

Универзитет у Београду  
Електротехнички факултет



# МОДЕЛ ЗА ПРЕДИКЦИЈУ ИНФАРКТА

## – ИЗВЕШТАЈ –

***Студент:***

Милинковић Андреа 3102/23

***Професор:***

Проф. др Мирослав Бојовић

***Предмет:***

Софтверско инжењерство великих база података

Београд 2024. године

# САДРЖАЈ

1. Увод.....	3
2. Анализа података.....	4
2.1. Преглед улазног фајла.....	4
2.2. Преглед улазних параметара .....	6
▪ gender.....	6
▪ hypertension .....	7
▪ heart_disease.....	7
▪ ever_married .....	8
▪ Residence_type .....	9
▪ work_type.....	9
▪ smoking_status .....	10
▪ age.....	11
▪ avg_glucose_level.....	12
▪ bmi .....	13
▪ stroke.....	14
3. Обрада података .....	15
3.1. Трансформација параметара .....	15
4. Избор модела.....	17
4.1. Random Forest .....	17
4.2. Light Gradient Boosting Machine .....	18
4.3. Logistic Regression .....	18
4.4. К најближих суседа .....	19
5. Преглед резултата .....	20
6. Закључак .....	21

# 1. Увод

Циљ овог пројекта је да се кроз анализу улазних параметара предвиди вероватноћа да особа доживи инфаркт. На основу два улазна фајла *train.csv* и *healthcare-dataset-stroke-data.csv* урађена је анализа параметара. Неке од параметара било је потребно додатно средити како би се добио што бољи резултат. Коришћено је неколико алгоритама машинског учења на датом проблему и њихови резултати су међусобно упоређени како би се одредио модел који даје најбоље резултате. Пројекат је рађен у програмском језику *Python*.

Рад се састоји од неколико целина чији је опис дат у наставку.

Прво поглавље је увод у проблематику пројекта. Дат је кратак опис проблема који је потребно решити, а који се додатно анализира у наставку. У уводу је дат и преглед свих поглавља извештаја.

Друго поглавље бави се анализом података. Дат је опис свих улазних параметара и на основу њиховог додатног проучавања одређено је који су параметри корисни, колико њих имају вредности које одступају од неке уобичајене вредности за дати параметар, колико има *NaN* вредности итд. Дат је приказ свих улазних параметара, њихова подела по категоријама ако су у питању категорички подаци, или њихова расподела уколико су у питању нумерички подаци и објашњено је на који начин они утичу на циљну променљиву.

У трећем поглављу приказана је обрада података тако што је приказана трансформација категоричких података у нумеричке вредности, док је над нумеричким подацима урађено скалирање на опсег од 0 до 1. Додатно, дата је слика *Python* кода који извршава горенаведене радње.

Четврти део рада даје опис алгоритама коришћених за тренирање модела који се користе у пројекту.

У петом поглављу анализирани су резултати више врста модела који су коришћени у пројекту. Њиховим упоређивањем одабран је један који има највећу тачност.

## 2. Анализа података

У овом поглављу биће описани и анализирани параметри дати у улазном фајлу *train.csv*. Дате су слике из конзоле као и графици који служе за додатну анализу проблема.

### 2.1. Преглед улазног фајла

Скуп улазних података који је дат у изазову садржи велики број испитаника (15304) који су описани следећим параметрима:

- **id** – јединствен идентификатор
- **gender** – пол испитаника (“Male”, “Female”, “Other”)
- **age** – број година испитаника
- **hypertension** – 0 ако испитаник нема хипертензију, 1 ако има хипертензију
- **heart\_disease** – 0 ако испитаник нема ниједну срчану ману, 1 ако има
- **ever\_married** – да ли је испитаник икад био у браку (“Yes”, “No”)
- **work\_type** – “children”, “Govt\_jov”, “Never\_worked”, “Private”, “Self-employed”.
- **Residence\_type** – да ли испитаник живи у руралној или урбаној средини (“Rural”, “Urban”)
- **avg\_glucose\_level** – просечан ниво шећера у крви
- **bmi** – индекс телесне масе (*body mass index*)
- **smoking\_status** – „formerly smoked”, “never smoked”, “smokes”, “Unknown”.
- **stroke** – 1 ако је испитаник доживео инфаркт, 0 ако није

Улазни фајл *train.csv* се састоји од 15304 редова и 11 колона. На сликама 2.1. и 2.2. приказано је првих пет и последњих пет редова овог фајла. Подаци из овог фајла даље су коришћени за тренирање модела.

First 5 rows:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
id											
0	Male	28.0	0	0	Yes	Private	Urban	79.53	31.1	never smoked	0
1	Male	33.0	0	0	Yes	Private	Rural	78.44	23.9	formerly smoked	0
2	Female	42.0	0	0	Yes	Private	Rural	103.00	40.3	Unknown	0
3	Male	56.0	0	0	Yes	Private	Urban	64.87	28.8	never smoked	0
4	Female	24.0	0	0	No	Private	Rural	73.36	28.8	never smoked	0

Слика 2.1 – Приказ првих пет редова улазног фајла *train.csv*

Last 5 rows:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
id											
15299	Female	22.0	0	0	No	Govt_job	Urban	72.63	19.5	never smoked	0
15300	Female	46.0	1	0	Yes	Private	Urban	101.19	32.1	never smoked	0
15301	Female	75.0	0	0	Yes	Self-employed	Urban	87.69	26.2	never smoked	0
15302	Male	46.0	0	0	Yes	Private	Rural	101.13	22.5	Unknown	0
15303	Female	14.0	0	0	No	Private	Rural	85.12	24.7	never smoked	0

Слика 2.2 – Приказ последњих пет редова улазног фајла *train.csv*

На слици 2.3. приказане су опште статистичке информације о нумеричким атрибутима.

General statistic information about attributes:

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	15304.000000	15304.000000	15304.000000	15304.000000	15304.000000	15304.000000
mean	41.417708	0.049726	0.023327	89.039853	28.112721	0.041296
std	21.444673	0.217384	0.150946	25.476102	6.722315	0.198981
min	0.080000	0.000000	0.000000	55.220000	10.300000	0.000000
25%	26.000000	0.000000	0.000000	74.900000	23.500000	0.000000
50%	43.000000	0.000000	0.000000	85.120000	27.600000	0.000000
75%	57.000000	0.000000	0.000000	96.980000	32.000000	0.000000
max	82.000000	1.000000	1.000000	267.600000	80.100000	1.000000

Слика 2.3 – Приказ општих статистичких информације о атрибутима

На слици 2.4. приказане су опште информације о колонама, број редова који су попуњени по колонама, као и тип података који се налазе у свакој колони. Видимо да су сви подаци попуњени тј. да не постоје *NaN* вредности (слика 2.5) што значајно олакшава даљи рад јер није потребно водити рачуна о начину попуњавања таквих вредности.

```
Info about train_dataset:

<class 'pandas.core.frame.DataFrame'>
Index: 15304 entries, 0 to 15303
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   gender                 15304 non-null object  
1   age                   15304 non-null float64
2   hypertension           15304 non-null int64  
3   heart_disease          15304 non-null int64  
4   ever_married           15304 non-null object  
5   work_type              15304 non-null object  
6   Residence_type         15304 non-null object  
7   avg_glucose_level      15304 non-null float64
8   bmi                   15304 non-null float64
9   smoking_status         15304 non-null object  
10  stroke                 15304 non-null int64  
dtypes: float64(3), int64(3), object(5)
memory usage: 1.4+ MB
None
```

Слика 2.4 – Приказ информација о улазним подацима

```
Number of missing values for each column:

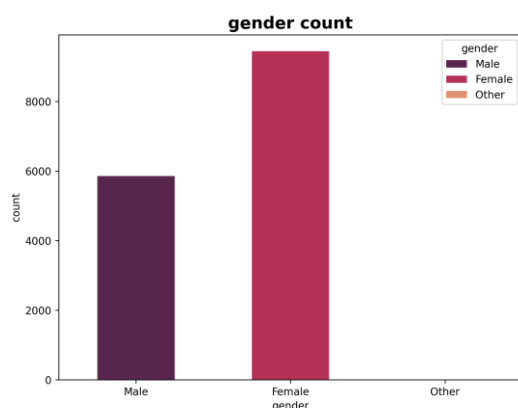
gender                0
age                   0
hypertension          0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                   0
smoking_status         0
stroke                 0
dtype: int64
```

Слика 2.5 – Приказ броја недостајућих вредности за сваку колону

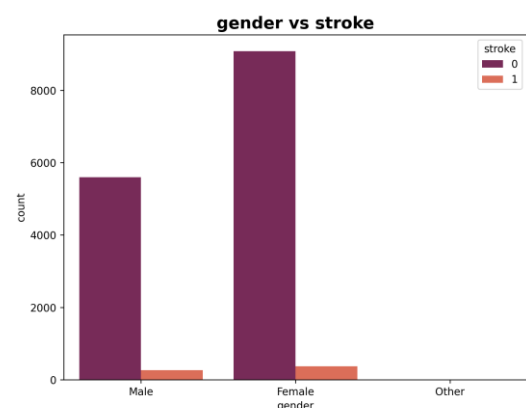
## 2.2. Преглед улазних параметара

У наставку рада биће приказане информације о сваком улазном атрибуту појединачно. Посматрајући ове атрибуте можемо закључити како они, и у којој мери утичу на крајњу предикцију инфаркта.

### ■ gender



Слика 2.6 – Визуелни приказ поделе испитаника на основу пола



Слика 2.7 – Визуелни приказ односа атрибута *gender* и *stroke*

<i>gender</i>	<i>Female</i>	<i>Male</i>	<i>Other</i>
<i>count</i>	9446	5857	1

Табела 2.1 – Приказ броја испитаника по полу

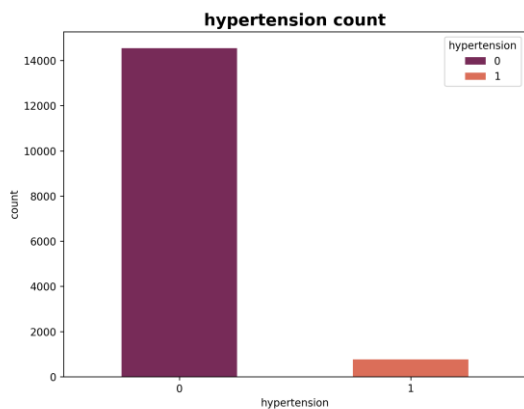
<i>gender \ stroke</i>	<i>0</i>	<i>1</i>
<i>Female</i>	0.961042	0.038958
<i>Male</i>	0.954926	0.045074
<i>Other</i>	1.000000	0.000000
<i>All</i>	0.958704	0.041296

Табела 2.2 – Приказ поделе испитаника на основу пола и доживљеног инфаркта у процентима

Са слике 2.6. можемо приметити да постоје три пола („*Female*“, „*Male*“, „*Other*“) и како су испитаници подељени по категоријама. Из табеле 2.1. видимо тачан број испитаника у свакој групи. Број жена у истраживању је доста већи од броја мушкараца, а као „*Other*“ се изјашњава само један испитаник.

Можемо приметити на основу графика са слике 2.7. и табеле 2.2. да 96% процената жена и 95% мушкараца није доживело инфаркт.

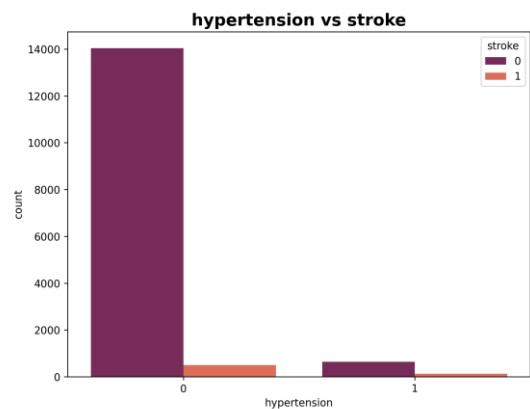
## ▪ hypertension



Слика 2.8 – Визуелни приказ броја испитаника који имају хипертензију

<i>hypertension</i>	<i>0</i>	<i>1</i>
<i>count</i>	14543	761

Табела 2.3 – Приказ броја испитаника у зависности од тога имају ли хипертензију или не



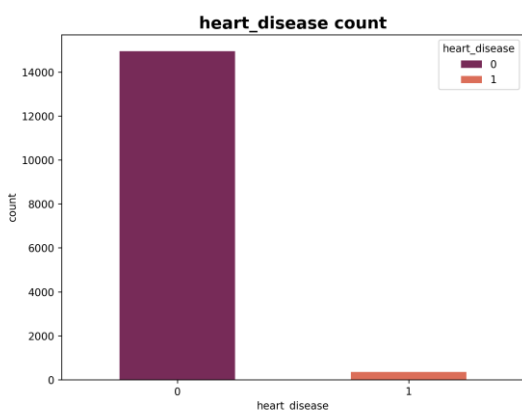
Слика 2.9 – Визуелни приказ односа атрибута *hypertension* и *stroke*

<i>hypertension \ stroke</i>	<i>0</i>	<i>1</i>
<i>0</i>	0.965344	0.034656
<i>1</i>	0.831800	0.168200
<i>All</i>	0.958704	0.041296

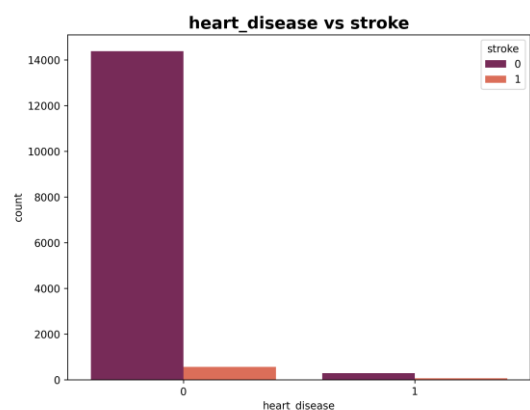
Табела 2.4 – Приказ поделе испитаника на основу хипертензије и доживљеног инфаркта у процентима

Видимо са слике 2.8. и из табеле 2.3. да је број испитаника који нема хипертензију (*14543*) значајно већи од броја оних који имају (*761*). Такође примећујемо да они људи који пате од хипертензије имају већу вероватноћу да доживе инфаркт (слика 2.9. и табела 2.4).

## ▪ heart\_disease



Слика 2.10 – Визуелни приказ броја испитаника који имају срчану ману



Слика 2.11 – Визуелни приказ односа атрибута *heart\_disease* и *stroke*

<i>heart_disease</i>	<i>0</i>	<i>1</i>
<i>count</i>	14947	357

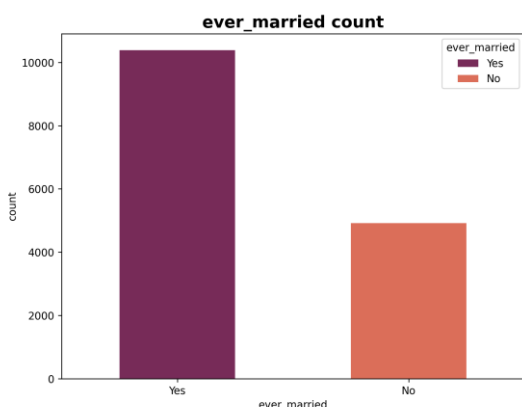
Табела 2.5 – Приказ броја испитаника у зависности од тога да ли имају срчану ману или не

<i>heart_disease</i> \ <i>stroke</i>	<i>0</i>	<i>1</i>
<i>0</i>	0.962133	0.037867
<i>1</i>	0.815126	0.184874
<i>All</i>	0.958704	0.041296

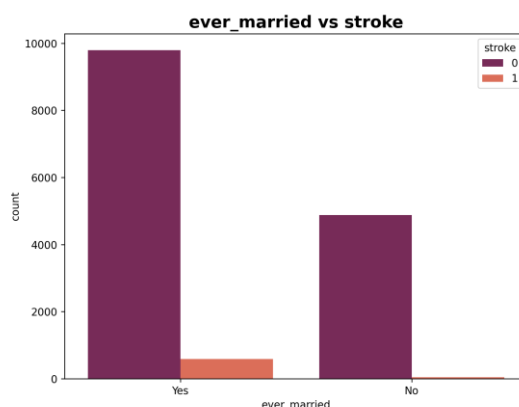
Табела 2.6 – Приказ поделе испитаника на основу постојања срчане мане и доживљеног инфаркта у процентима

Број испитаника који нема срчану ману (14947) је много већи од броја оних који је имају (357), што можемо уочити на основу графика са слике 2.10 и из табеле 2.5. Људи који имају срчану ману имају и већу шансу да доживе инфаркт (слика 2.11. и табела 2.6).

## ▪ ever\_married



Слика 2.12 – Визуелни приказ поделе испитаника на основу брачног статуса



Слика 2.13 – Визуелни приказ односа атрибута *ever\_married* и *stroke*

<i>ever_married</i>	<i>No</i>	<i>Yes</i>
<i>count</i>	4919	10385

Табела 2.7 – Приказ броја испитаника у зависности од тога да ли су били у браку или не

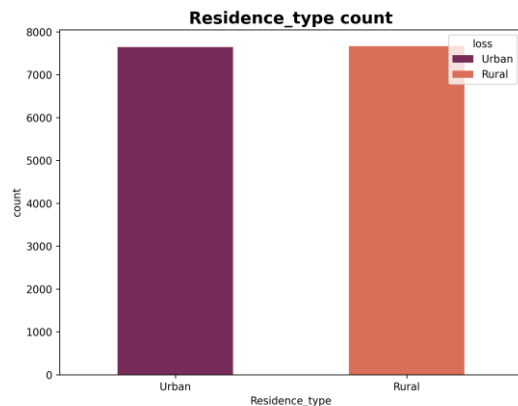
<i>ever_married</i> \ <i>stroke</i>	<i>0</i>	<i>1</i>
<i>No</i>	0.991665	0.008335
<i>Yes</i>	0.943091	0.056909
<i>All</i>	0.958704	0.041296

Табела 2.8 – Приказ поделе испитаника на основу брачног статуса и доживљеног инфаркта у процентима

Број испитаника који никад нису били у браку (4919) је много мањи од броја оних који јесу (10385), што можемо уочити на основу графика са слике 2.12 и из табеле 2.7. Људи који су бар једном били у браку имају већу шансу да доживе инфаркт (слика 2.13. и табела 2.8).



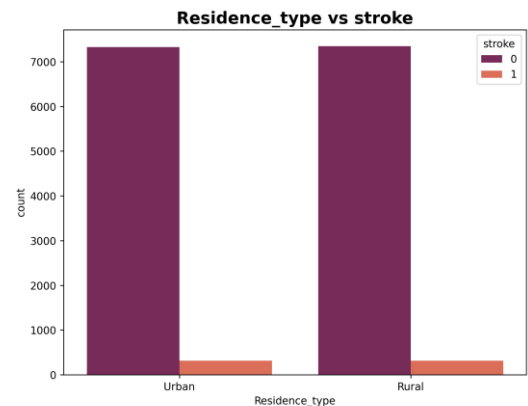
## ▪ Residence\_type



Слика 2.14 – Визуелни приказ поделе испитаника на основу типа пребивалишта

<i>Residence_type</i>	<i>Rural</i>	<i>Urban</i>
<i>count</i>	7664	7640

Табела 2.9 – Приказ броја испитаника у зависности од типа пребивалишта



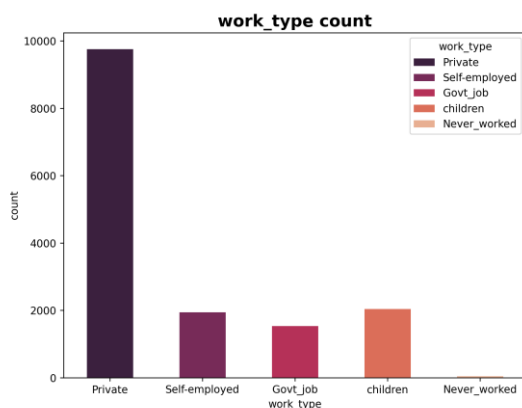
Слика 2.15 – Визуелни приказ односа атрибута *Residence\_type* и *stroke*

<i>Residence_type</i> \ <i>stroke</i>	<i>0</i>	<i>1</i>
<i>Rural</i>	0.958638	0.041362
<i>Urban</i>	0.958770	0.041230
<i>All</i>	0.958704	0.041296

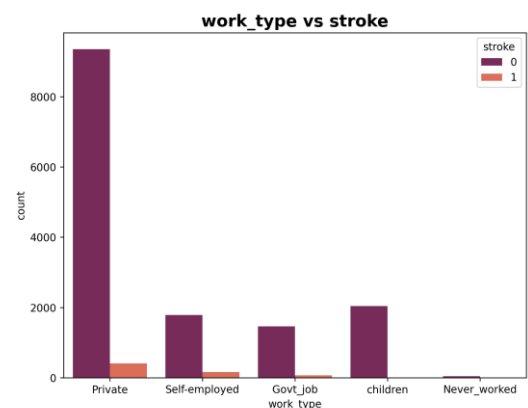
Табела 2.10 – Приказ поделе испитаника на основу типа пребивалишта и доживљеног инфаркта у процентима

Може се приметити да атрибут *Residence\_type* има подједнак утицај на резултујућу променљиву, 95.9% (слика 2.15 и табела 2.10). Самим тим можемо избацити овај атрибут из модела јер не утиче на коначни исход.

## ▪ work\_type



Слика 2.16 – Визуелни приказ поделе испитаника на основу типа посла



Слика 2.17 – Визуелни приказ односа атрибута *work\_type* и *stroke*

<i>work_type</i>	<i>Govt job</i>	<i>Never worked</i>	<i>Private</i>	<i>Self employed</i>	<i>Children</i>
<i>count</i>	1533	42	9752	1939	2038

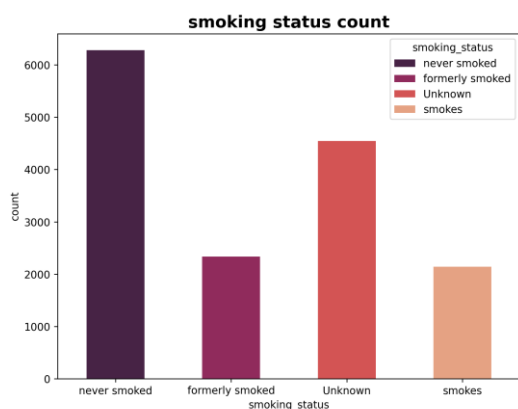
Табела 2.11 – Приказ броја испитаника у зависности од тога којом врстом посла се баве

<i>work_type</i> \ <i>stroke</i>	<i>0</i>	<i>1</i>
<i>Govt job</i>	0.954990	0.045010
<i>Never worked</i>	1.000000	0.000000
<i>Private</i>	0.958573	0.041427
<i>Self employed</i>	0.918515	0.081485
<i>Children</i>	0.999509	0.000491
<i>All</i>	0.958704	0.041296

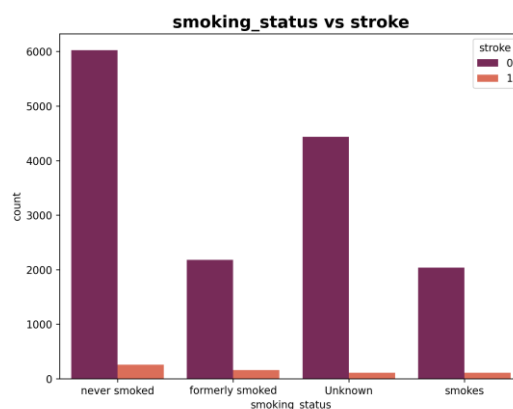
Табела 2.12 – Приказ поделе испитаника на основу типа посла и доживљеног инфаркта у процентима

Људи који никад нису радили имају најмање шансе да доживе срчани удар, док су они који су *self\_employed* под већим ризиком да га доживе (слика 2.17. и табела 2.12). Видимо из приложеног да овај атрибут има већи утицај на резултујућу промељиву.

## ■ smoking\_status



Слика 2.18 – Визуелни приказ поделе испитаника на основу тога да ли су пушачи или не



Слика 2.19 – Визуелни приказ односа атрибута *smoking\_status* и *stroke*

<i>smoking_status</i>	<i>Unknown</i>	<i>formerly smoked</i>	<i>never smoked</i>	<i>smokes</i>
<i>count</i>	4543	2337	6281	2143

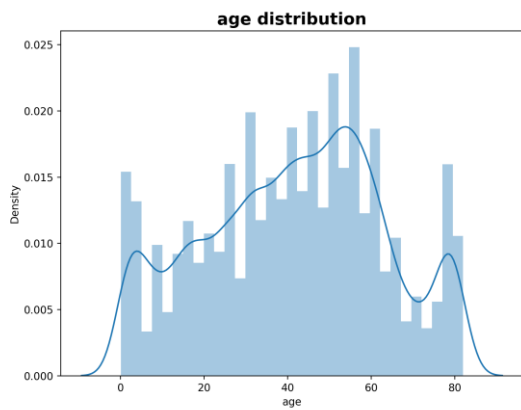
Табела 2.13 – Приказ броја испитаника у зависности од тога да ли су пушачи или не

<i>smoking_status</i> \ <i>stroke</i>	<i>0</i>	<i>1</i>
<i>Unknown</i>	0.976227	0.023773
<i>formerly smoked</i>	0.931964	0.068036
<i>never smoked</i>	0.959083	0.040917
<i>smokes</i>	0.949603	0.050397
<i>All</i>	0.958704	0.041296

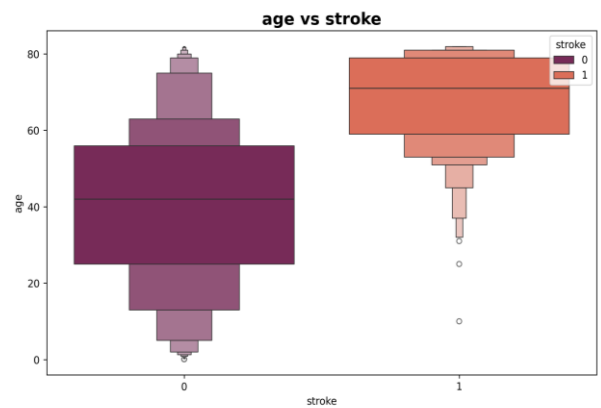
Табела 2.14 – Приказ поделе испитаника на основу тога да ли су пушачи и доживљеног инфаркта у процентима

Мању вероватноћу да доживе инфаркт имају испитаници чији је статус *Unknown* или *never smoked*, док они са статусом *formerly smoked* имају највећу шансу (табела 2.14).

▪ **age**



Слика 2.20 – Визуелни приказ расподеле испитаника на основу старости



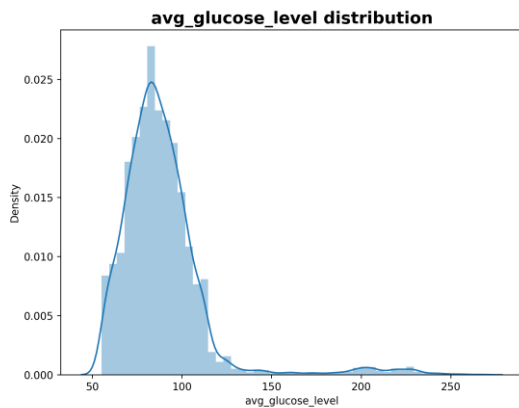
Слика 2.21 – Визуелни приказ односа атрибута *age* и *stroke*

<i>count</i>	15304.000000
<i>mean</i>	41.417708
<i>std</i>	21.444673
<i>min</i>	0.080000
<i>25%</i>	26.000000
<i>50%</i>	43.000000
<i>75%</i>	57.000000
<i>max</i>	82.000000

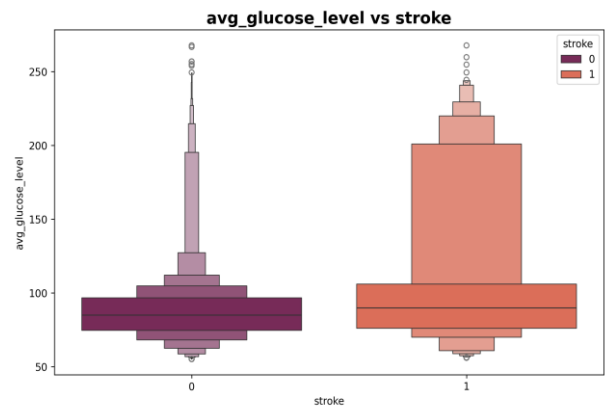
Табела 2.15 – Приказ додатних информација о расподели испитаника на основу старости

На основу слике 2.20. највећи број испитаника има између 40 и 60 година. Видимо да су испитаници који су доживели инфаркт у највећој мери имали преко 60 година (слика 2.21). На основу ових информација, можемо закључити да је овај атрибут важан за предвиђање потенцијаног инфаркта.

- **avg\_glucose\_level**



Слика 2.22 – Визуелни приказ расподеле испитаника на основу просечног нивоа шећера у крви



Слика 2.23 – Визуелни приказ односа атрибута *avg\_glucose\_level* и *stroke*

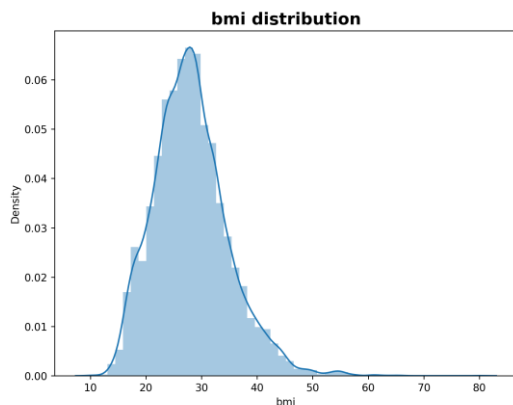
<b>count</b>	15304.000000
<b>mean</b>	89.039853
<b>std</b>	25.476102
<b>min</b>	55.220000
<b>25%</b>	74.900000
<b>50%</b>	85.120000
<b>75%</b>	96.980000
<b>max</b>	267.600000

Табела 2.16 – Приказ додатних информација о расподели испитаника на основу просечног нивоа шећера у крви

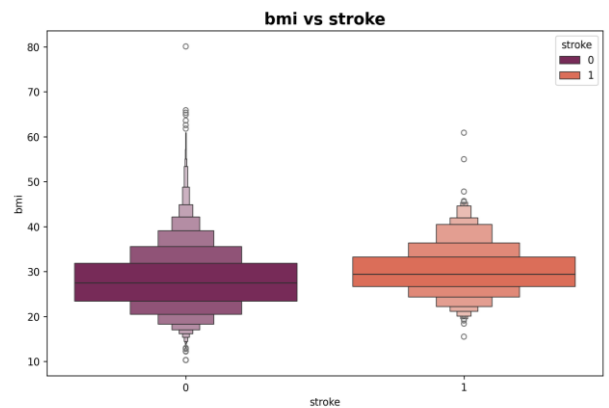
Расподела испитаника на основу просечног нивоа шећера у крви није равномерна. Видимо да је код највише испитаника ниво шећера у крви између 60 и 120, док је оних којима је ниво шећера преко 120 доста мање (слика 2.22).

Што је ниво шећера у крви већи то је вероватноћа да особа доживи инфаркт већа (слика 2.23).

## ▪ **bmi**



Слика 2.24 – Визуелни приказ расподеле испитаника на основу индекса телесне масе



Слика 2.25 – Визуелни приказ односа атрибута *bmi* и *stroke*

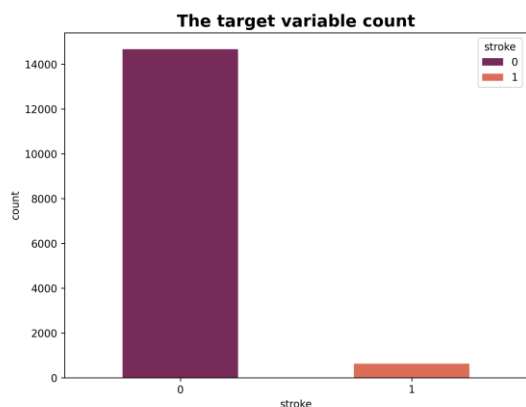
<b>count</b>	15304.000000
<b>mean</b>	28.112721
<b>std</b>	6.722315
<b>min</b>	10.300000
<b>25%</b>	23.500000
<b>50%</b>	27.600000
<b>75%</b>	32.000000
<b>max</b>	80.100000

Табела 2.17 – Приказ додатних информација о расподели испитаника на основу индекса телесне масе

Са слике 2.24. примећује се да расподела испитаника на основу индекса телесне масе није једнака. Видимо да највише испитаника има *bmi* између 25 и 35.

Овај атрибут није од превеликог значаја за предвиђање инфаркта, јер можемо приметити да иако велики број испитаника који имају низак индекс телесне масе није доживео инфаркт, такође постоје испитаници који имају висок индекс телесне масе а исто нису доживели инфаркт (слика 2.25). Закључак је да овај атрибут нећемо користити у тренирању модела.

## ■ stroke



Слика 2.26 – Визуелни приказ поделе испитаника на основу тога да ли су доживели инфаркт или не

<i>stroke</i>	<i>0</i>	<i>1</i>
<i>count</i>	14672	632

Табела 2.18 – Приказ односа броја испитаника који су доживели инфаркт и оних који нису

Видимо да је број испитаника који нису доживели инфаркт много већи од број оних који јесу. Како би резултат био што тачнији, улазном фајлу *train.csv* придружићемо додане испитанике извучене из другог улазног фајла *healthcare-dataset-stroke-data.csv*, али тако што ћемо бирати само оне испитанике који су доживели инфаркт.

## 3. Обрада података

Овај одељак бави се обрадом и трансформацијом улазних параметара од значаја како би крајња предикција била што тачнија. Из претходне анализе података видели смо да можемо да уклонимо атрибуте *Residence\_type* и *bmi*, јер они скоро да и немају утицај на крајње предвиђање, па њихово укључивање у модел неће допринети побољшању перформанси. Над категоријским подацима извршићемо енковање у нумеричке вредности како бисмо могли да их користимо у тренирању модела. Нумеричке вредности ћемо нормализовати/скалирати како би све вредности биле у истом опсегу. У те сврхе коришћен је *MinMaxScaler* како би се све вредности поставиле у опсег од 0 до 1.

### 3.1. Трансформација параметара

На слици 3.1. приказан је део *Python* кода који приказује на који начин је извршено енковање категоријских података и како су скалирани нумерички подаци.

Најпре су раздвојени нумерички подаци од категоријских и смештени у променљиве *num\_col* и *cat\_col* респективно. Затим је позвана функција *tr* која врши трансформацију ових података тако што за нумеричке податке позива *MinMaxScaler()* док се за категоријске позива *OrdinalEncoder()*.

```
num_cols = ["age", "avg_glucose_level"]
cat_cols = X.columns.difference(num_cols)

num_pipe = Pipeline([
    ('scaler', MinMaxScaler())
])

tr = ColumnTransformer([
    ("num", num_pipe, num_cols),
    ("cat", OrdinalEncoder(), cat_cols)
])

X = tr.fit_transform(X)
df_test = tr.transform(df_test)
```

Слика 3.1 – Приказ кода за трансформацију категоријских и нумеричких података

Трансформација је урађена и на скупу података за тренирање и на скупу података за тестирање. Добијени модификовани подаци се даље користе за тренирање модела и предикцију крајњег резултата.



## 4. Избор модела

У овом делу рада побројани су сви алгоритми машинског учења који су коришћени за тернирање модела за предикцију крајњег резултата проблема. Дат је кратак опис сваког од њих.

### 4.1. Random Forest

*Random Forest* је „ансамбл“ алгоритам машинског учења. Развио се као унапређење стабла одлучивања (*Decision Tree*) и усмерен је на превазилажење његових недостатака као што су *overfitting* и осетљивост на мање промене у подацима. Овај алгоритам користи више стабала одлучивања како би побољшао тачност.

Свако стабло које улази у *Random Forest* креирано је на основу различитог скупа података и за свако се бира другачији подскуп карактеристика за раздвајање по чворовима. Описани процес раздвајања се рекурзивно понавља док стабло не достигне максималну дубину или постављени лимит. Када се сва стабла обуче (дају своје предикције), *Random Forest* комбинује њихове резултате и гласањем долази до коначног резултата. У случају класификације, класа са највише гласова постаје класификација *Random Forest* модела, док у контексту регресије, најчешће се користи средња вредност свих предикција стабла. Баш због оваког начина одлучивања гласањем, овај алгоритам је прецизнији и бољи од једног стабла одлучивања.

*Random Forest* често даје добре перформансе и може се користити за различите врсте података, укључујући и оне са великим бројем атрибута. Овај алгоритам се широко користи у пракси због своје ефикасности и способности да се носи са различитим типовима података. Додатно треба напоменути да од избора броја стабала и параметара зависи да ли ћемо постићи оптималне перформансе за одређени проблем.

У пројекту је овај алгоритам коришћен у моделима *RandomForestClassifier* и *XGBRFClassifier*.

## 4.2. Light Gradient Boosting Machine

*LightGBM* је „ансамбл“ метода за појачавање која се базира на стаблима одлучивања. Као и друге методе базиране на стаблима одлучивања, *LightGBM* се може користити и за класификацију и за регресију. *LightGBM* је оптимизован за високе перформансе у дистрибуираним системима.

*LightGBM* комбинује више слабих модела како би се формирао снажан модел. Тренирање модела се врши итеративно, а свако ново стабло фокусира се на исправљање преосталих грешака модела ансамбла. Нова стабла се генеришу све док се не достигне задати критеријум (максималан број стабала или минимално побољшање функције губитка).

Свако ново стабло, уместо да расте ниво по ниво (*level-wise*), расте тако што се шири према делу са највећим добитком. Ова стратегија често резултира краћим путем ка оптималном решењу, што доприноси бржем тренирању модела.

*LightGBM* користи хистограмски приступ при чему се подаци групишу у бинове помоћу хистограма дистрибуције. Тиме се смањује време обраде и побољшава ефикасност алгоритма, посебно на великим скуповима података.

*LightGBM* је алгоритам који је ефикасан, има добре перформансе и може да подржи рад са великим бројем улазних података.

## 4.3. Logistic Regression

Логистичка регресија је тип надгледаних алгоритама машинског учења. Њену примену можемо видети код проблема класификације, када је потребно предвидети којој класи неки примерак припада на основу једног или више његових атрибута (континуалних или категоријских). Постоје три облика Логистичке регресије. Најпростији облик је бинарна логистичка регресија када је излаз једна од две могуће вредности (0 или 1). Она користи логистичку (*sigmoid*) функцију да моделује зависну излазну променљиву (вероватноћу припадности класи 1) на основу улазних података. У случају да постоји више од две могуће излазне вредности, онда је реч о мултиномијалној логистичкој регресији (*softmax*

функција уместо *sigmoid* функције), а уколико је могуће успоставити редослед излазних вредности, онда је реч о ординалној логистичкој регресији.

Логистичка регресија је генерализовани линеарни модел (утврђује одлике класа) који користи формулу линеарне зависности излазног податка од улазних, као и линеарна регресија:

$$y = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$

Ову формулу касније употребљава као аргумент *sigmoid* функције за предикцију излаза:

$$h(x) = \frac{1}{1 + e^{-y}} = \frac{1}{1 + e^{-(w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n)}}$$

Излаз Линеарне регресије може бити било која континуална вредност у опсегу  $[-\infty, +\infty]$ , док *sigmoid* функција доводи те вредности у опсег  $[0, 1]$ .

#### 4.4. К најближих суседа

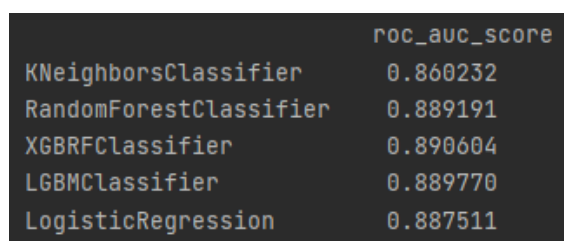
Алгоритам К најближих суседа (*K-Nearest Neighbors – KNN*) је алгоритам за класификацију који се заснива на принципу да слични подаци теже да буду један до другог у простору атрибута. Припада групи надгледаних алгоритама машинског учења. Алгоритам класификује одређену инстанцу на основу класификације К најближих инстанци простим пребројавањем сваке класе.

Принцип рада алгоритма је следећи:

- **Бирање вредности К** – важно је одабрати оптималну вредност за К
- **Мерење сличности** – за сваки податак који треба класификовати, мери се сличност између тог податка и свих података из обучавајућег скупа. Неке од најпопуларнијих метрика за проналажење најближих суседа између две тачке су Еуклидска раздаљина, Менхетн раздаљина и Чебишева раздаљина.
- **Избор К најближих суседа (KNN)** – идентификују се К најближих суседа том податку на основу мере сличности.
- **Гласање** – класа новог податка се одређује гласањем (мањинским или већинским) између класа његових К најближих суседа.

## 5. Преглед резултата

Слика 5.1. приказује поређење пет тестираних алгорита машинског учења. Евалуација је вршена на основу параметра који је дефинисан као површина испод ROC (*Receiver Operating Characteristic*) криве, која представља криву са стопом лажно позитивних (*False Positive Rate*) на x-оси и стопом тачно позитивних (*True Positive Rate*) на y-оси при свим праговима класификације.



	roc_auc_score
KNeighborsClassifier	0.860232
RandomForestClassifier	0.889191
XGBRFClassifier	0.890604
LGBMClassifier	0.889770
LogisticRegression	0.887511

Слика 5.1 – Резултати тестирања пет различитих модела

*roc\_auc\_curve* је функција која се често користи у контексту оцењивања перформанси класификационих модела. ROC-AUC (*Receiver Operating Characteristic - Area Under the Curve*) представља метрику која мери способност класификационог модела да раздвоји класе и да измери квалитет његових предикција.

Посматрајући резултате (слика 5.1) можемо закључити да најбољи резултат даје модел *XGBRFClassifier* (*Extreme Gradient Boosting Random Forest*), а одмах иза њега и *LGBMClassifier* (*Light Gradient Boosting Machine*). Коефицијент детерминације је ~0.89 што представља задовољавајући резултат с обзиром да је тренутно најбољи остварен резултат на овом изазову 0.9.

## 6. Закључак

Циљ овог рада био је да се на основу скупа улазних података предвиди вероватноћа да нека особа доживи инфаркт. Кроз рад је урађена анализа улазних података и одређено је како који од атрибута утиче на исход предвиђања. Одабрани су атрибути од значаја за моделовање, након чега је извршена њихова трансформација како би се добио што бољи резултат предвиђања. За крај, тестирано је пет различитих модела над истим скупом података како би се утврдило који од њих даје најбоље резултате.