.
Skin Cancer: Minimal variation in heart disease presence related to skin cancer.
General Health: Better self-rated health correlates with lower heart disease incidence.
Asthma: Notable count of heart disease in individuals with asthma.
Diabetes: Higher heart disease presence among diabetics.
Physical Activity: Less heart disease in physically active individuals.
Mobility Issues: Greater heart disease presence in those with difficulty walking.
Sex: Observable sex difference in heart disease prevalence.
Alcohol Consumption: Lower heart disease in those with less/no alcohol consumption.
Stroke History: Strong association between stroke history and heart disease.
Smoking: Higher heart disease presence among smokers.



IMBALANCE ISSUE & MODEL BUILDING:





- There is a severe class imbalance (heart disease vs healthy)
- Imbalanced attributes : HeartDisease, Alcohol drinking, Sroke, DiffWalking, Race, Diabetic, PhysicalActivity, Asthma, KidneyDisease, SkinCancer.

IMBALANCE ISSUE & MODEL BUILDING:

- Imbalanced Models: Logistic Regression Vs Random Forest

	LogReg_Imbalanced	RF_Imbalanced
Accuracy	0.910579	0.901153
Precision	0.534024	0.368577
Recall	0.105602	0.126956
F1-Score	0.176334	0.188860
AUC	0.829571	0.782174

- Over 90% accuracy misleading due to majority class bias.
- Imbalanced models favor precision over recall, risk missing many true positives.

IMBALANCE ISSUE & MODEL BUILDING:

- Balanced Models: Logistic Regression Vs Random Forest
- Weight Balancing

	LogReg_Balanced	RF_Balanced
Accuracy	0.740103	0.740103
Precision	0.225193	0.225193
Recall	0.765102	0.765102
F1-Score	0.347968	0.347968
AUC	0.829929	0.829929

- Lower accuracy (74%), yet more reflective of true model capability.
- Greatly enhance recall, critical for detecting minority class.
- Slightly higher F1-scores in balanced models indicate improved handling of the minority class.

IMBALANCE ISSUE & MODEL BUILDING:

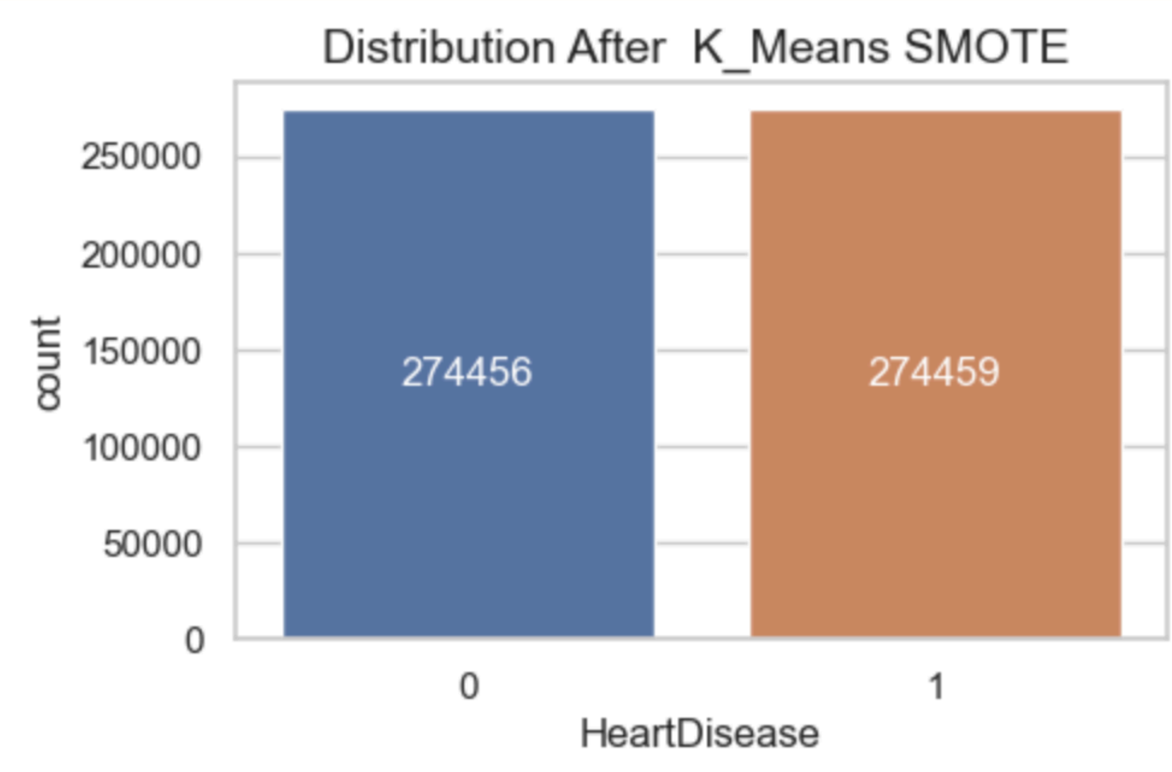
- Balanced Models: Logistic Regression Vs Random Forest
- SMOTE

	LogReg_Smote	RF_Smote
Accuracy	0.740130	0.879133
Precision	0.224975	0.290979
Recall	0.763639	0.232119
F1-Score	0.347557	0.258238
AUC	0.828730	0.778652

- Comparing with weight balanced model, Logistic Regression model is almost identical.
- In SMOTE Regression Model we got increased accuracy but lower Recall, F1-Score and AUC, indicating a reduced ability to distinguish between classes.

IMBALANCE ISSUE & MODEL BUILDING:

- Balanced Models: Logistic Regression
- K-Means SMOTE



	LogReg_KSmote
Accuracy	0.903687
Precision	0.905353
Recall	0.901833
F1-Score	0.903590
AUC	0.956998

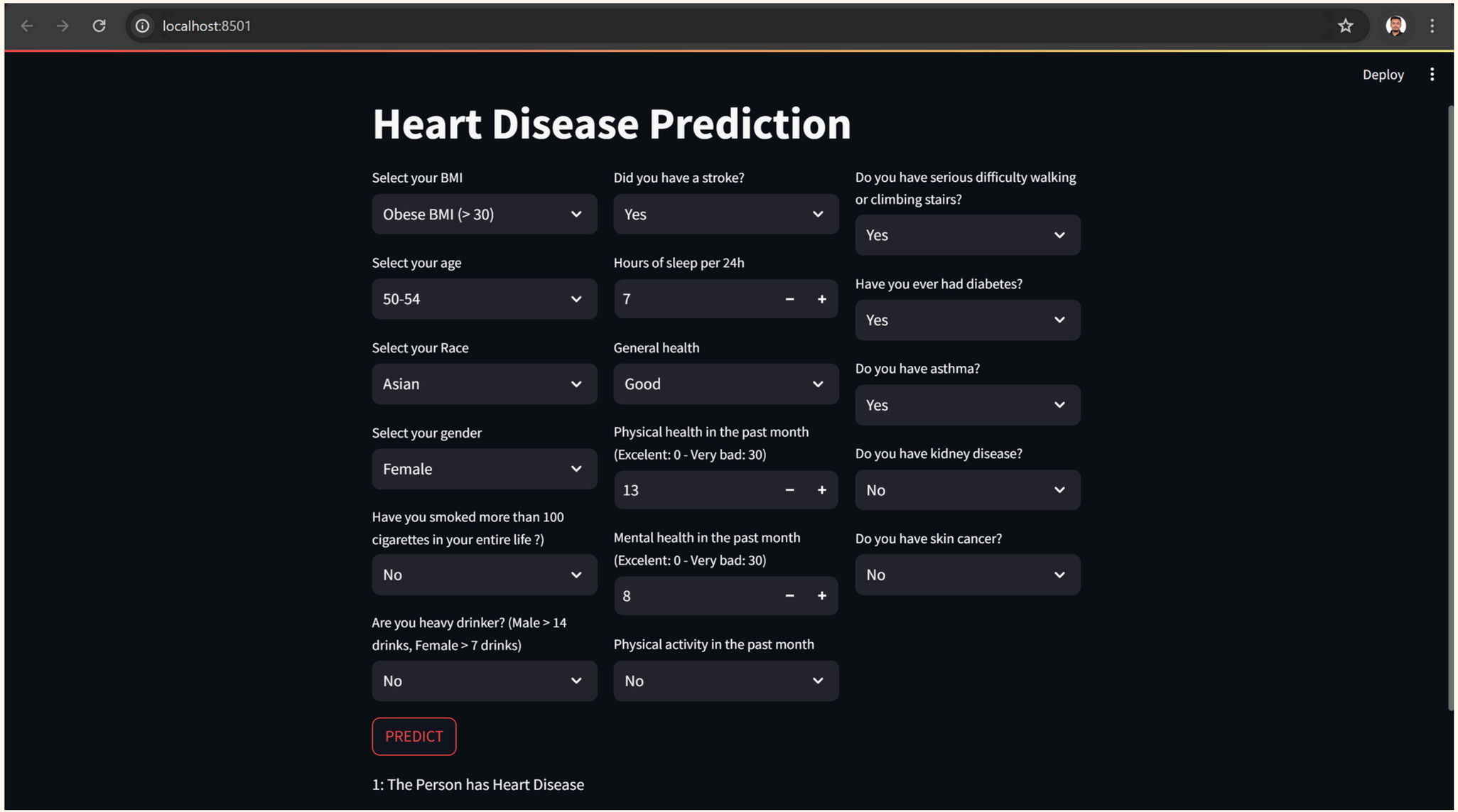
IMBALANCE ISSUE & MODEL BUILDING:

- Comparison :

	LogReg_Imbalanced	RF_Imbalanced	LogReg_Balanced	RF_Balanced	LogReg_Smote	RF_Smote	LogReg_KSmote
Accuracy	0.910579	0.901153	0.740103	0.740103	0.740130	0.879133	0.903687
Precision	0.534024	0.368577	0.225193	0.225193	0.224975	0.290979	0.905353
Recall	0.105602	0.126956	0.765102	0.765102	0.763639	0.232119	0.901833
F1-Score	0.176334	0.188860	0.347968	0.347968	0.347557	0.258238	0.903590
AUC	0.829571	0.782174	0.829929	0.829929	0.828730	0.778652	0.956998

- The KMeansSMOTE technique applied to Logistic Regression greatly enhances recall and AUC, suggesting excellent identification of the minority class and strong discrimination between classes, without a significant trade-off in precision.

STREAMLIT WEB APP:



THANK YOU !