Smoking and drinking

prediction model

Temirlan Yeslamov
Azamat Galidenov
Zholdas Aldanbergen

Topic explanation

This project aims to **predict alcohol consumption** using machine learning by analyzing a dataset with various **physical parameters**. Before model training, our team conducted comprehensive research on these parameters **to identify patterns and dependencies** that influence drinking behavior. The insights gained from this research are used to **formulate hypotheses**, which are then be tested using the dataset **to approve or disprove them**.

This approach ensures that our predictive model is grounded in both **empirical data and existing literature**.

Data Origin

Drinking Dataset with body signal

This dataset is collected from National Health Insurance Service in Korea
https://www.data.go.kr/data/15007122/fileData.do

sex [male, female]	age [age]	height [m]	weight [kg]	waistline [cm]	sight_left -
Male 53% Female 47%	20 – 85	130 – 190	25 – 140	8 – 999	0.1 – 9.9
Male	35	170	75	90.0	1.0
Male	30	180	80	89.0	0.9
Male	40	165	75	91.0	1.2
Male	50	175	80	91.0	1.5
Male	50	165	60	80.0	1.0
Male	50	165	55	75.0	1.2
Female	45	150	55	69.0	0.5
Male	35	175	65	84.2	1.2
Male	55	170	75	84.0	1.2
Male	40	175	75	82.0	1.5
Male	45	155	55	79.2	1.0
Male	65	155	75	98.0	1.2
Female	55	150	55	72.3	1.2
Male	30	175	75	88.0	1.2
Female	30	160	50	76.0	0.9
Male	40	170	65	80.0	1.0
Female	25	160	65	73.0	1.2
Male	25	170	65	78.0	1.2

Motivation

By using **machine learning** to predict alcohol consumption based on specific **physical parameters**, we can gain a deeper understanding of the factors influencing **alcohol consumption behavior**. This knowledge can become the basis for targeted interventions and policies aimed at **mitigating the negative effects of alcohol abuse**.

Methods

01. Literature Review and Hypothesis Formation

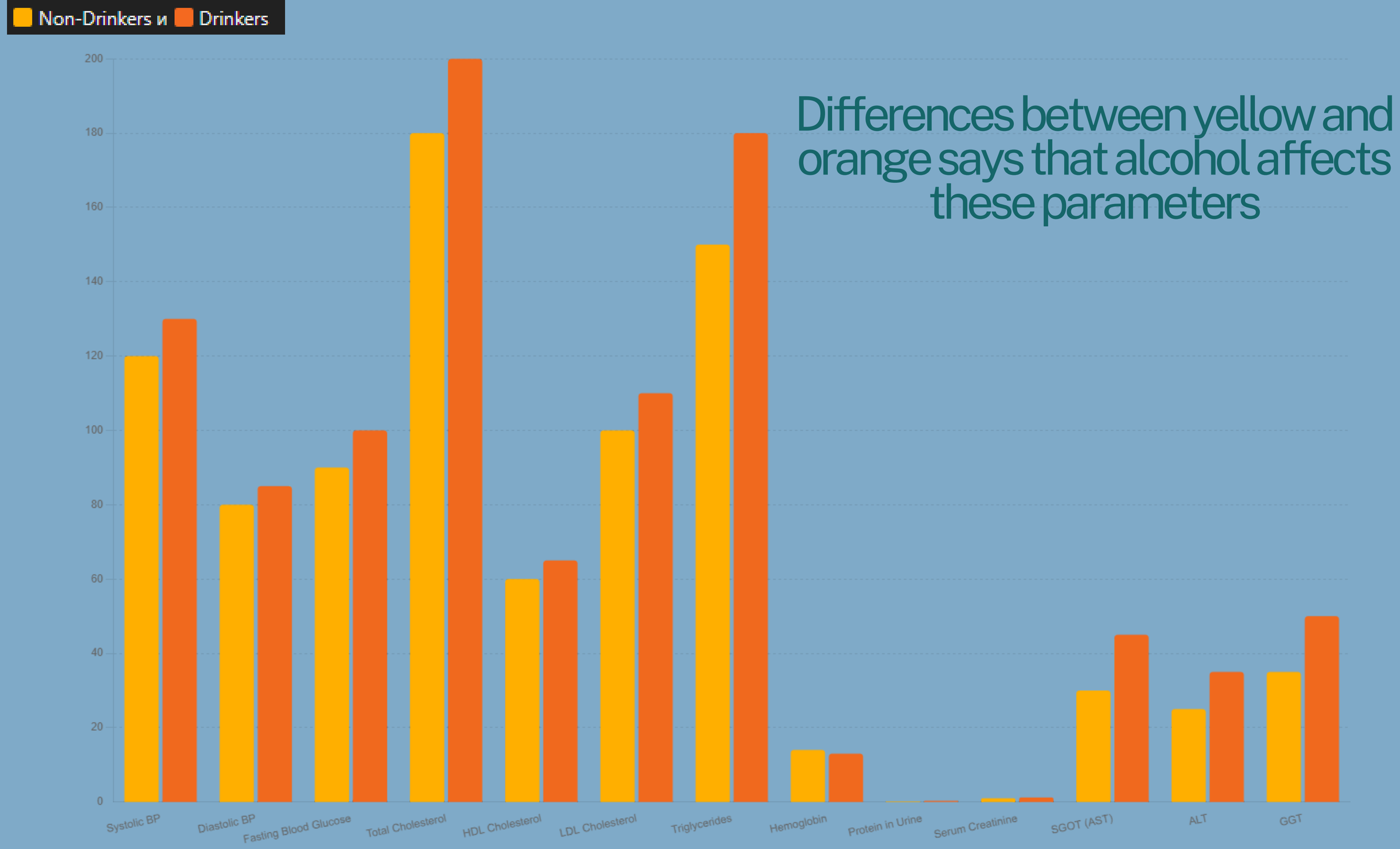
02. Pattern Research and Data Analysis

03. Preprocessing and Feature selection

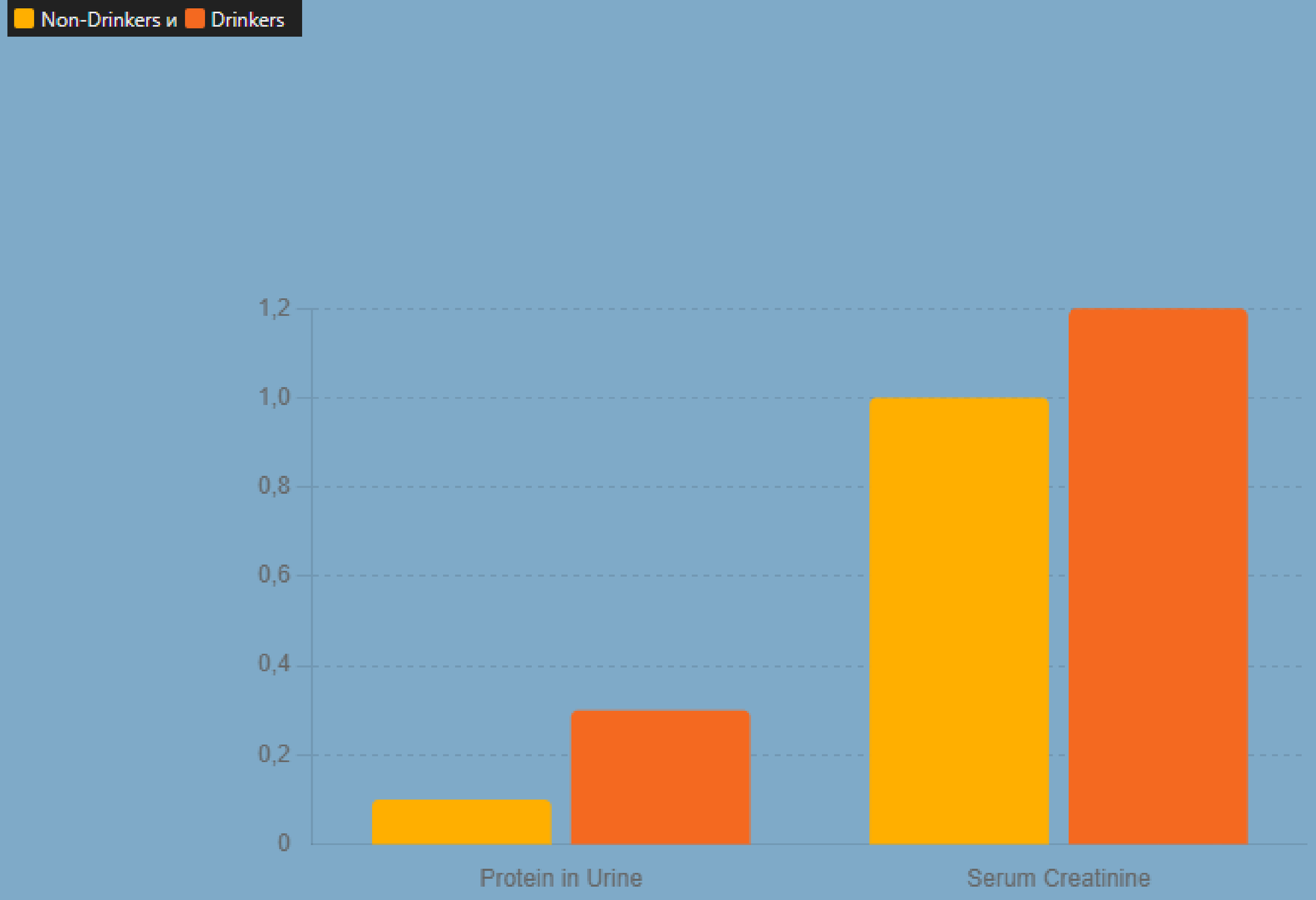
04. Model Training and Validation

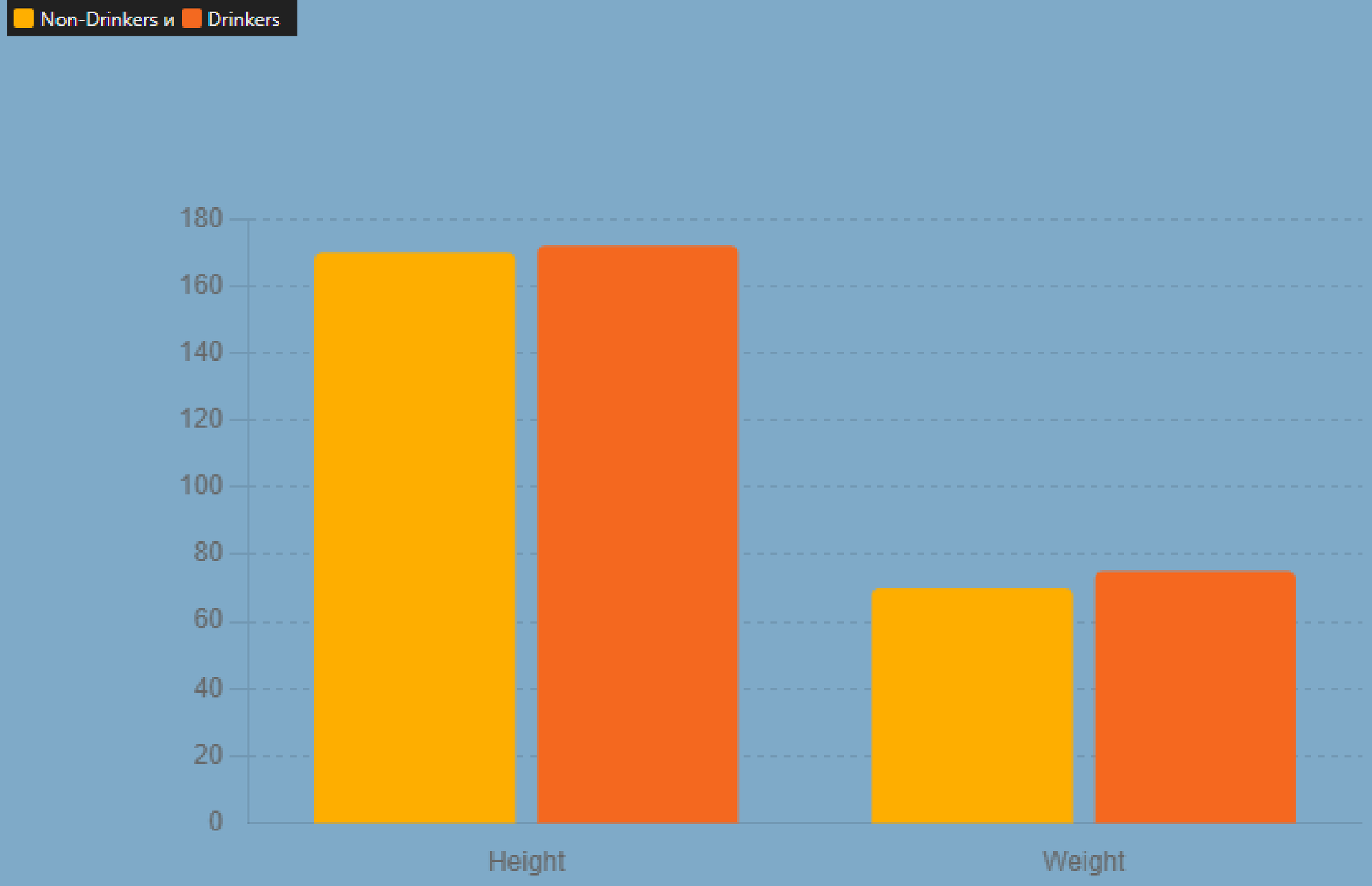
Literature Review and Hypothesis Formation

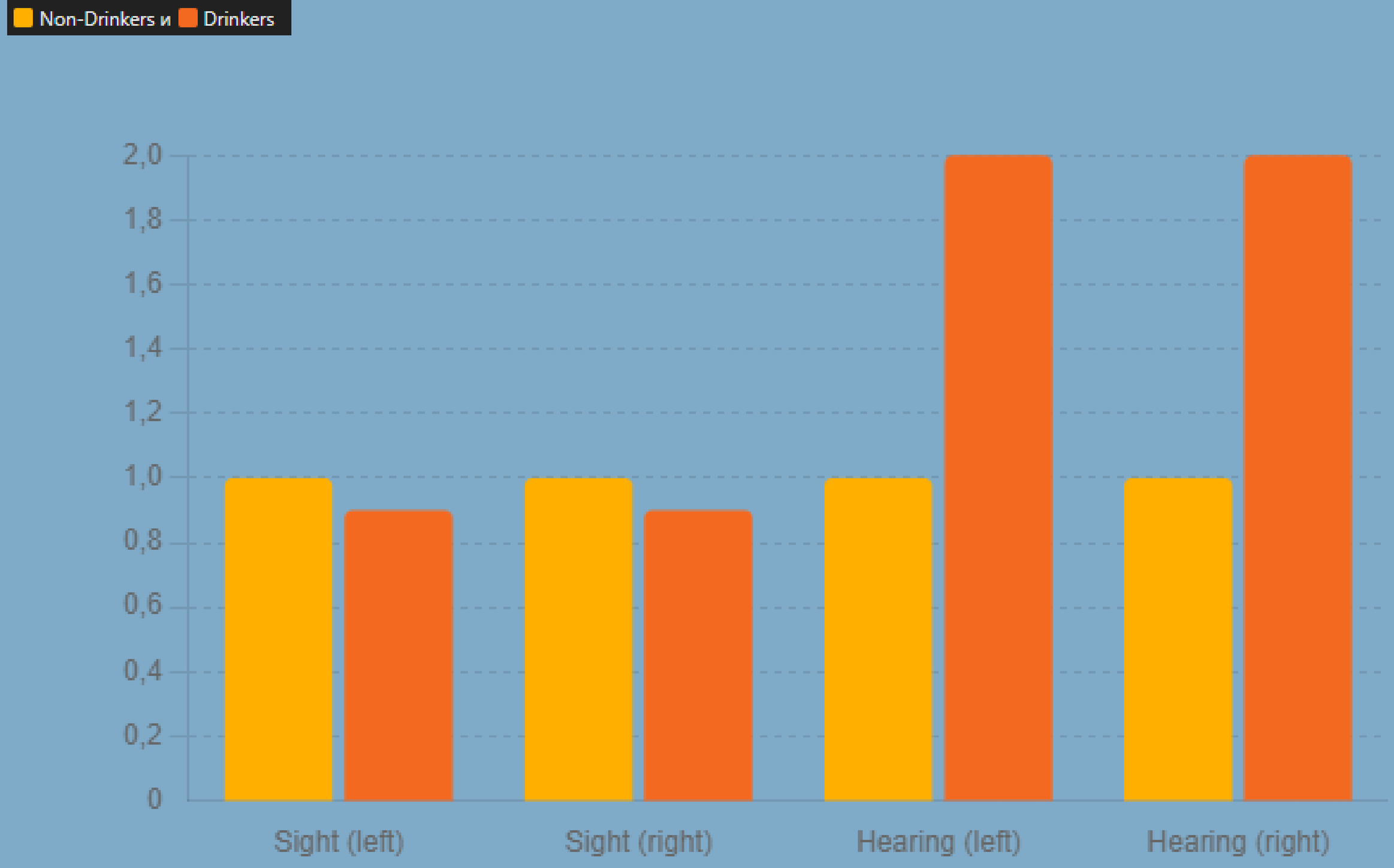
National Library of Medicine of United States of America



Differences between yellow and orange says that alcohol affects these parameters


Non-Drinkers и
Drinkers
1,2
1,0
0,8
0,6
0,4
0,2
0
Protein in Urine
Serum Creatinine


Non-Drinkers и
Drinkers
180
160
140
120
100
80
60
40
20
0
Height
Weight


Non-Drinkers и
Drinkers
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0
Sight (left)
Sight (right)
Hearing (left)
Hearing (right)

Hypothesis

Parameters affected by drinking alcohol:

Systolic BP
Diastolic BP
Fasting Blood Glucose
HDL Cholesterol
LDL Cholesterol
Triglicerides
SGOT(AST)
ALT
Gamma GTP
Protein in Urine
Serum Creatinine
Hearing

Parameters not affected by drinking alcohol:

Height
Weight

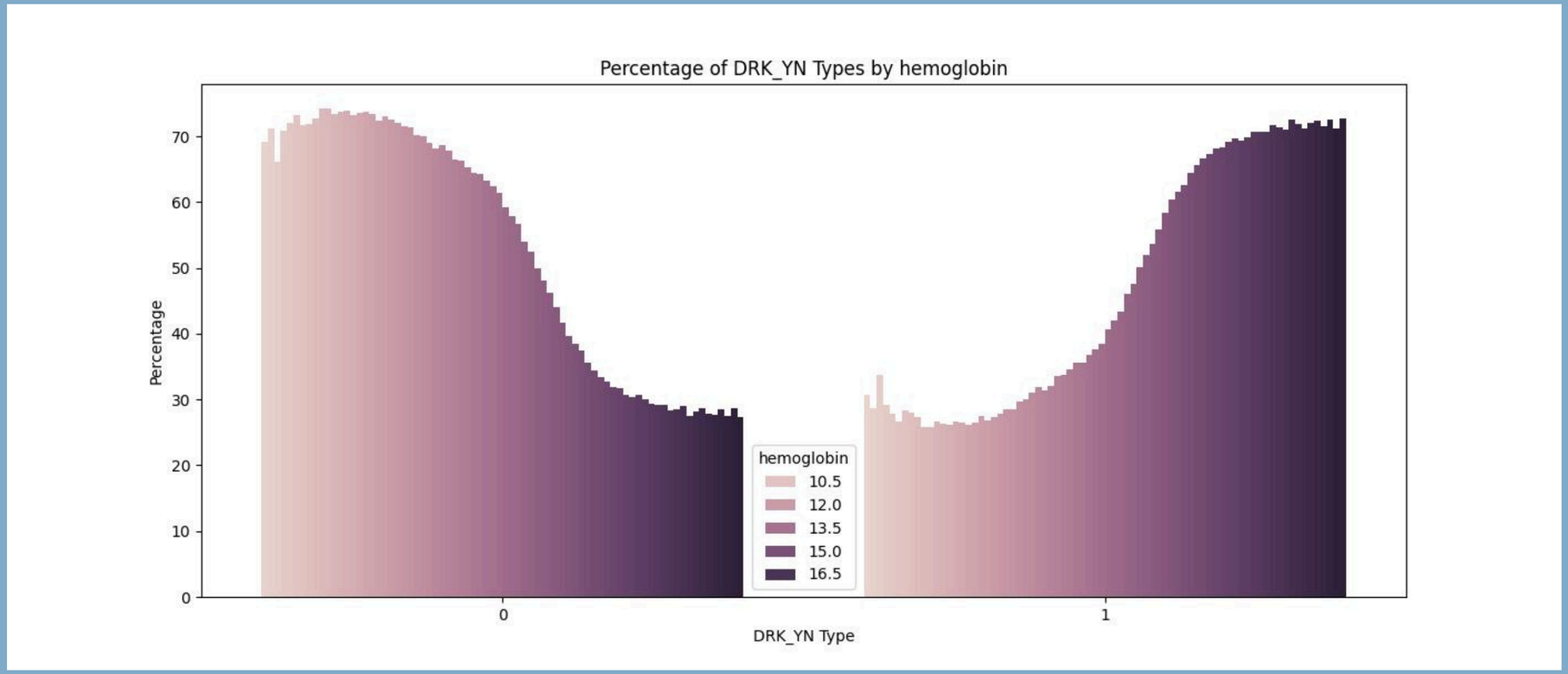
Parameters table

Parameter	Description
Sex	Male, Female
Age	Round up to 5 years
Height	Round up to 5 cm [cm]
Weight	[kg]
Sight (left)	Eyesight (left)
Sight (right)	Eyesight (right)
Hearing (left)	Hearing left, 1(normal), 2(abnormal)
Hearing (right)	Hearing right, 1(normal), 2(abnormal)
Systolic blood pressure (SBP)	Systolic blood pressure [mmHg]
Diastolic blood pressure (DBP)	Diastolic blood pressure [mmHg]
Fasting blood glucose (BLDS or FSG)	BLDS or FSG (fasting blood glucose) [mg/dL]
Total cholesterol	Total cholesterol [mg/dL]
HDL cholesterol	HDL cholesterol [mg/dL]
LDL cholesterol	LDL cholesterol [mg/dL]
Triglyceride	Triglyceride [mg/dL]
Hemoglobin	Hemoglobin [g/dL]
Protein in urine	Protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
Serum creatinine	Serum creatinine [mg/dL]
SGOT (AST)	SGOT (AST) [IU/L]
ALT	ALT [IU/L]
Gamma-GTP	Gamma-GTP [IU/L]
Smoking state	Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke)
Drinker or Not	Drinker or Not

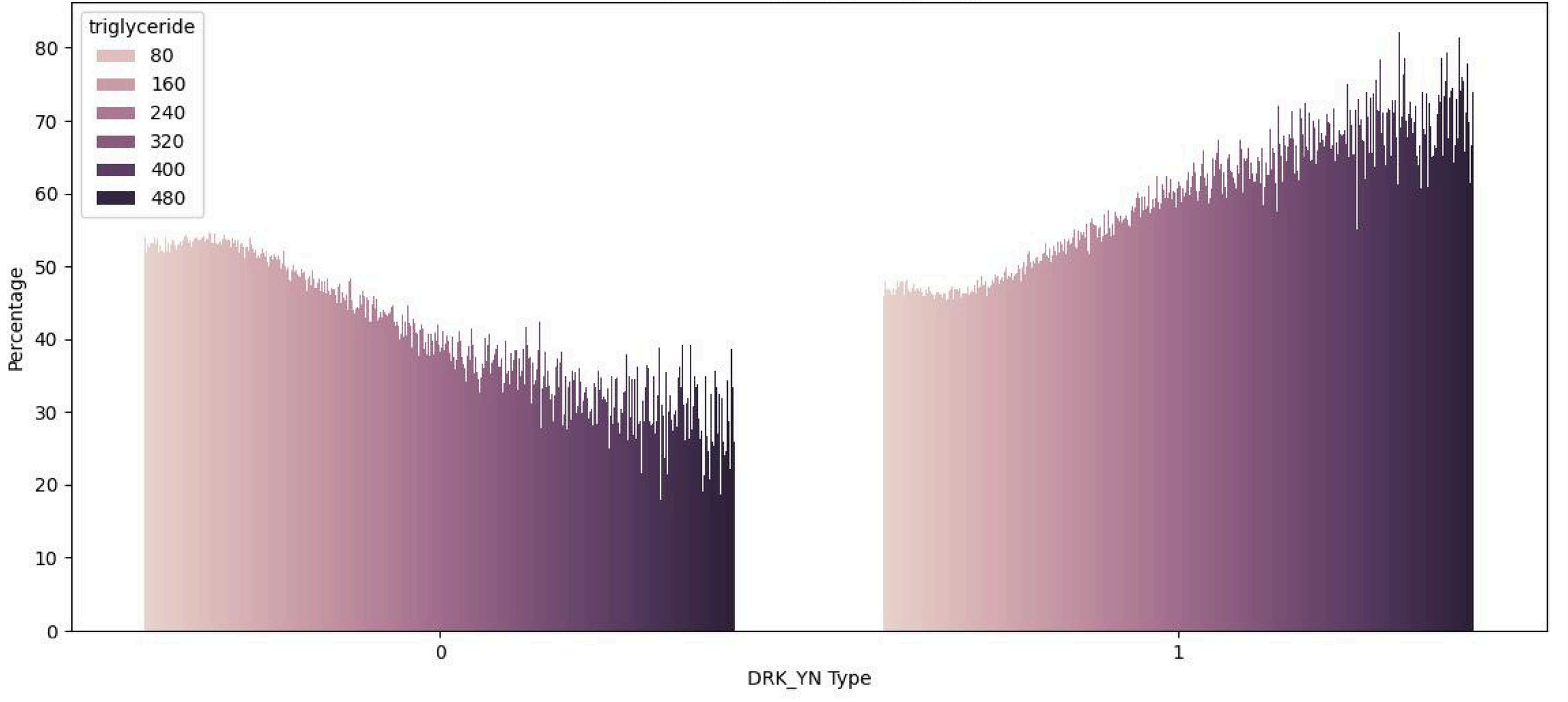
Pattern Research and Data Analysis

the e-Government of the Republic of Korea. (South)

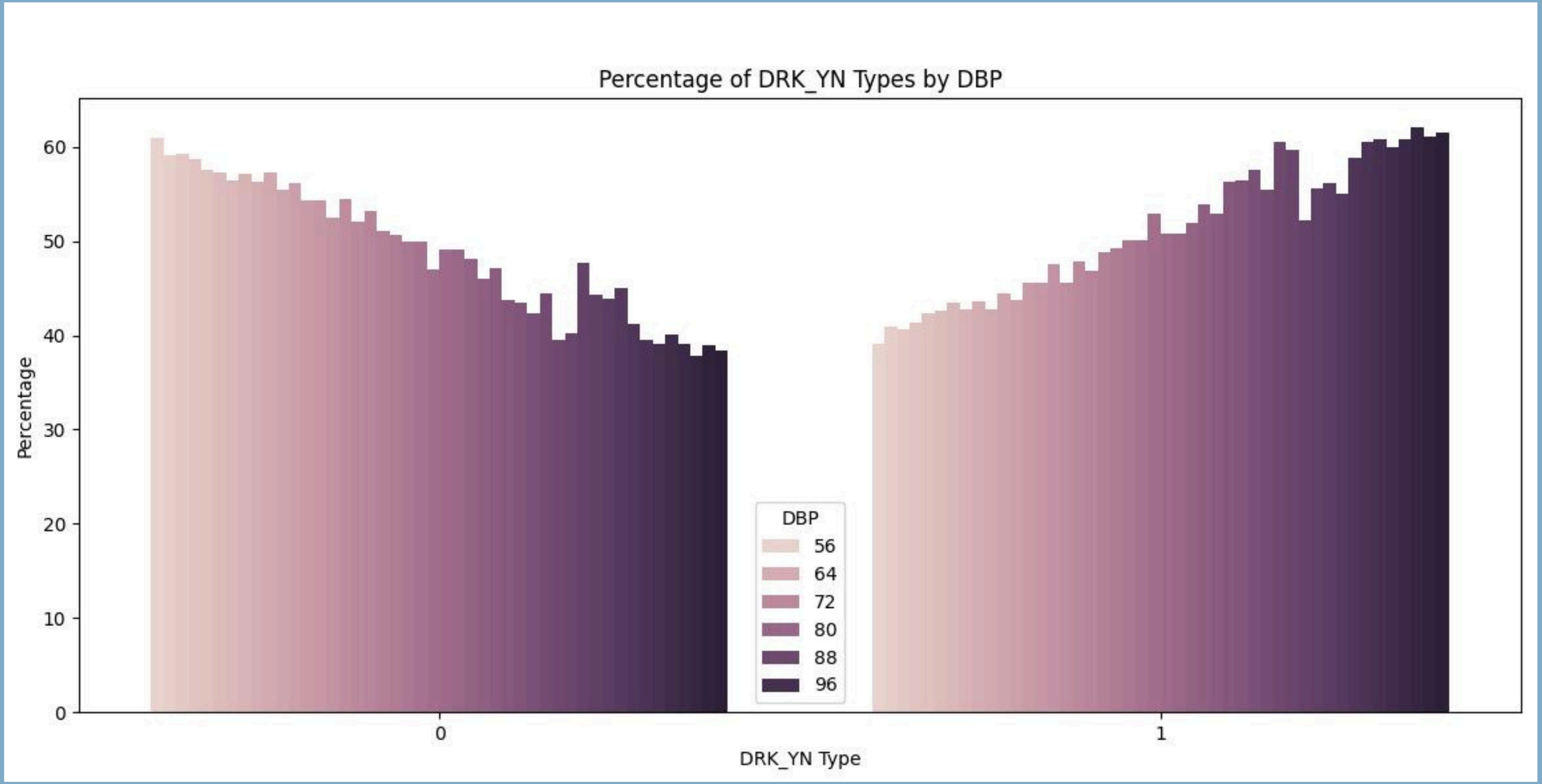
Features affected by drinking


Percentage of DRK_YN Types by hemoglobin
Percentage
70
60
50
40
30
20
10
0
0
1
DRK_YN Type
hemoglobin
10.5
12.0
13.5
15.0
16.5

Triglyceride percentage

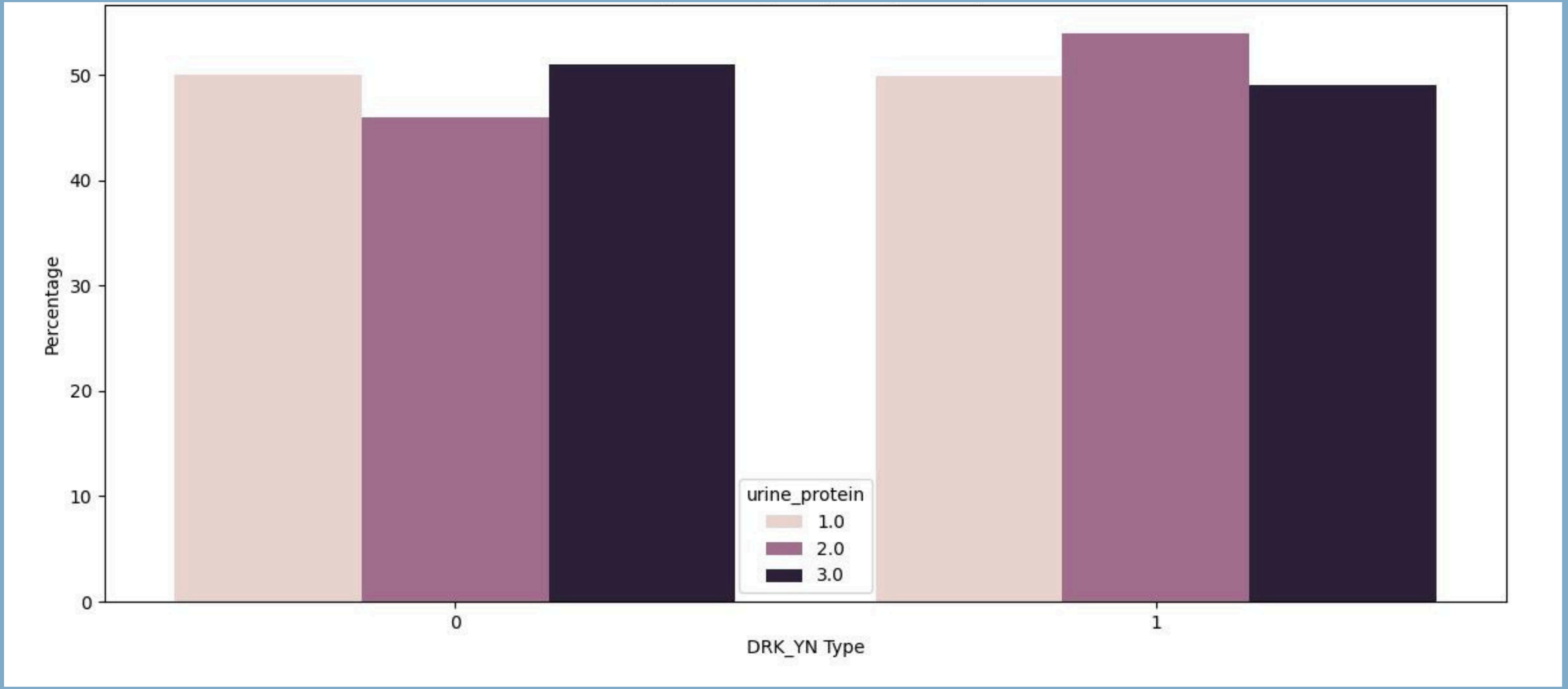


Diastolic Blood pressure

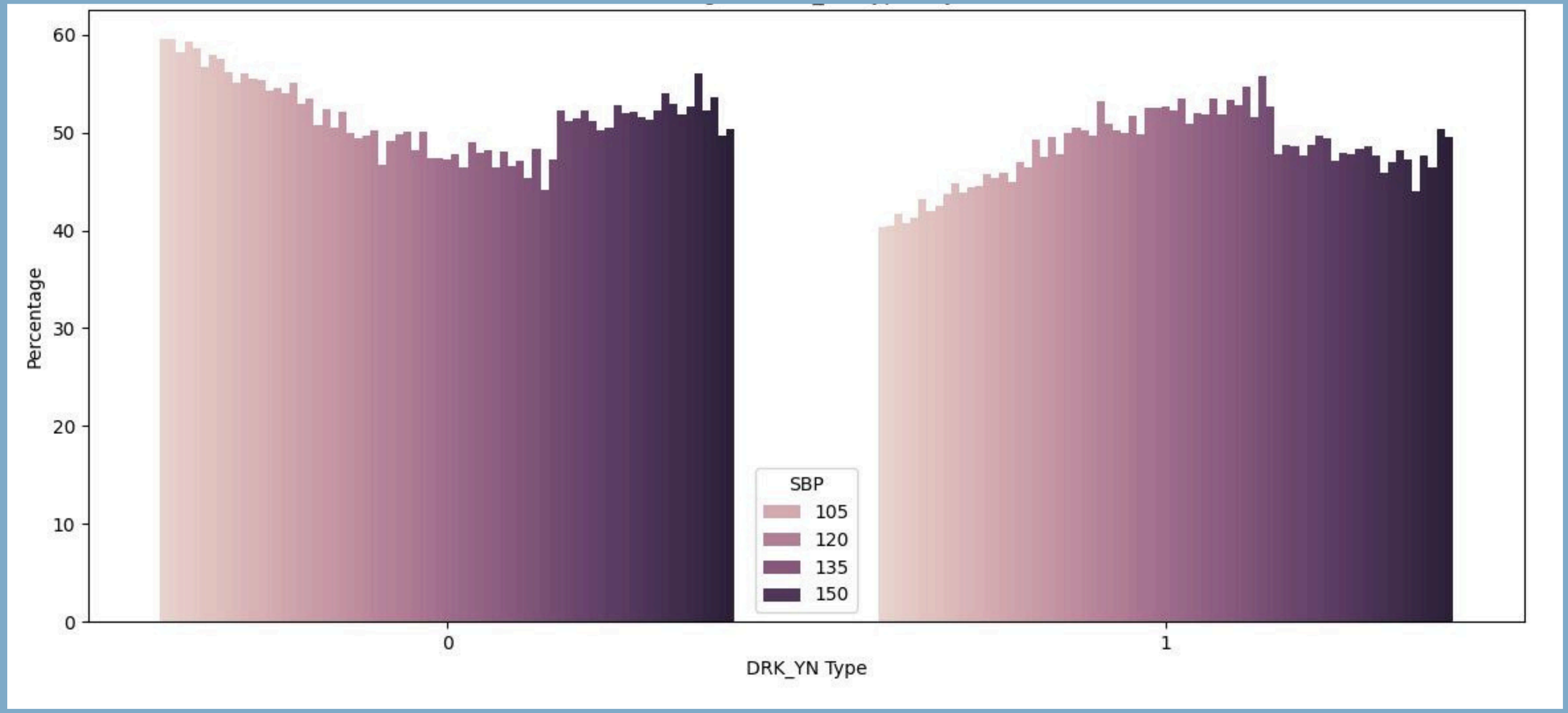


Features not affected by drinking

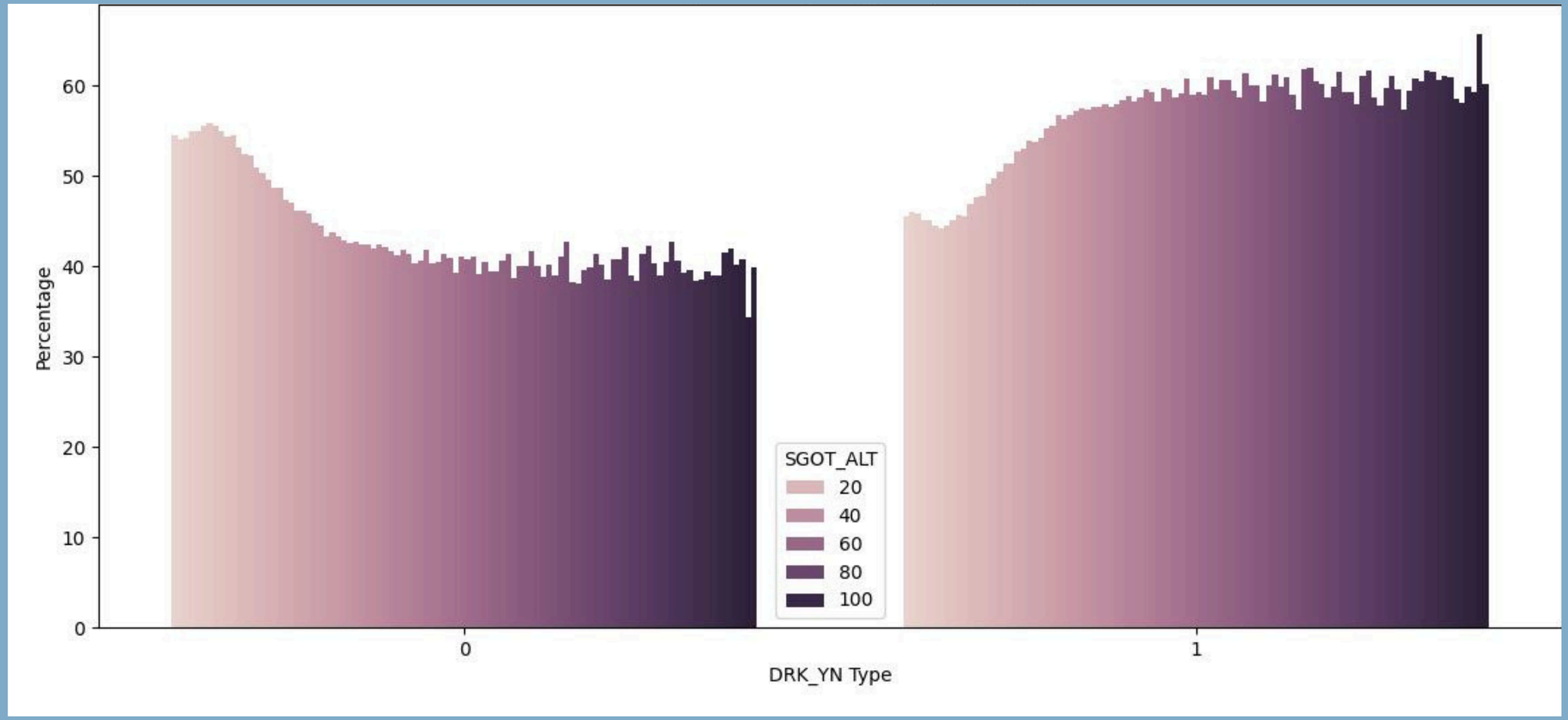
Protein in Urine



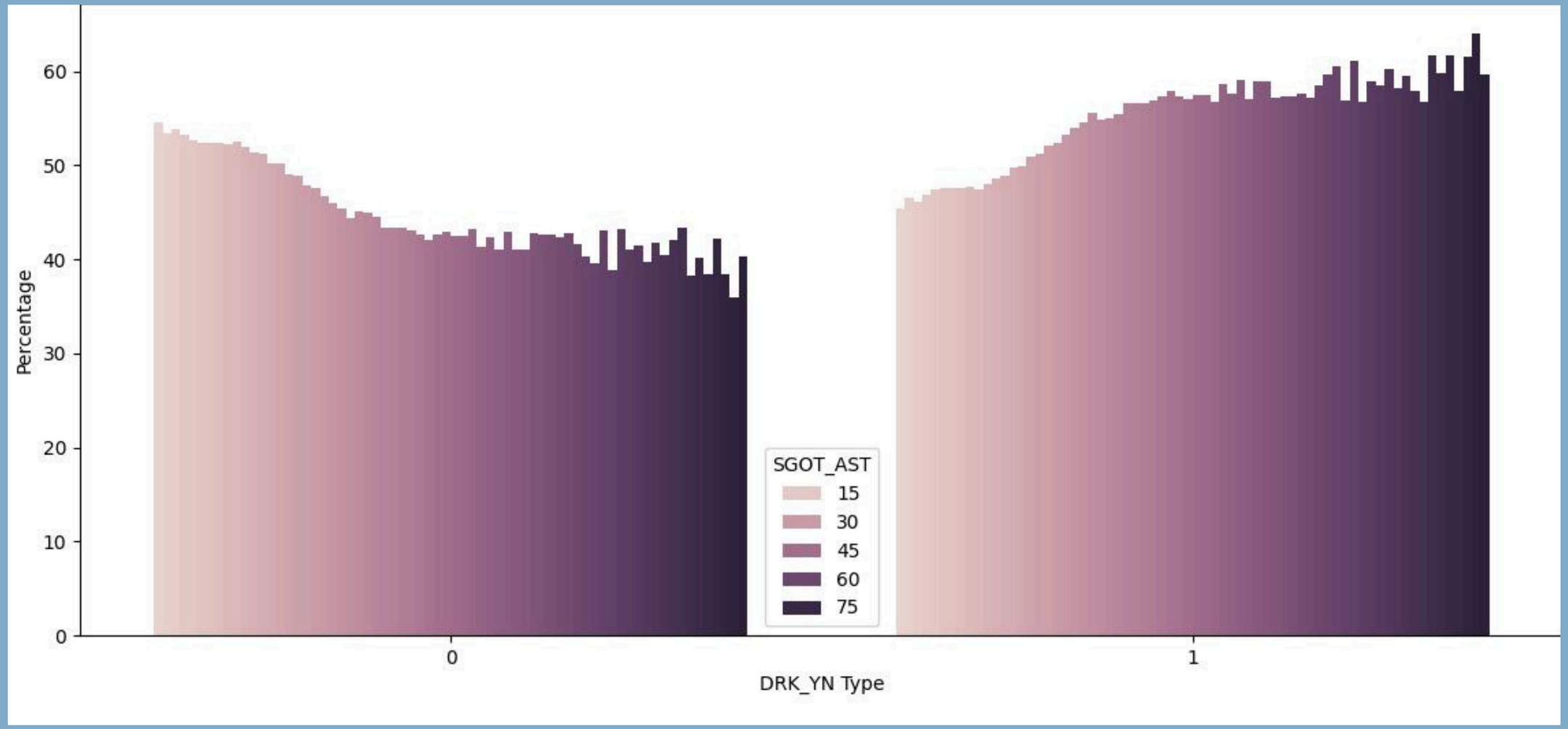
Systolic Blood Pressure



SGOT(ALT)


Percentage
60
50
40
30
20
10
0
0
1
DRK_YN Type
SGOT_ALT
20
40
60
80
100

SGOT(AST)


60
50
40
30
20
10
0
Percentage
SGOT_AST
15
30
45
60
75
0
1
DRK_YN Type

Approval and disapproval of the Hypothesis

Parameters affected by drinking alcohol:

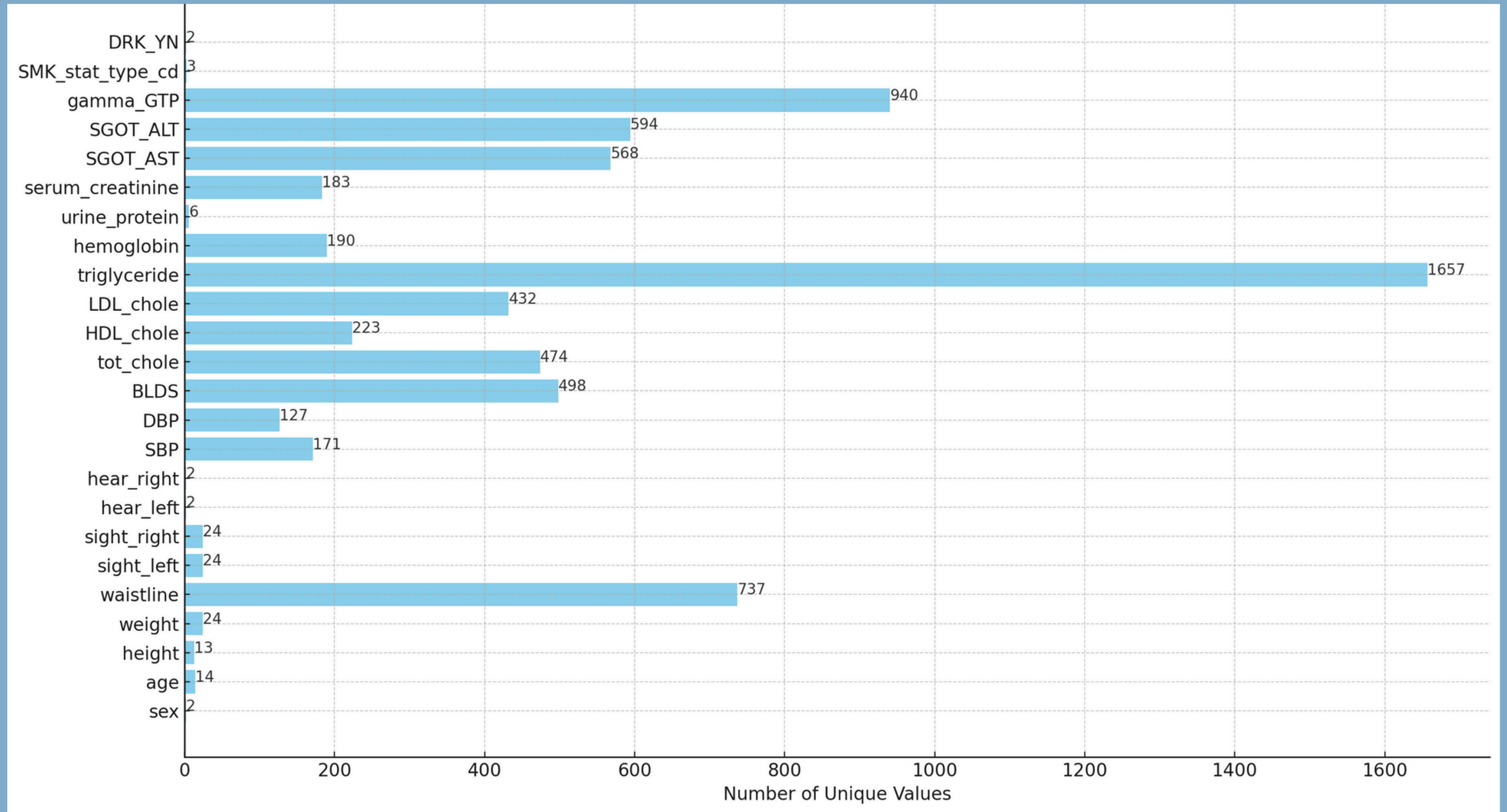
Systolic BP - is not affected
Diastolic BP
Fasting Blood Glucose
HDL Cholesterol
LDL Cholesterol
Triglicerides
SGOT(AST) - is not affected
SGOT(ALT) - is not affected
Gamma GTP
Protein in Urine - is not affected
Serum Creatinine
Hearing

Parameters not affected by drinking alcohol:

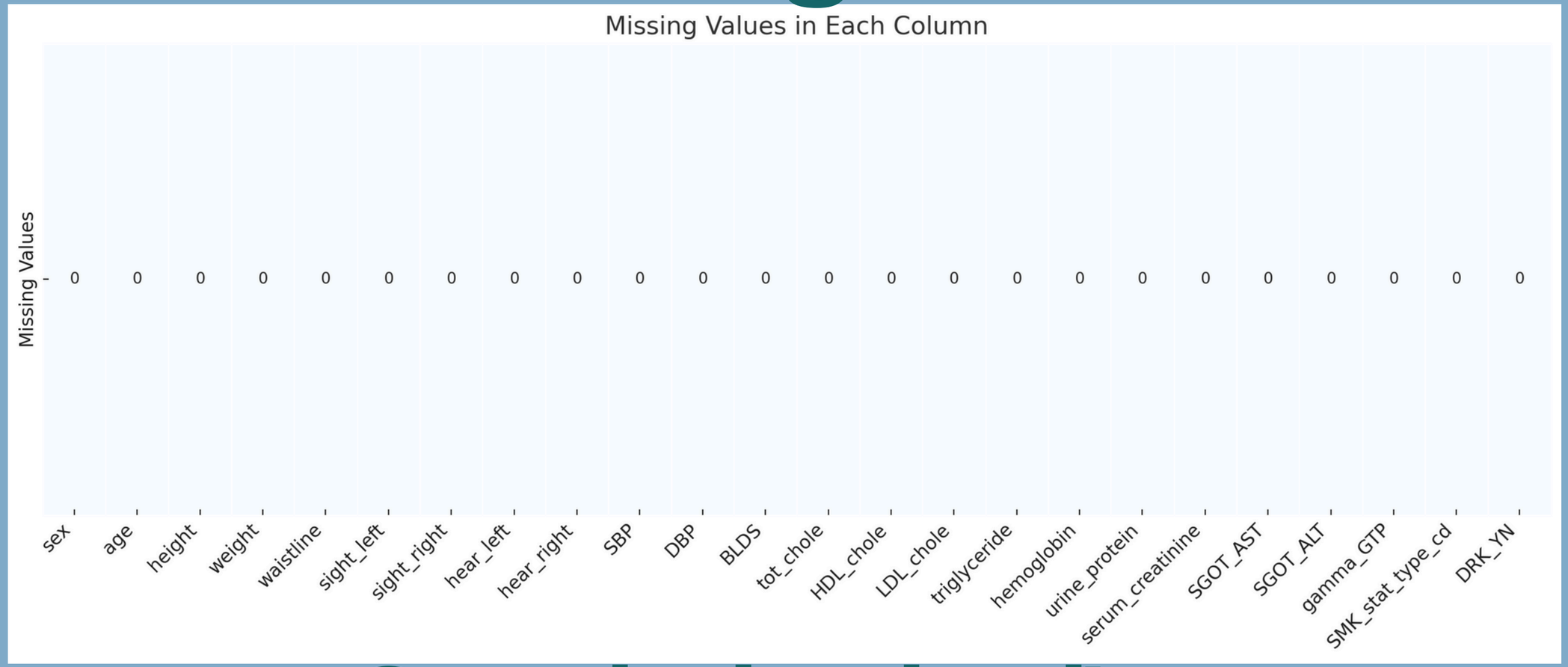
Height - is affected
Weight - is affected

Preprocessing and Feature selection

Unique values



Missing data



Sample data loading

sex	age	height	weight	waistline	SBP	DBP	BLDS	tot_chole	HDL_chole	hemoglobin	urine_protein	serum_creatinine	SGOT_AST	SGOT_ALT	gamma_GTP	SMK_stat_type	DRK_YN	sight	hear
Male	35	170	75	90.0	120.0	80.0	99.0	193.0	48.0	17.1	1.0	1.0	21.0	35.0	40.0	1.0	Y	1.0	1.0
Male	30	180	80	89.0	130.0	82.0	106.0	228.0	55.0	15.8	1.0	0.9	20.0	36.0	27.0	3.0	N	1.05	1.0
Male	40	165	75	91.0	120.0	70.0	98.0	136.0	41.0	15.8	1.0	0.9	47.0	32.0	68.0	1.0	N	1.35	1.0
Male	50	175	80	91.0	145.0	87.0	95.0	201.0	76.0	17.6	1.0	1.1	29.0	34.0	18.0	1.0	N	1.35	1.0
Male	50	165	60	80.0	138.0	82.0	101.0	199.0	61.0	13.8	1.0	0.8	19.0	12.0	25.0	1.0	N	1.1	1.0

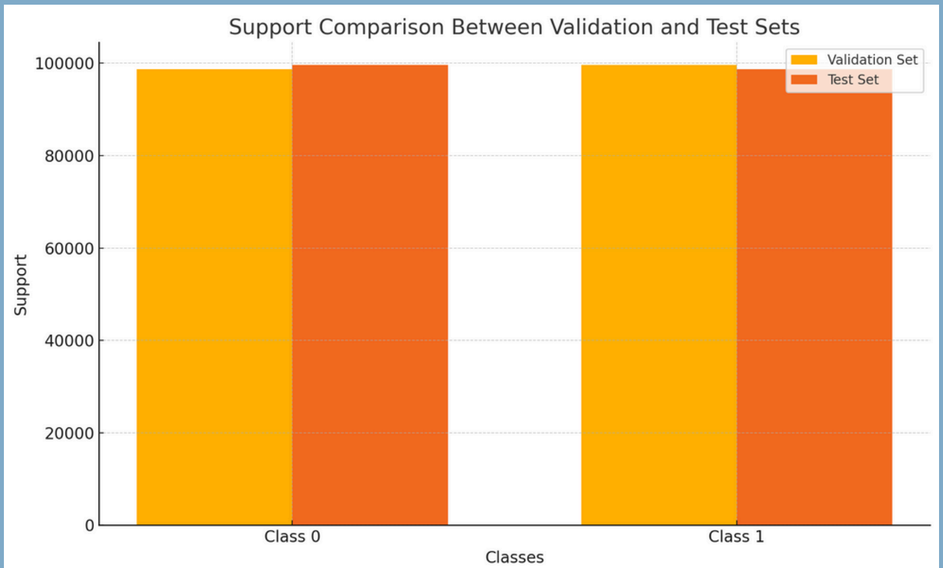
Backward feature elimination

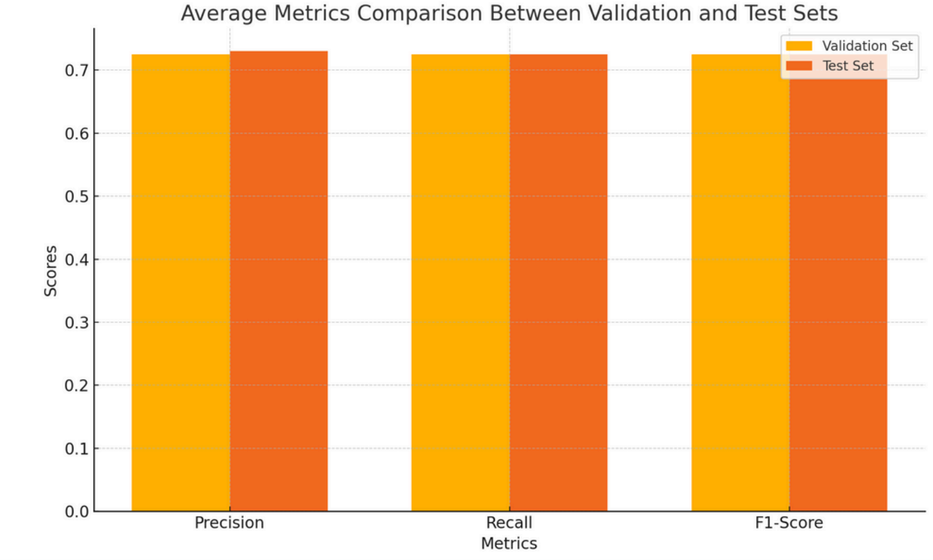
The Algorithms have given result to remain these feature to the dataset:

‘sex', 'age', 'height', 'weight', 'waistline', 'DBP', 'HDL_chole', 'LDL_chole', 'triglyceride', 'hemoglobin', 'gamma_GTP', 'SMK_stat_type_cd', 'sight', 'hear'

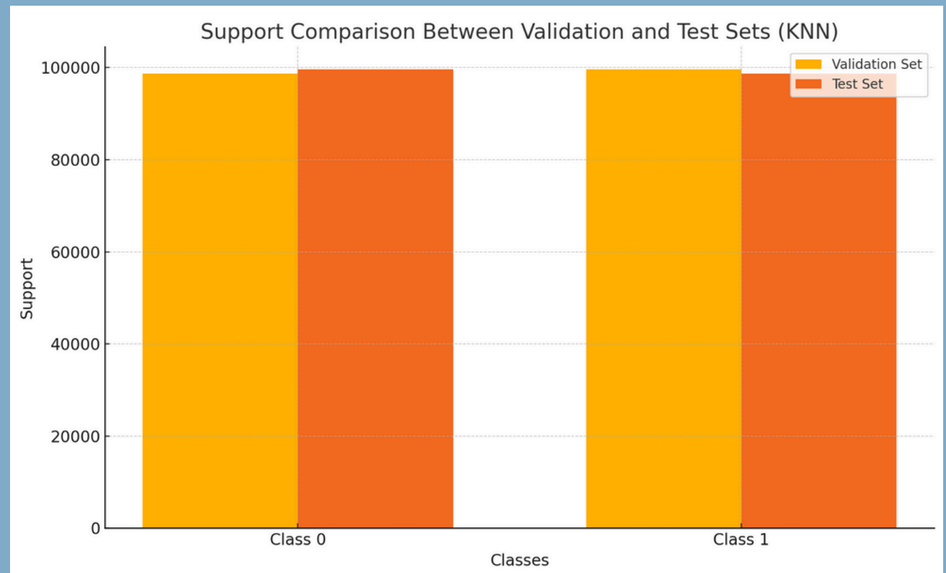
Model Training and Validation

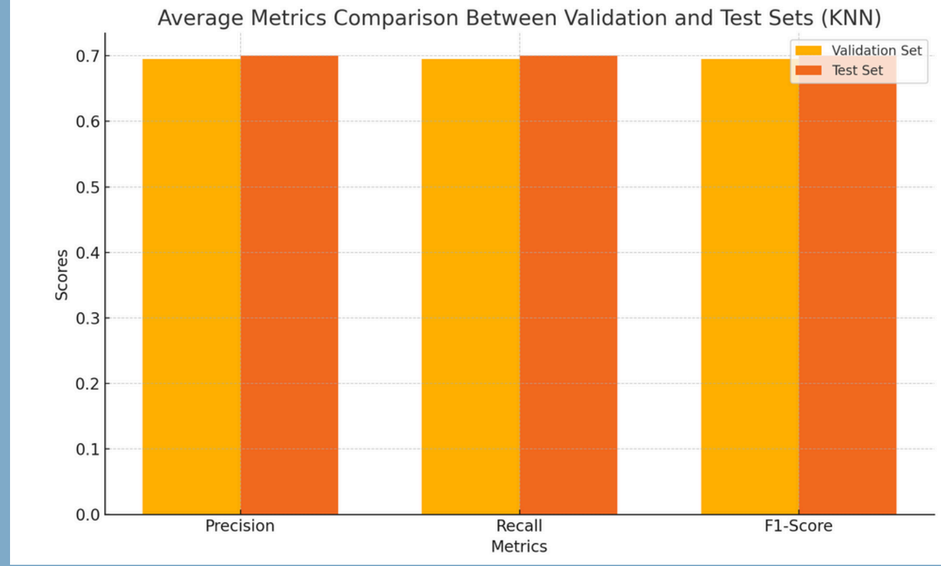
Random Forest



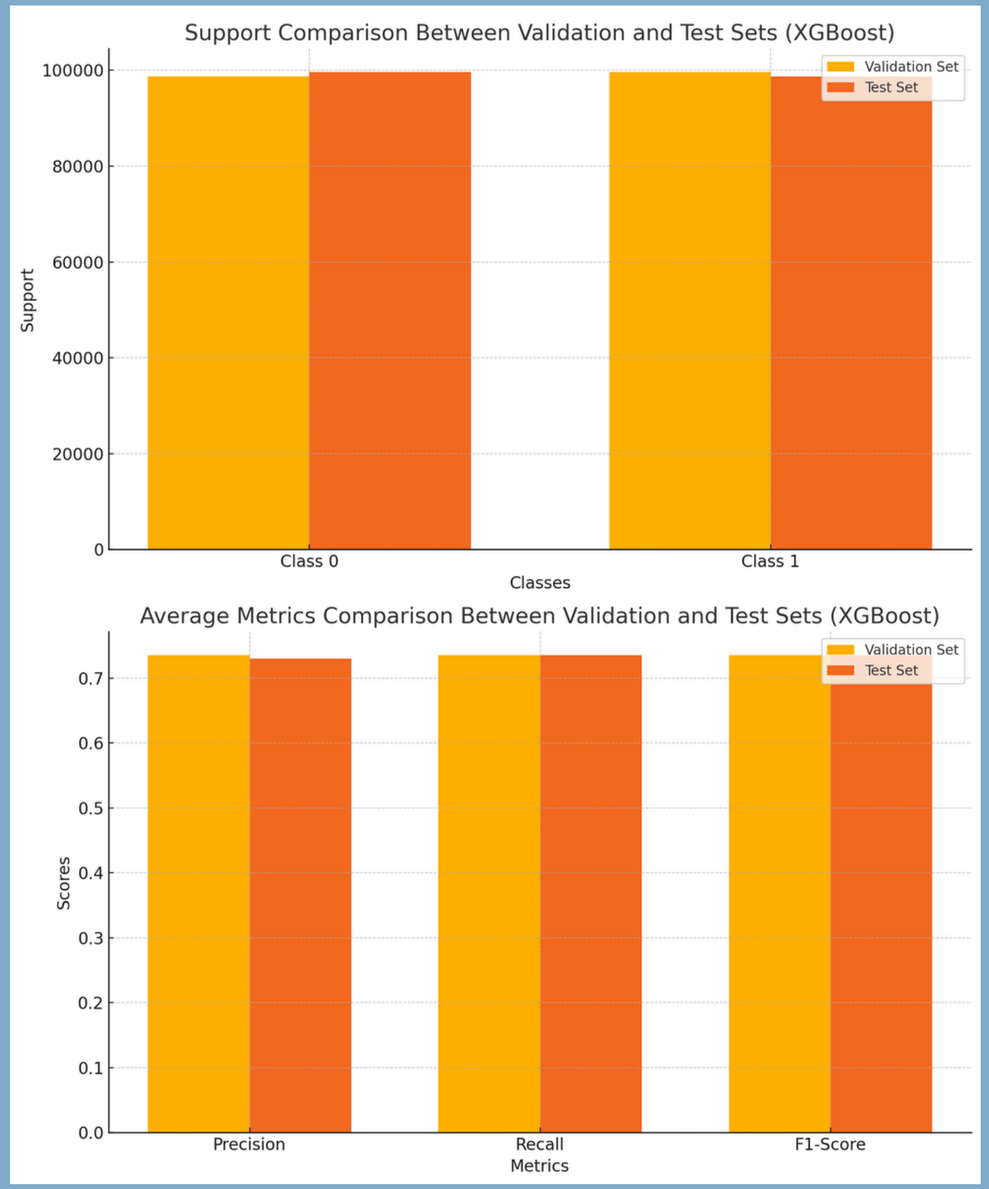


KNN





XGBoost


Support Comparison Between Validation and Test Sets (XGBoost)
Validation Set
Test Set
100000
80000
60000
40000
20000
0
Support
Class 0
Class 1
Classes
Average Metrics Comparison Between Validation and Test Sets (XGBoost)
Validation Set
Test Set
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Scores
Precision
Recall
F1-Score
Metrics

Random Forest
Validation Accuracy: 0.7285152999208147
Test Accuracy: 0.7276743834165532

KNN
Validation Accuracy: 0.695675
Test Accuracy: 0.69885

XGBoost
Validation Accuracy: 0.7329032778699646
Test Accuracy: 0.7322186916830584

For each:
Number of training samples: 594807
Number of validation samples: 198269
Number of test samples: 198270

No specification about who drinks how often

References

Systolic and Diastolic Blood Pressure
https://pubmed.ncbi.nlm.nih.gov/32216847/
Cholesterol Levels (Total, HDL, LDL, Triglycerides)
https://pubmed.ncbi.nlm.nih.gov/30591638/
Hemoglobin
https://pubmed.ncbi.nlm.nih.gov/30409799/

Protein in Urine and Serum Creatinine
https://pubmed.ncbi.nlm.nih.gov/31044751/
Liver Enzymes (SGOT (AST), ALT, Gamma-GTP):
https://pubmed.ncbi.nlm.nih.gov/31119674/
Height and Weight:
https://pubmed.ncbi.nlm.nih.gov/30571171/
Sight and Hearing:
https://pubmed.ncbi.nlm.nih.gov/29772234/