DATA SCIENCE PORTFOLIO

By: Camytha Octanuryati Rochmad

My name is Camytha Octanuryati Rochmad. I have completed my bachelor degree in mechanical engineering major at Sampoerna University and University of Arizona in 2021. I also have a keen interest in data science and currently graduated from a Data Science bootcamp at Dibimbing. Then, I have good knowledge of python, R, and SQL languages.

Certifications

- **RevoU (Aug 2021)**

Intro to Data Analytics Mini-Course

- **Certiport (Jun-Aug 2021)**

Microsoft Office Specialist (MOS) Ms. Excel 2016

- **G2Academy (May-June 2021)**

Full-Stack Data Pre-Bootcamp

- **RMDS Narasio Data (May 2021)**

First Journey To Data Analysis RMDS Narasio Data

Data Science Mini Project

House Price Prediction – Regularized Regression

- Identify and integrate the needs of data
- Minimize the machine learning model errors using regularization
- Train multiple models using Ridge and LASSO regressions
- Evaluate the model on test data using MAE, MAPE, and RMSE

Admission Status – Exploratory Data Analysis

- Clean, manipulate, manage data
- Conduct data deep-dive understanding
- Gain insights that mostly students who have research experiments and good score in schools are admitted to the university

Data Science End-to-End Project

Stroke Prediction and Classification

Supervised Machine Learning

Tools

Python for analysis and modelling

Ms. Excel for load data and data understanding

Outline

01

Business Background and Objectives

02

Data Preparation and Feature Engineering

03

Modelling and Evaluation

04

Conclusion and Recommendation

01

Business Background and Objectives

Introduction and Problems

Stroke has become a significant global public health issue in recent years. One solution is to control metabolic factors. However, the medical staffs have difficulty predicting people getting stroke unless it is obviously abnormal. Therefore, it is necessary to predict stroke using modelling and valid data.

Objectives

- What factors affect stroke?
- What machine learning algorithms are suitable for predicting strokes?

02

Data Preparation and Feature Engineering

Dataset Information

43400

rows

11

features

Numerical

- id
- age
- hypertension
- heart_disease
- avg_glucose_level
- bmi

Categorical

- Gender
- ever_married
- work_type
- residence_type
- smoking_status

1

target

Stroke

Source dataset: https://data.mendeley.com/datasets/x8ygrw87jw/1

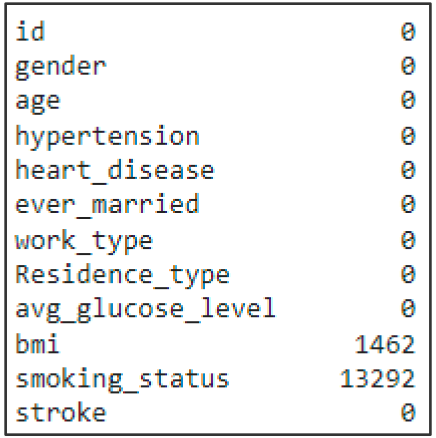
Dataset Attribute Information

Column name	Description
Id	unique identifier
Gender	"Male", "Female" or "Other"
Age	age of the patient in years
Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
Ever_married	"No" or "Yes"
Work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
Avg_glucose_level	average glucose level in blood
Bmi	body mass index
Smoking_status	"formerly smoked", "never smoked", "smokes"
Stroke	1 if the patient had a stroke and 0 if not

Data Cleansing

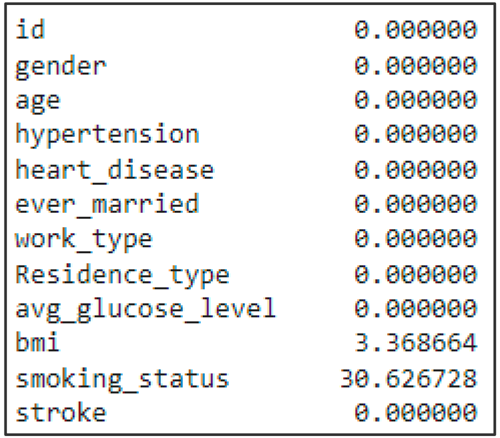
Missing Value Handling

Value



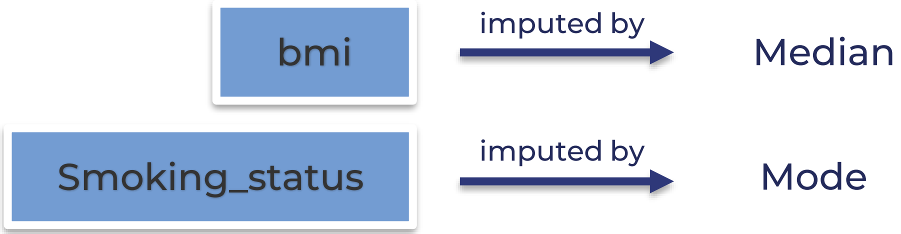
```
id                       0
gender                   0
age                      0
hypertension             0
heart_disease            0
ever_married             0
work_type                0
Residence_type           0
avg_glucose_level        0
bmi                   1462
smoking_status       13292
stroke                   0
```

Value (%)



```
id                    0.000000
gender                0.000000
age                   0.000000
hypertension          0.000000
heart_disease         0.000000
ever_married          0.000000
work_type             0.000000
Residence_type        0.000000
avg_glucose_level     0.000000
bmi                   3.368664
smoking_status       30.626728
stroke                0.000000
```

There is no column that has more than 35% NaN missing values. Therefore, none of the columns need to be dropped. In the other hand, the imputation is conducted.

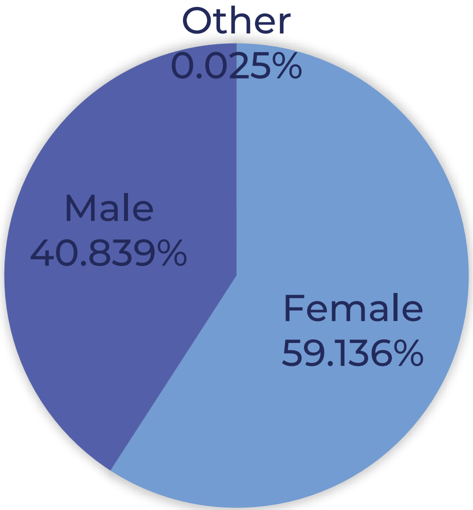


bmi imputed by → Median

Smoking_status imputed by → Mode

Data Cleansing

Missing Value Handling



‘Other’ in gender column is only 0.025% of the data. Therefore, it can be omitted and focus on ‘Female’ and ‘Male’

Duplicate Value Handling

0

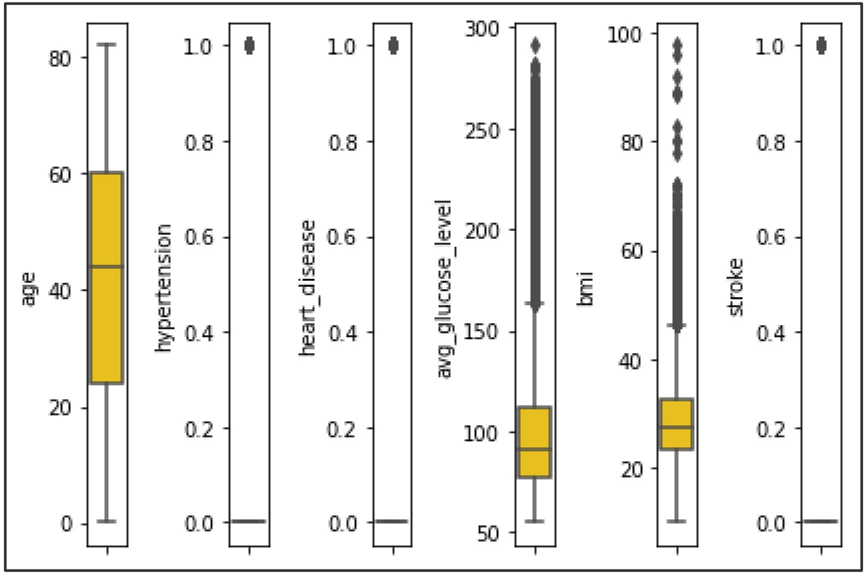
Data Prepocessing

Boxplot for numerical data

Before filtering,

43400

rows



After filtering,

42305

rows

Outliers in avg_glucose_level:

11.47%

More than 5%, using clip()

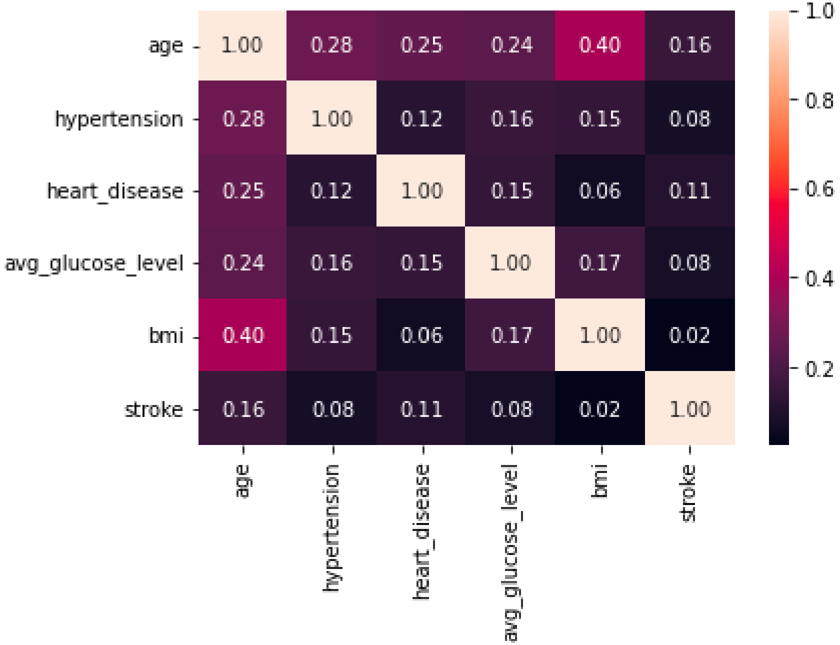
Outliers in bmi:

2.5%

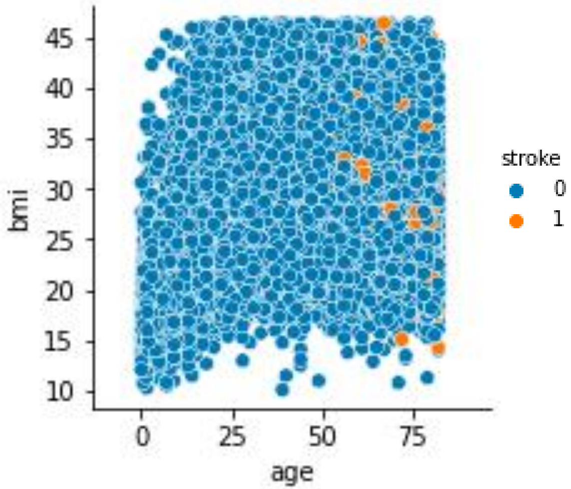
Less than 5%, drop

Exploratory Data Analysis Insight

Numerical data



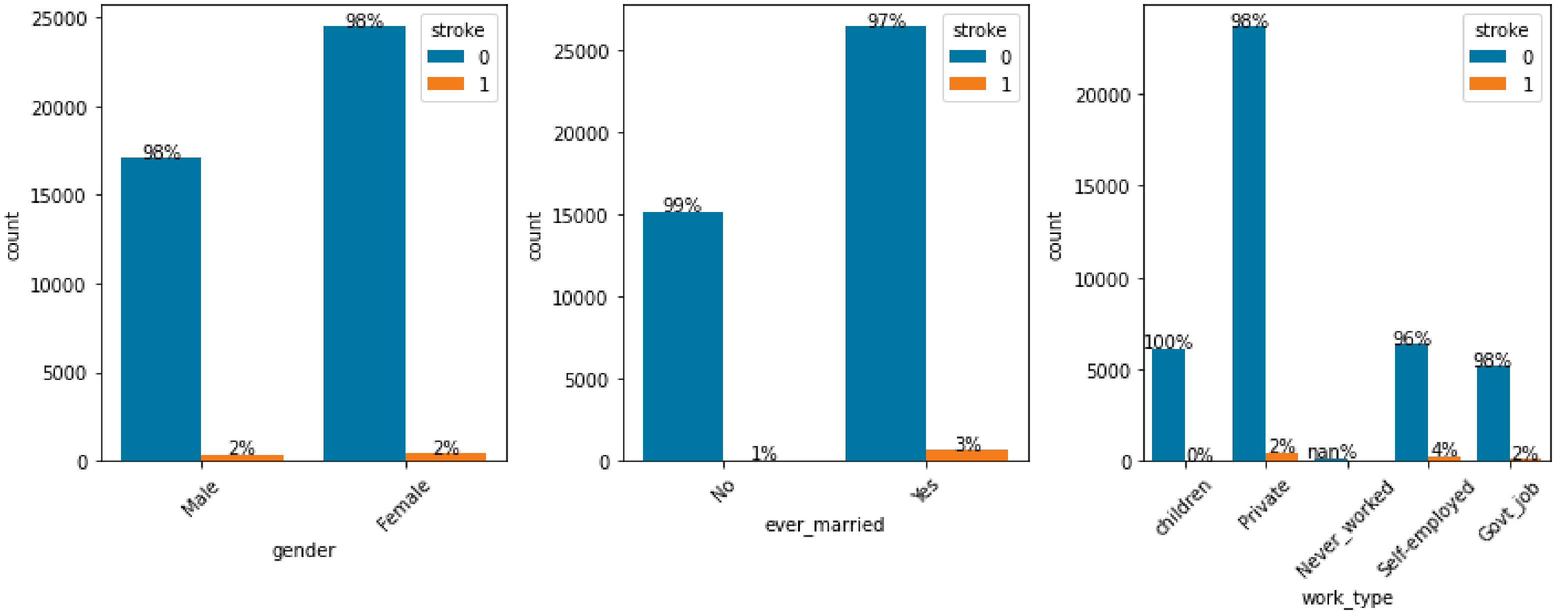
	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1.00	0.28	0.25	0.24	0.40	0.16
hypertension	0.28	1.00	0.12	0.16	0.15	0.08
heart_disease	0.25	0.12	1.00	0.15	0.06	0.11
avg_glucose_level	0.24	0.16	0.15	1.00	0.17	0.08
bmi	0.40	0.15	0.06	0.17	1.00	0.02
stroke	0.16	0.08	0.11	0.08	0.02	1.00



- There is no redundant feature(s)
- Age has highest correlation towards the stroke (16%)
- Age and bmi have highest correlation (40%)
  The higher the age, no matter what their BMI is, the higher their chances of having a stroke

Exploratory Data Analysis Insight

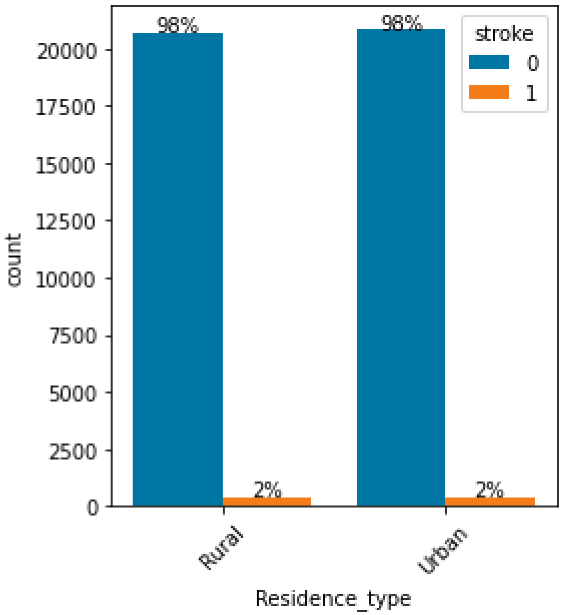
Categorical data

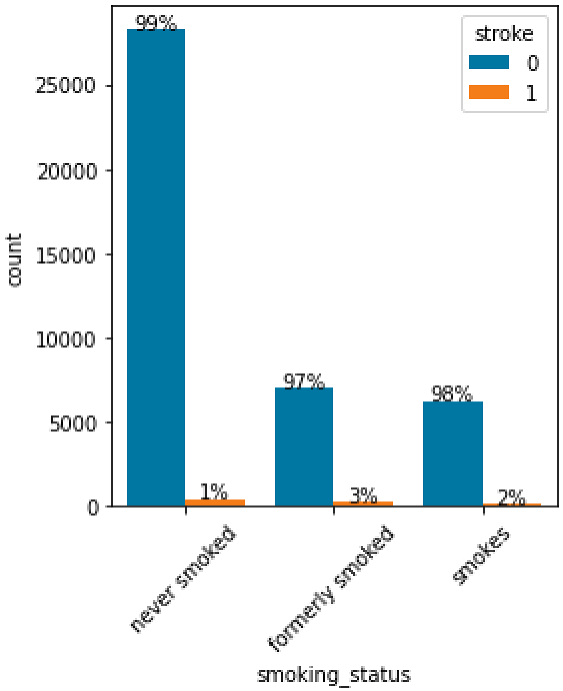


- Male has same opportunity as female to have stroke (depends on genetic)
- People who has married is more likely to have stroke (3%)
- People work as self-employed has highest possibility to have stroke (4%)

Exploratory Data Analysis Insight

Categorical data





- Rural and urban people has same opportunity to have stroke (depends on lifestyle)
- People who formerly smoked tend to have stroke (3%)

Feature Engineering

Label Encoding

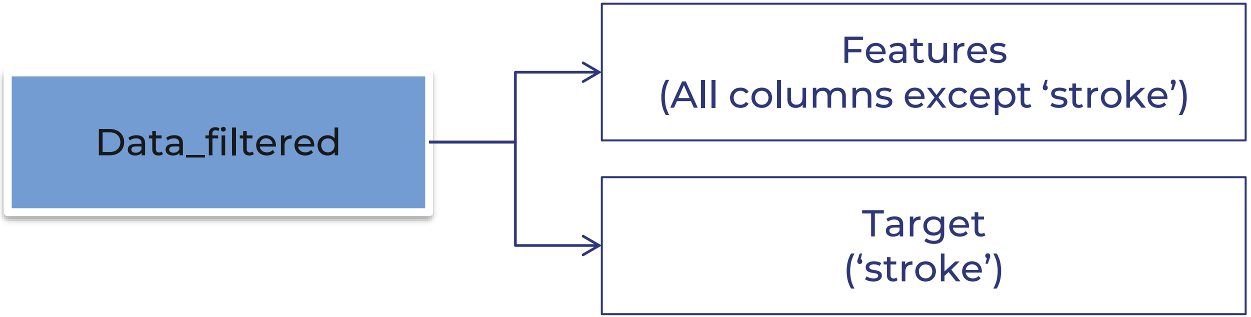
Features	0	1
Gender	Male	Female
Residence_type	Rural	Urban
Ever_married	No	Yes

Features	Features after One Hot Encoding				
Work_type	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children
smoking_status	smoking_status_formerly smoked	smoking_status_never smoked	smoking_status_smokes		

03

Modelling and Evaluation

Features and Target Splitting

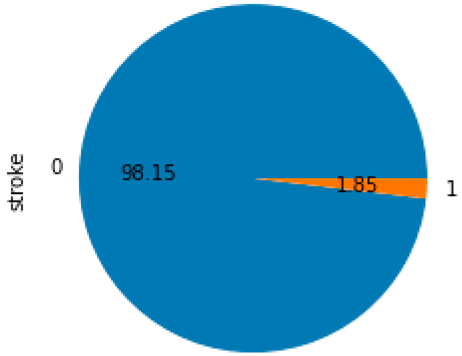


Data_filtered

Features
(All columns except ‘stroke’)

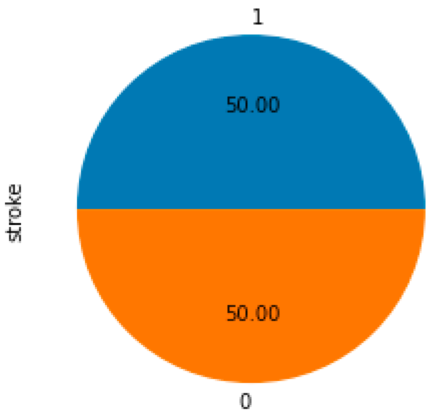
Target
(‘stroke’)

Imbalanced Data



```
0      33219
1        625
```

Resampling by Random Over Sampler and SMOTE Algorithm



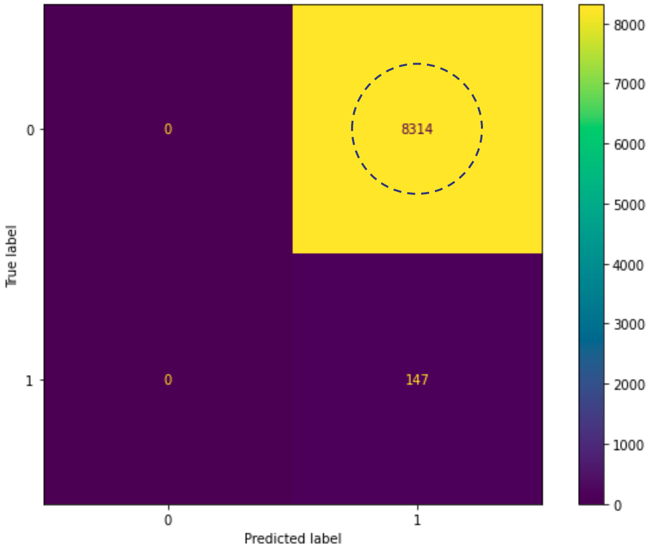
```
1      33219
0      33219
```

Logistic Regression and Decision tree are two algorithms which are better than others since they have high recall score

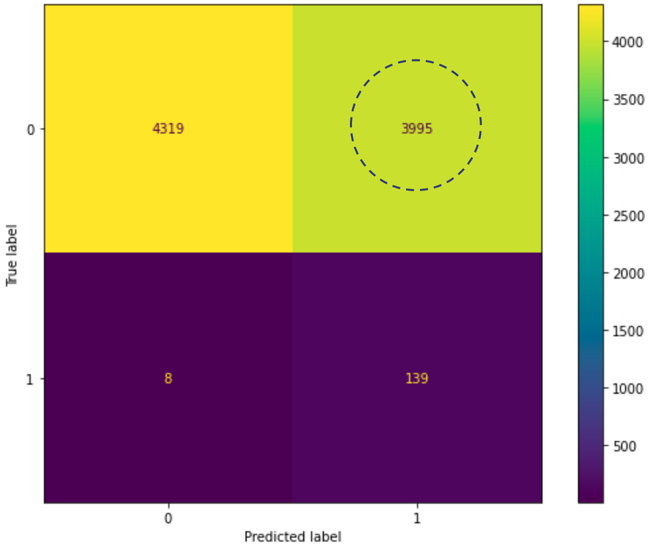
Algorithm	Time to Run (s)	Metrics		
		Recall	F1 score	Precision
Logistic Regression	55.5	1.00	0.03	0.02
Decision Tree	50.5	0.95	0.06	0.03
Random Forest	364	0.79	0.10	0.05
SVM (Support Vector Machine)	63	0.51	0.11	0.06
XGB (Extreme Gradient Boosting)	737	0.46	0.11	0.06

Then, what is the best algorithm?

Logistic Regression

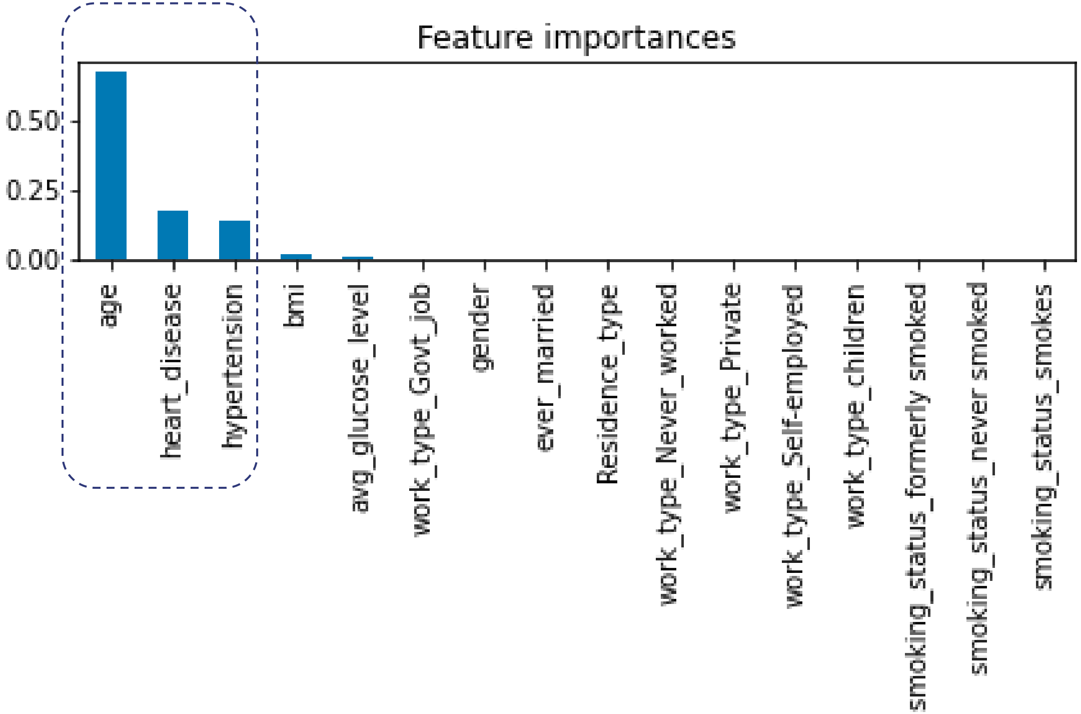


Decision Tree



False positive in confusion matric:
Model predicts people **having stroke**, but actually they **do not have stroke**

Decision Tree is a suitable algorithm for predicting the tendention of the stroke



Age, heart disease, and hypertension are the highest three factors that affect stroke

04

Conclusion and Recommendation

Conclusions

- All features in the dataset are used to analysing (no redundant features)
- Individual's age is the highest factor that affects stroke
- Gender, residence do not have much effect on having stroke, it depends on the genetics and lifestyle
- Best algorithm is decision tree classifier

Recommendations

Adding more features about genetics and lifestyle. Example:

- Stroke history from their parents
- Daily food
- Physical activity

Script: https://colab.research.google.com/drive/1m3H5_qEtrwwNMsArxcuSh8iKDlSMAQae?usp=sharing

Thank you

Gmail : camytha.octa2@gmail.com
LinkedIn : www.linkedin.com/in/camytha-octanuryati-rochmad/
Github : https://github.com/camythaocta