Rakamin Academy x id/x partners

Credit Risk Prediction Model

ID/X Partners - Data Scientist

Presented by
Irpan Maulana

Irpan Maulana

Hi! I am Irpan Maulana, a junior data scientist. i have a great interest in building predictive models that can be used for decision making.

Skill :

- Machine Learning (Python)
- SQL
- Pengolahan Data
- Data Analysis

✉ Irpanmaulana038.im@gmail.com

in https://linkedin.com/in/irpan-maulana-87a1332b5/

COURSES

Belajar Dasar Data Science(Dicoding) | https://www.dicoding.com/certificates/NVP74GY8GPR0

<09,2024>

Belajar Dasar Structured Query Language (SQL) | https://www.dicoding.com/certificates/EYX4JYL1WZDL

<10,2024>

Data Science Course Level Basic(ITBOX) | https://itbox.id/certificate-verifier/139407D25-1395F7679-1277DFD43/

<10,2024>

Data Science Course Level Intermediate | https://itbox.id/certificate-verifier/139407D25-1395F7995-1277DFD43/

<02,2025>

Data Scince Course Level Advanced | https://itbox.id/certificate-verifier/139407D25-1395F9FD8-1277DFD43/

<05,2025>

ABOUT COMPANY

ID/X Partners is a consulting firm specialising in information technology solutions. specialising in leveraging data analytics and decision making (DAD) solutions combined with risk management and integrated marketing disciplines to help clients optimise portfolio profitability and business processes

PROJECT PORTOFOLIO

This project is to develop a model to predict credit risk to improve the accuracy of assessing and managing credit risk, so that they can optimize their business decisions and reduce the potential losses of lending companies (Multifinance). In developing this model using the Logistic Regression and Random Forest algorithms using the loan dataset

Link Code :

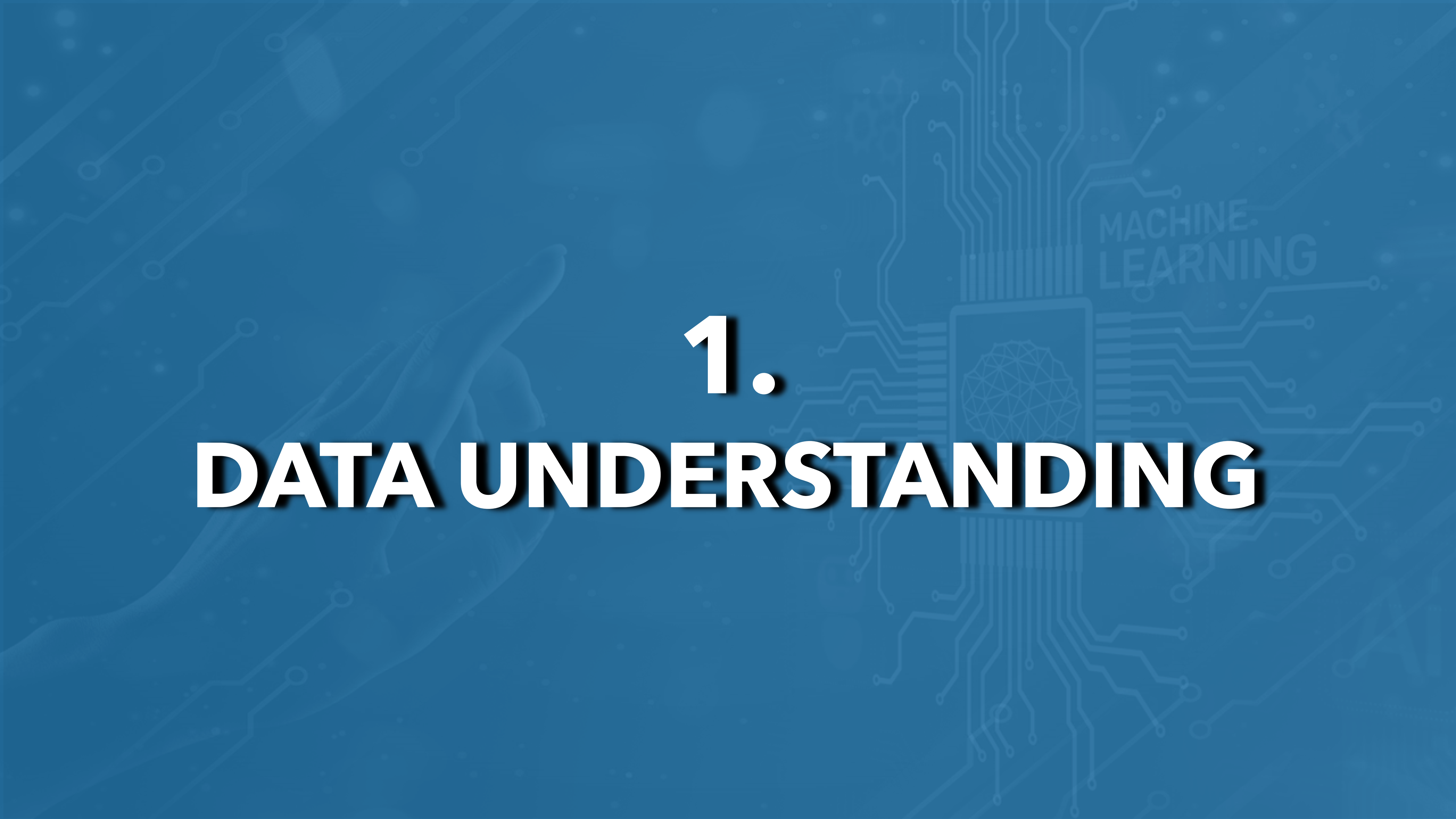
https://github.com/irpanmaulana038/Credit_Risk_Loan

Link code drive :

https://drive.google.com/drive/folders/1b7Inyoi4GA-mEvRzD4LrgvO1EbgHZeuw?usp=sharing

Link Vidio :

https://youtu.be/8V8-o5JeyMY

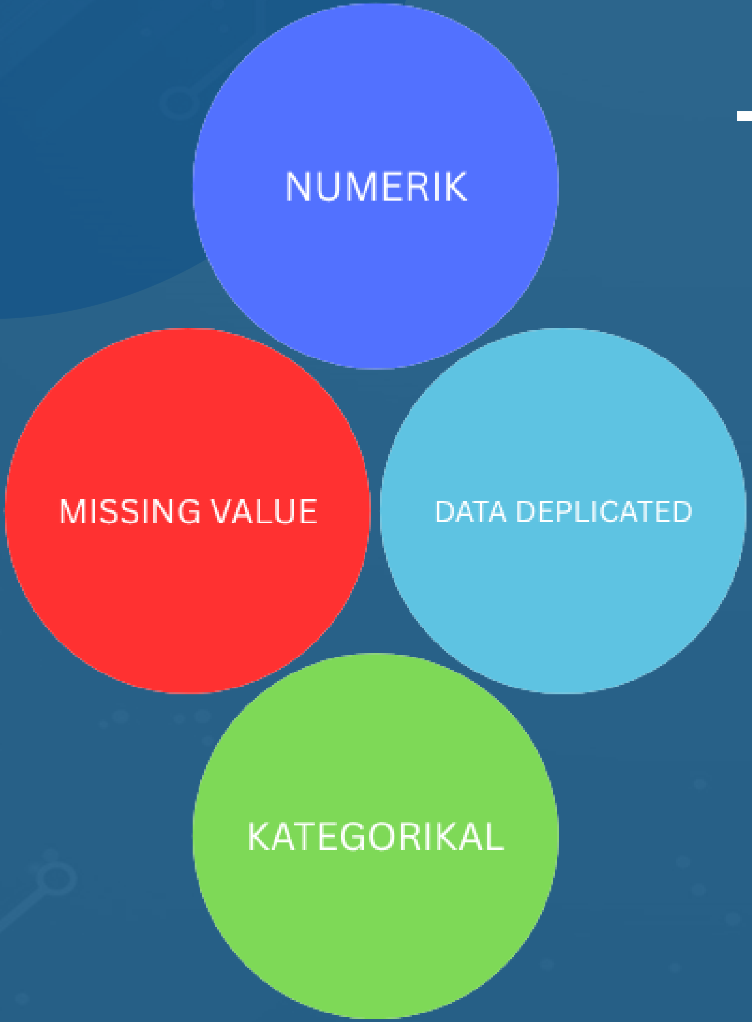


1.
DATA UNDERSTANDING

DATASET

This dataset contains information about borrowers, starting from financial profile, credit history and borrower status. This data is data from 2007 – 2014 and has 466285 rows and 75 columns

DATASET



NUMERIK

MISSING VALUE

DATA DEPLICATED

KATEGORIKAL

🟥 **Missing Value**

40 columns with Missing Value

■ **Data Duplikat**

No duplicated Data

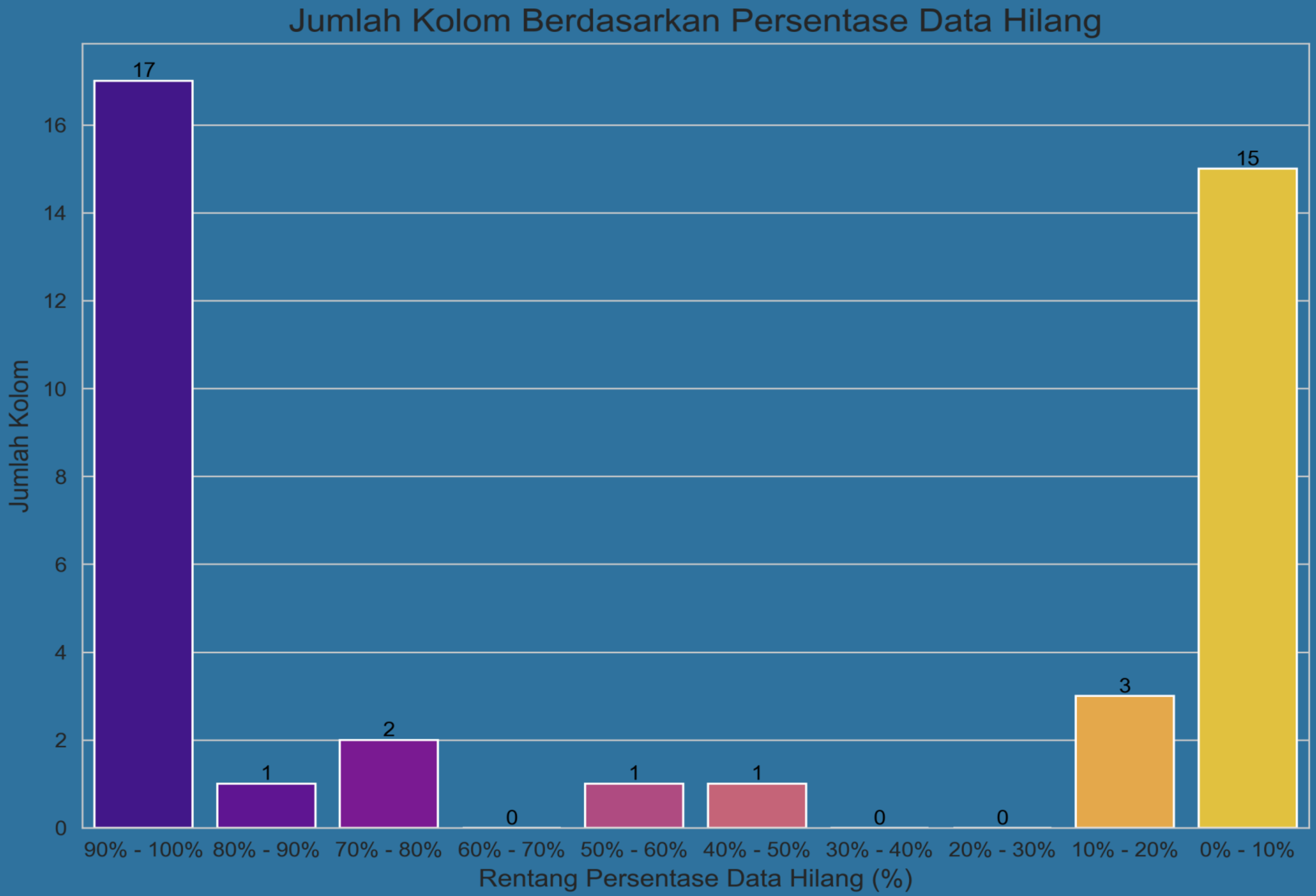
🟦 **Numerik**

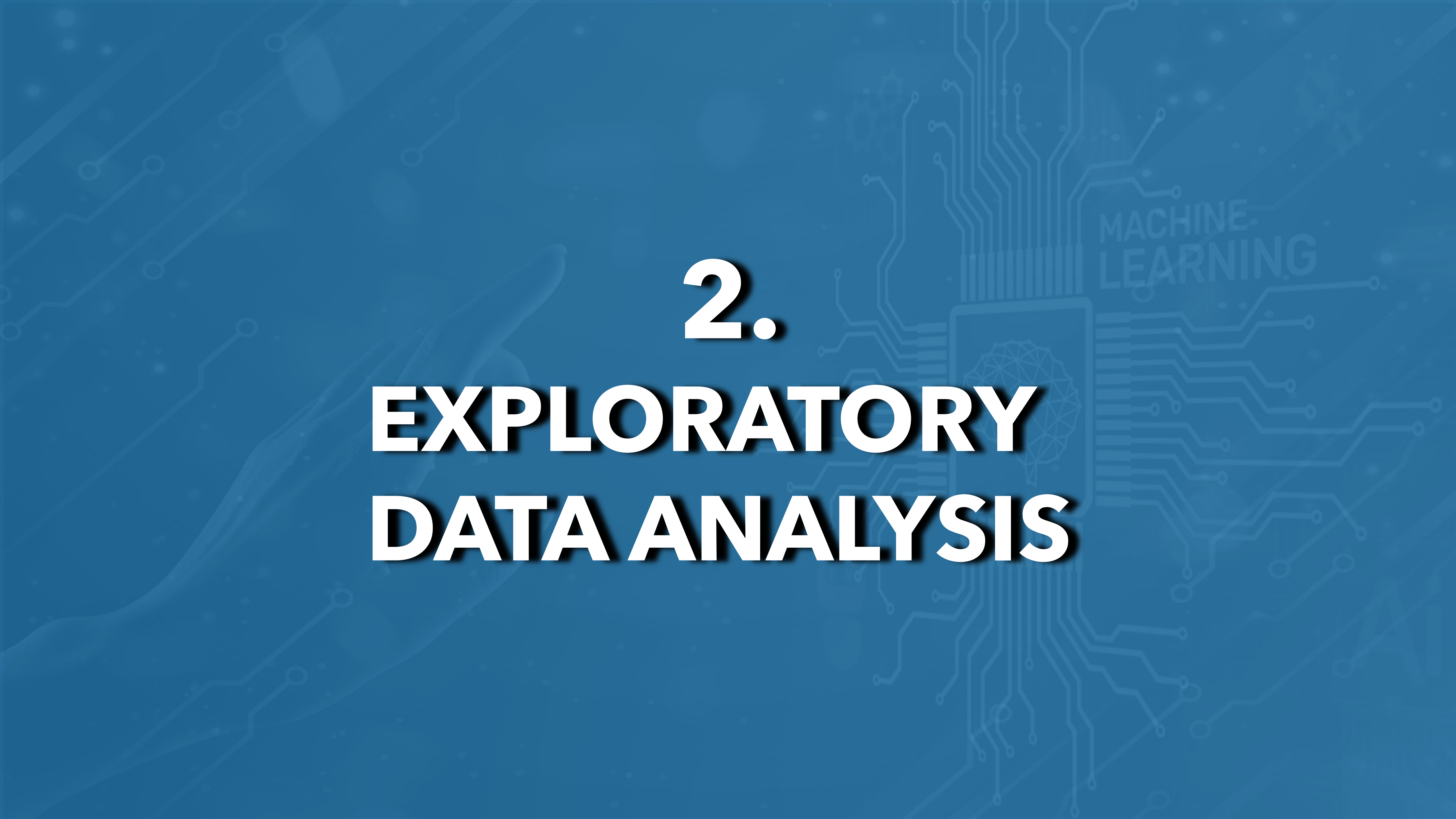
53 Numeric columns

🟩 **Kategorikal**

22 Categorical columns

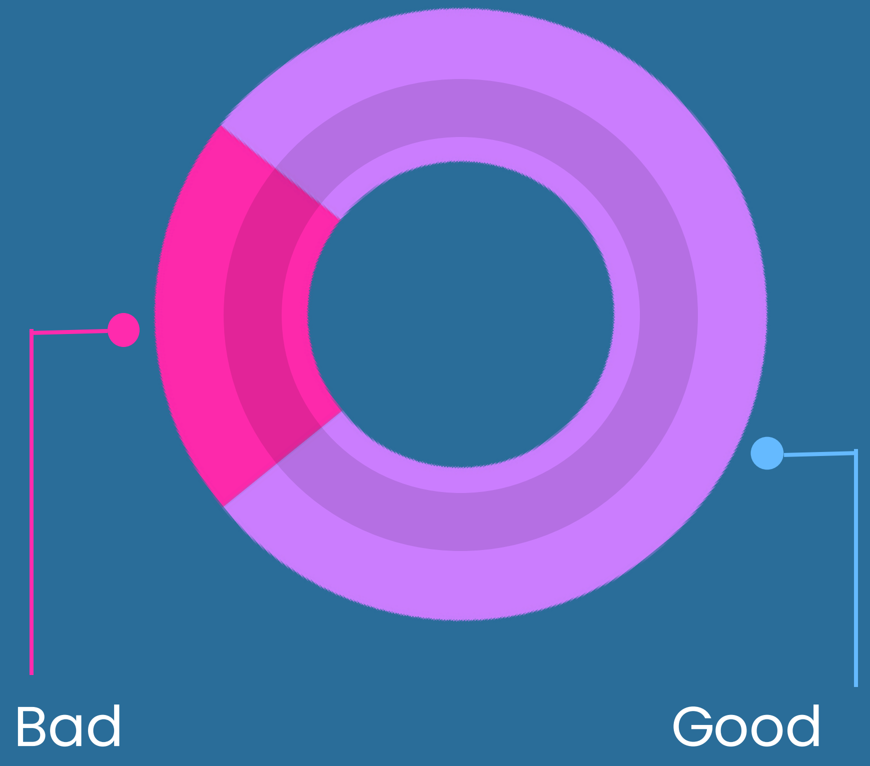
MISSING VALUE





2.
EXPLORATORY DATA ANALYSIS

Credit Risk

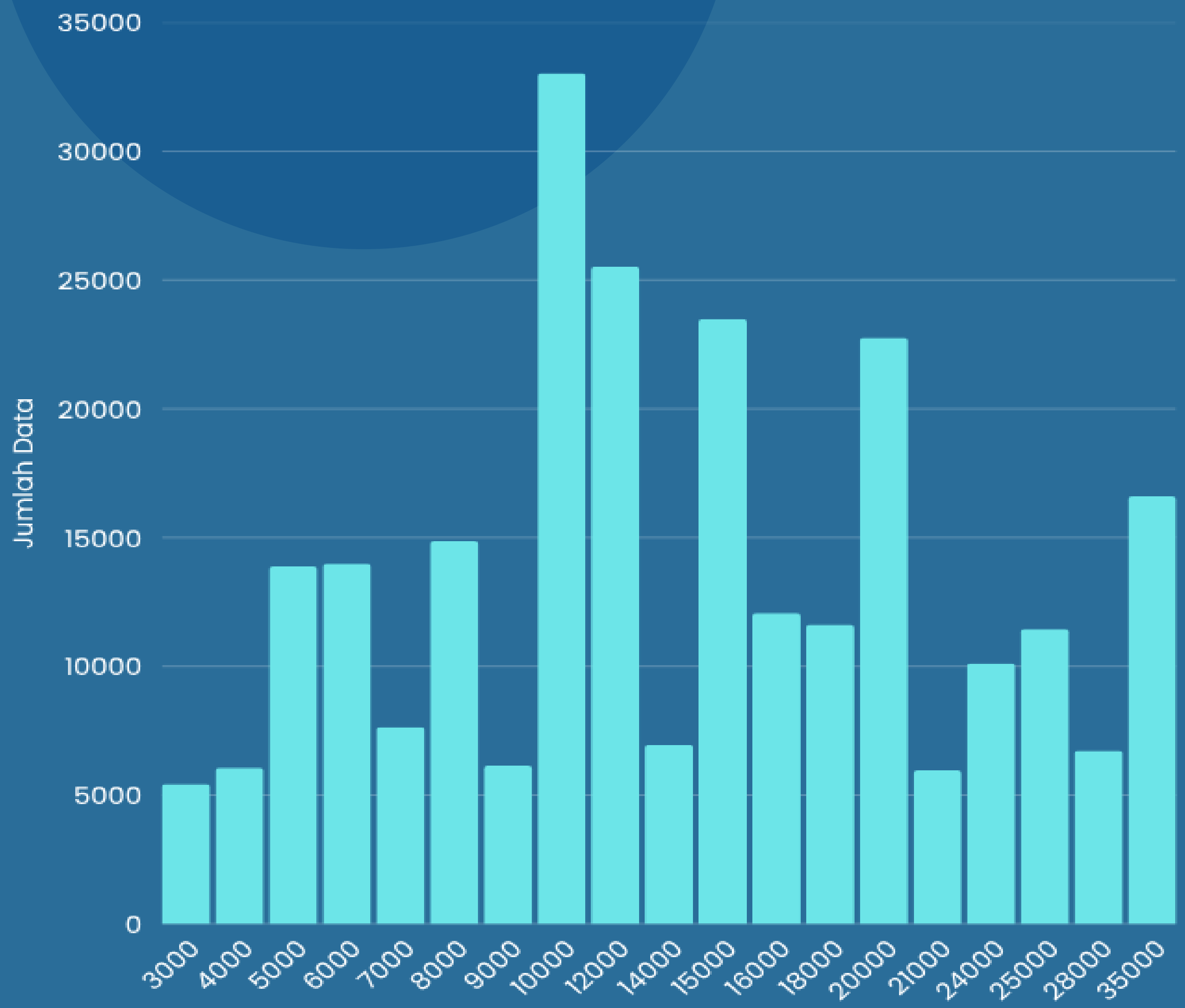


🟪 **Good 78,56%**

Good memiiki 186727 data

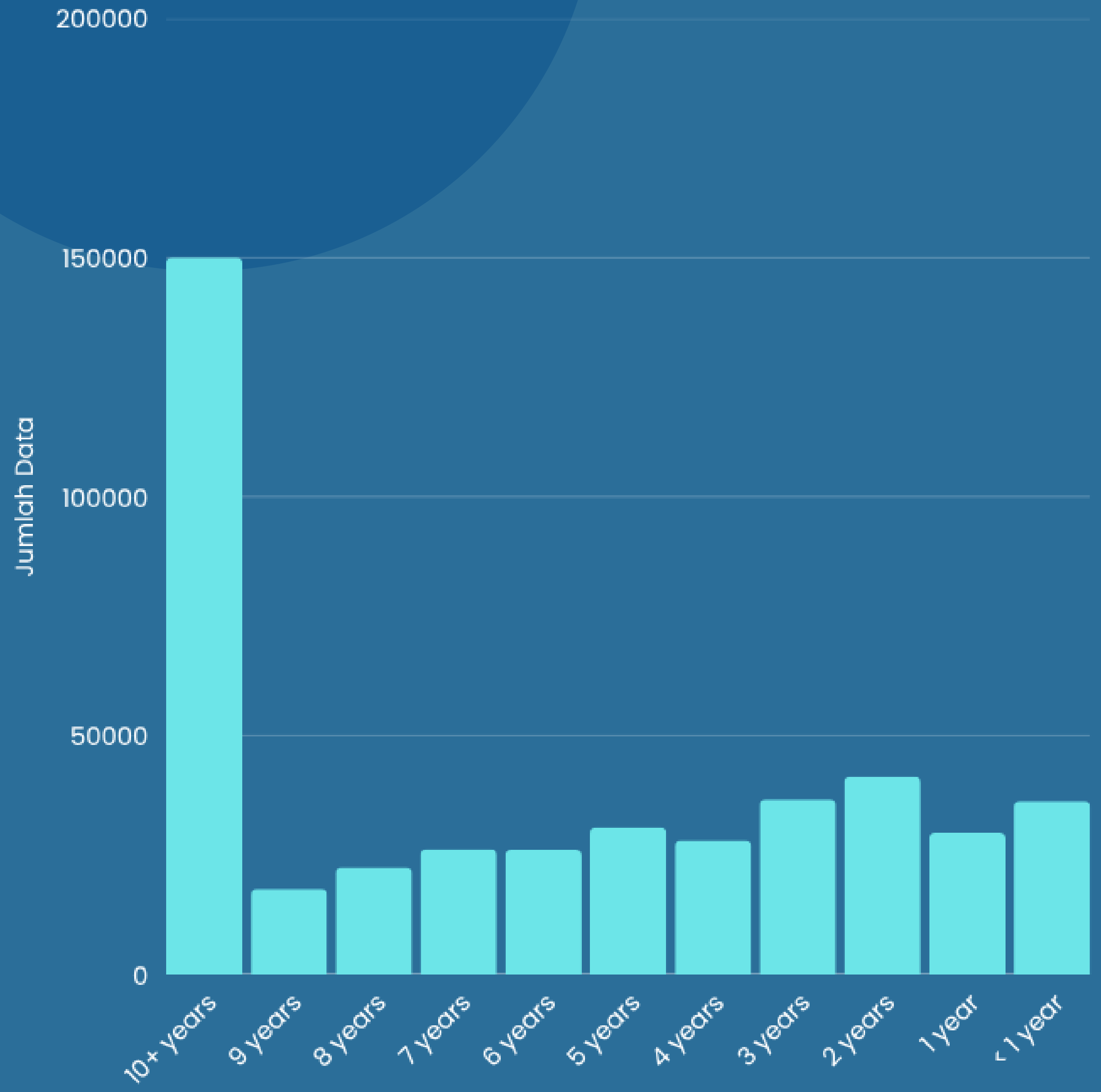
🟥 **Bad 21,44%**

Bad memiliki 50968



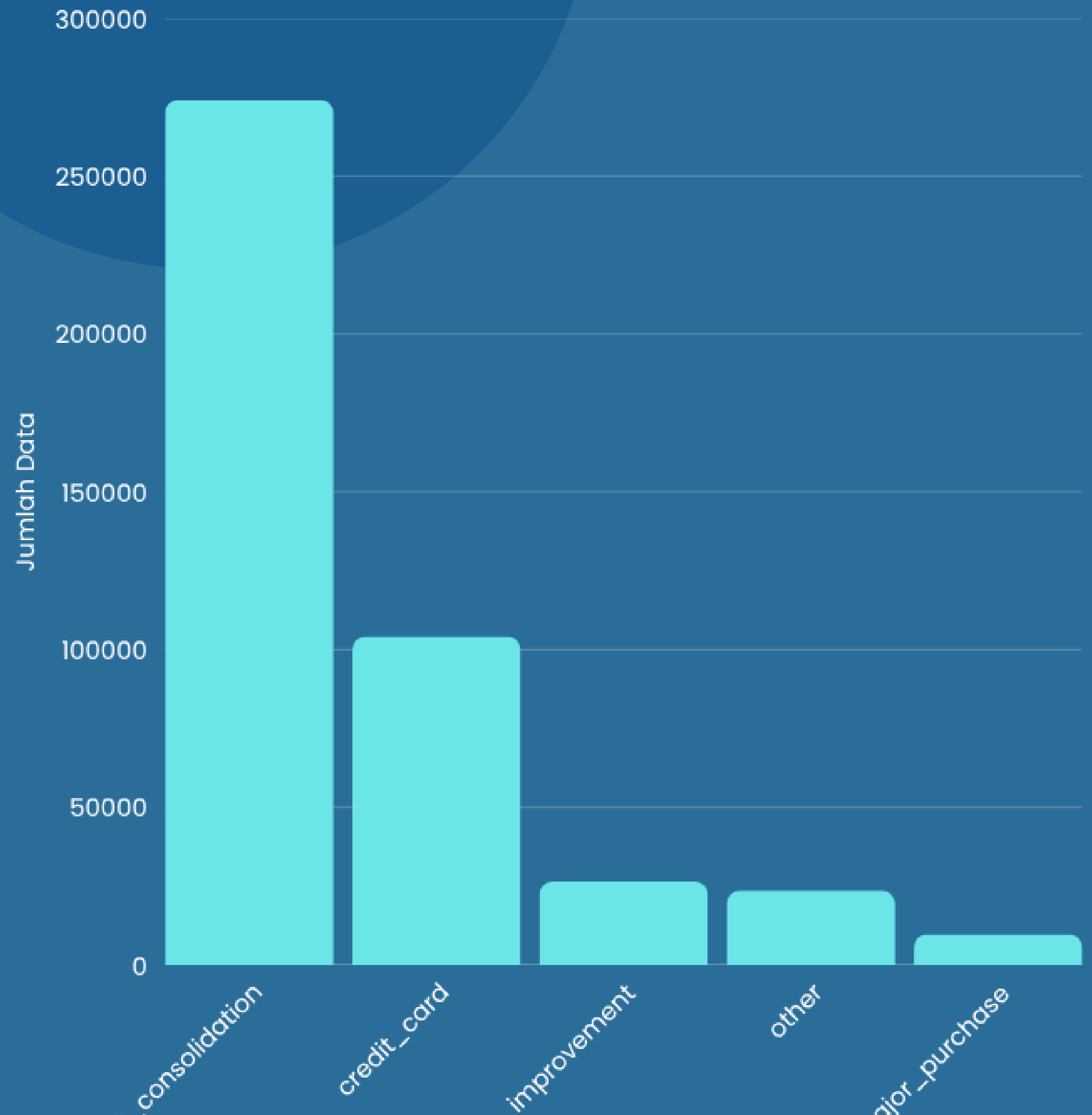
LOAN_AMNT

loan_amnt the amount of loan applied for by the borrower. Based on the graph, the majority of the amount applied for is 10000. This amount is below the average, where the average for loan amount is 14300



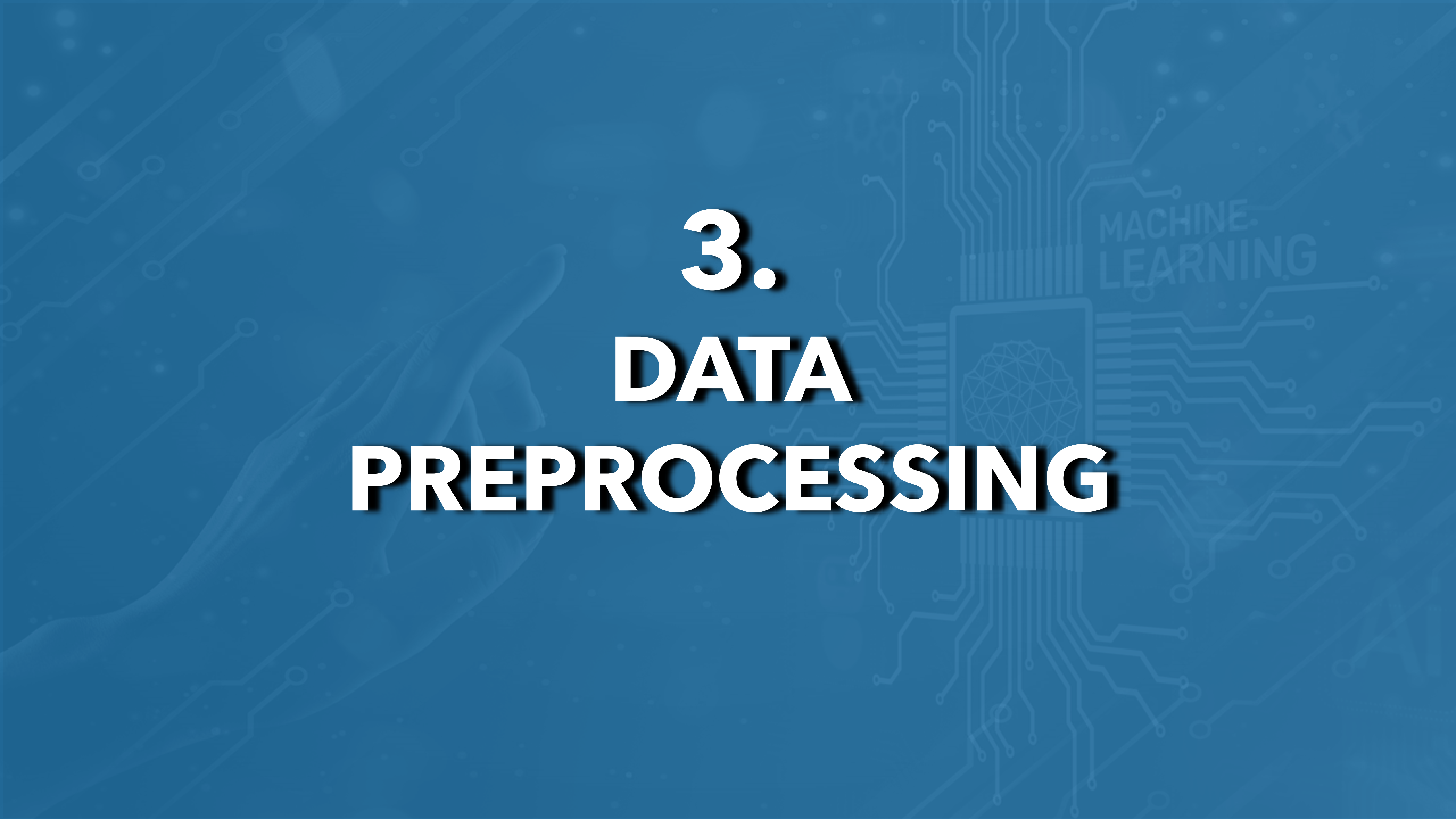
EMP_LENGTH

Based on the graph, the majority of borrowers have worked for more than 10 years.



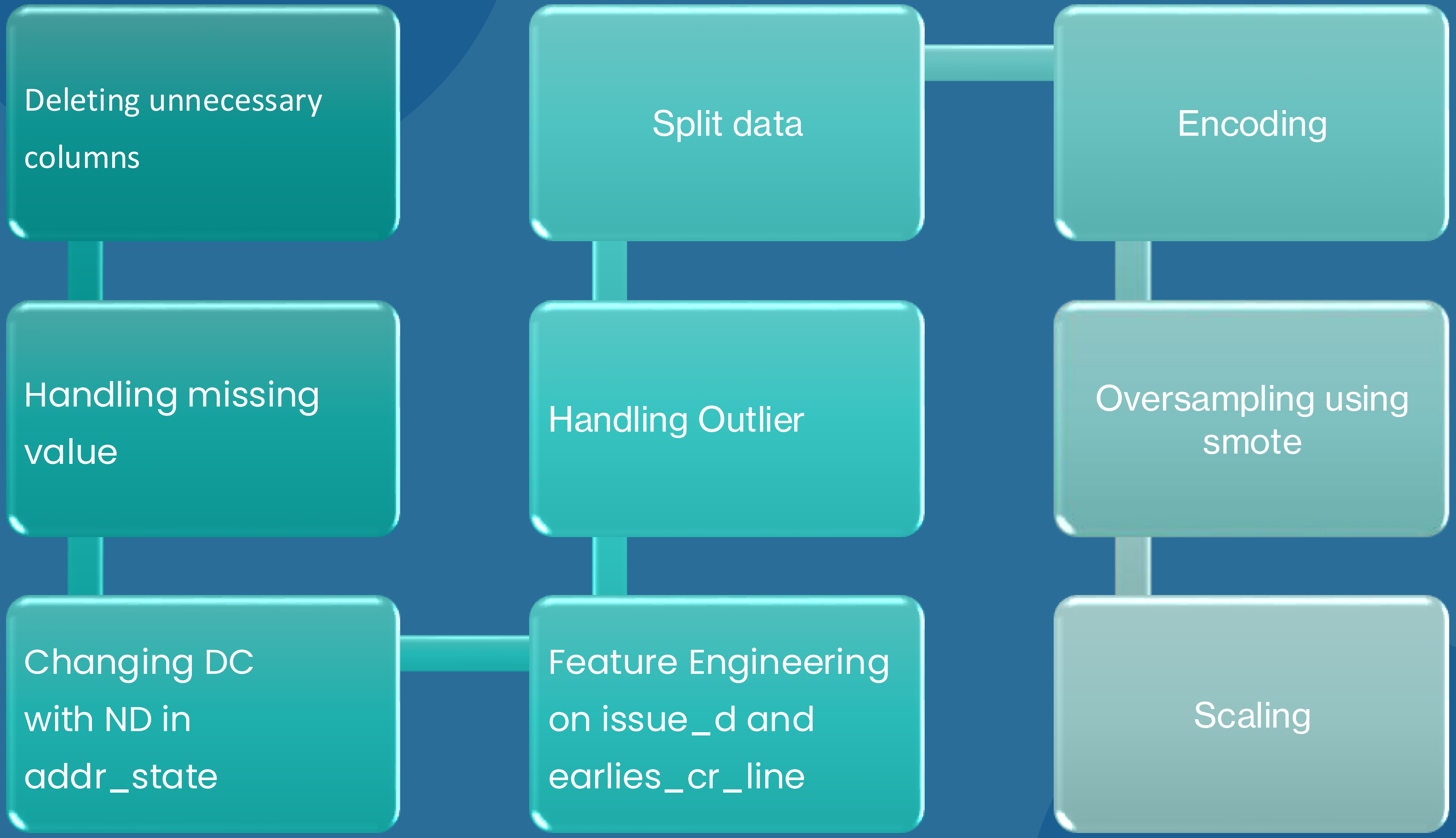
TOP 5 PURPOSE

Based on the graph, the majority of borrowers' goals are debt consolidation and credit cards.



3.
DATA PREPROCESSING

DATA PREPROCESSING



Deleting unnecessary columns

Handling missing value

Changing DC with ND in addr_state

Feature Engineering on issue_d and earlies_cr_line

Handling Outlier

Split data

Encoding

Oversampling using smote

Scaling

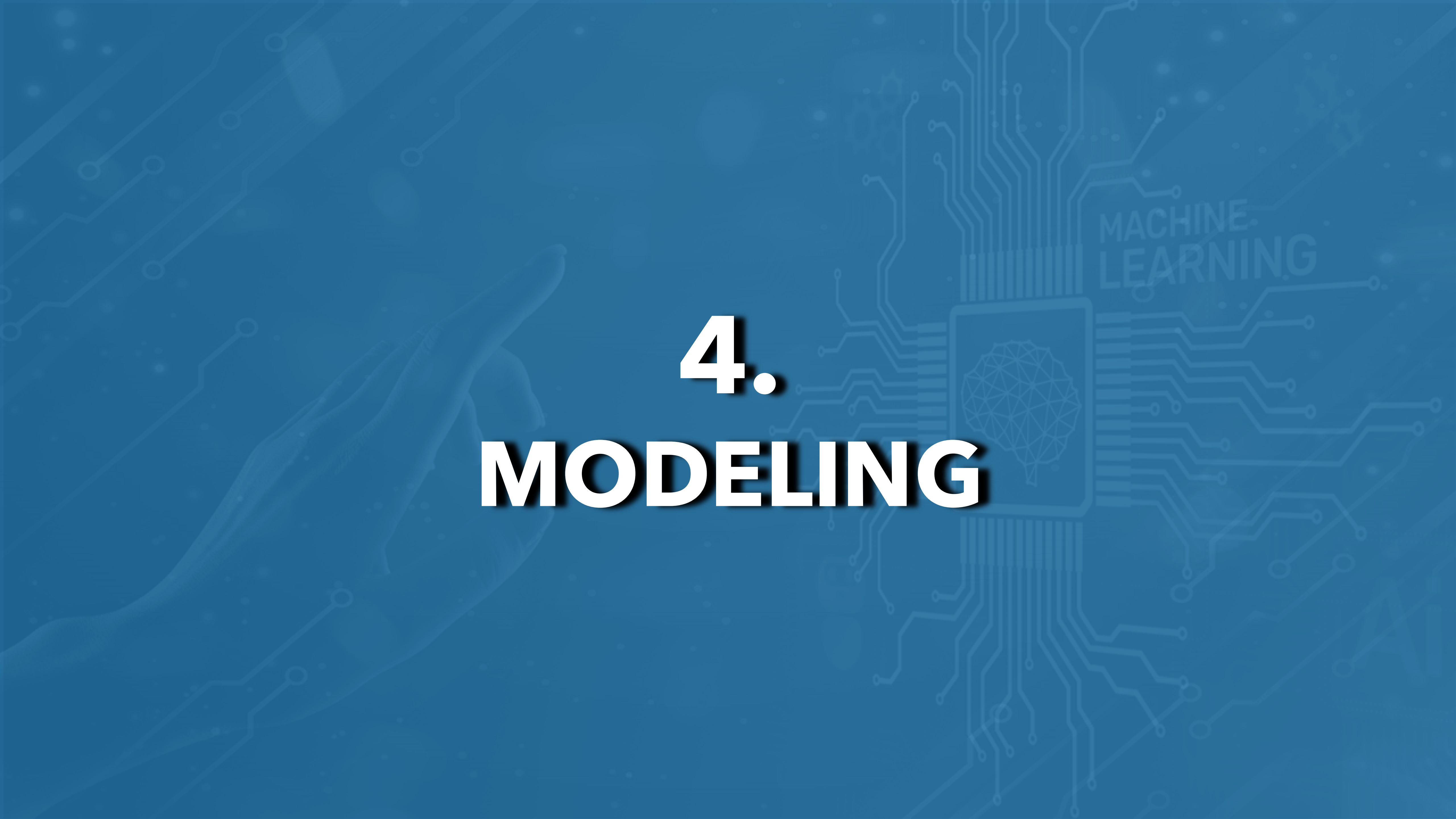
LABELLING

Good

- Fully Paid
- Does not meet the credit policy. Status:Fully Paid'

Bad

- Charged Off
- Default
- Does not meet the credit policy. Status:Charged Off
- Late (31-120 days)



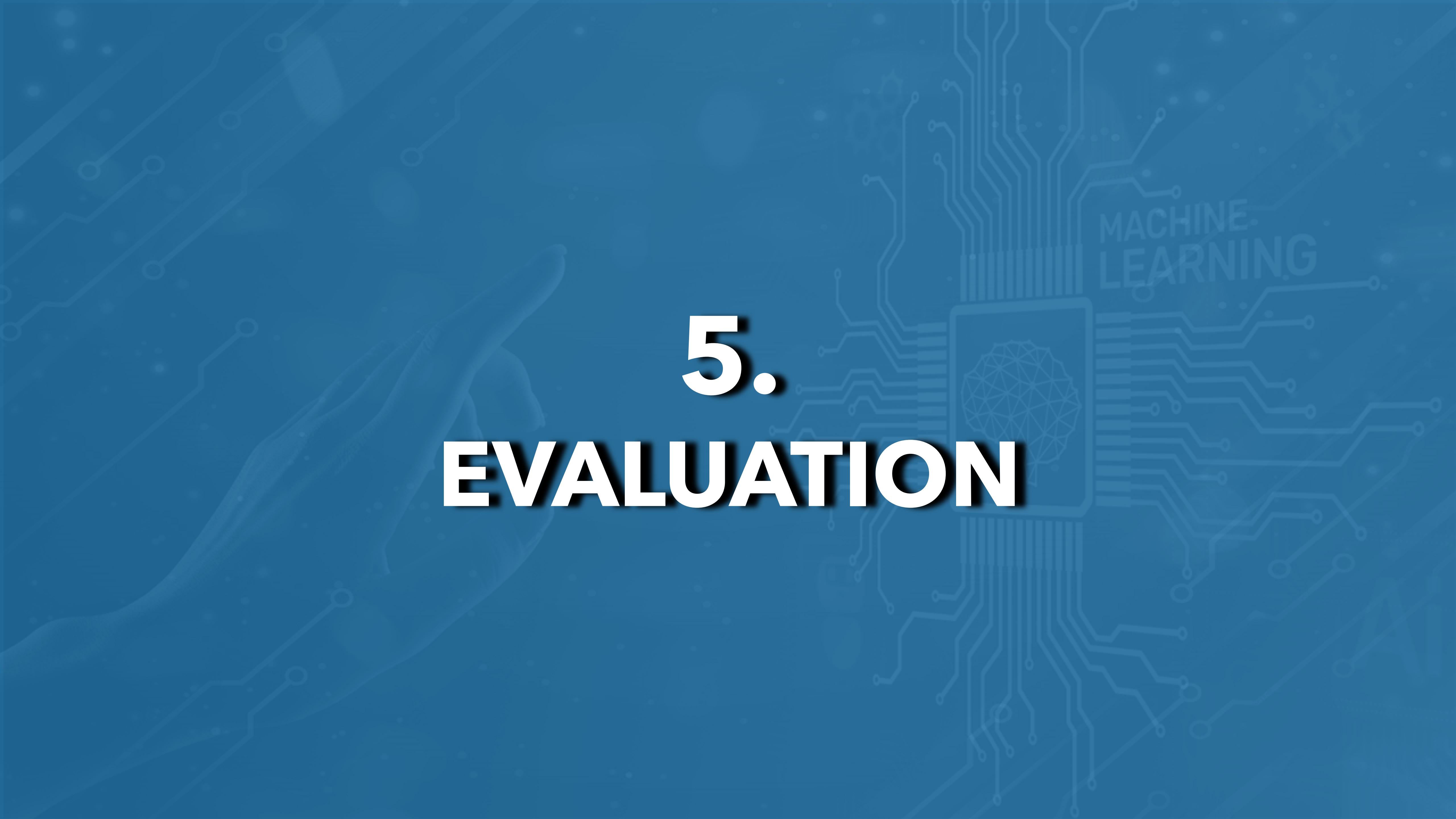
4. MODELING

MODELING

Logistic Regression		
Parameter	Hyperparameter	Metode
Class_weight = ‘Balanced	'C: [100,10, 1, 0.1, 0.01, 0.001], solver: [liblinear, 'saga']	GridSearch

Random Fores		
Parameter	Hyperparameter	Metode
Class_weight = ‘Balanced	'n_estimators': [300, 400], 'max_depth': [10, 20], 'min_samples_split': [5,7]	GridSearch

5.
EVALUATION



EVALUATION

Data Training (SMOTE)					
Model	Label Bad(1),Good(0)	PRECISION	RECALL	F1 SCORE	ACCURACY
Logistic Regression	**0**	0.66	0.66	0.66	0.66
	1	0.66	0.67	0.66	
Random Forest	**0**	0.89	1.00	0.94	0.94
	1	1.00	0.87	0.93	

EVALUATION

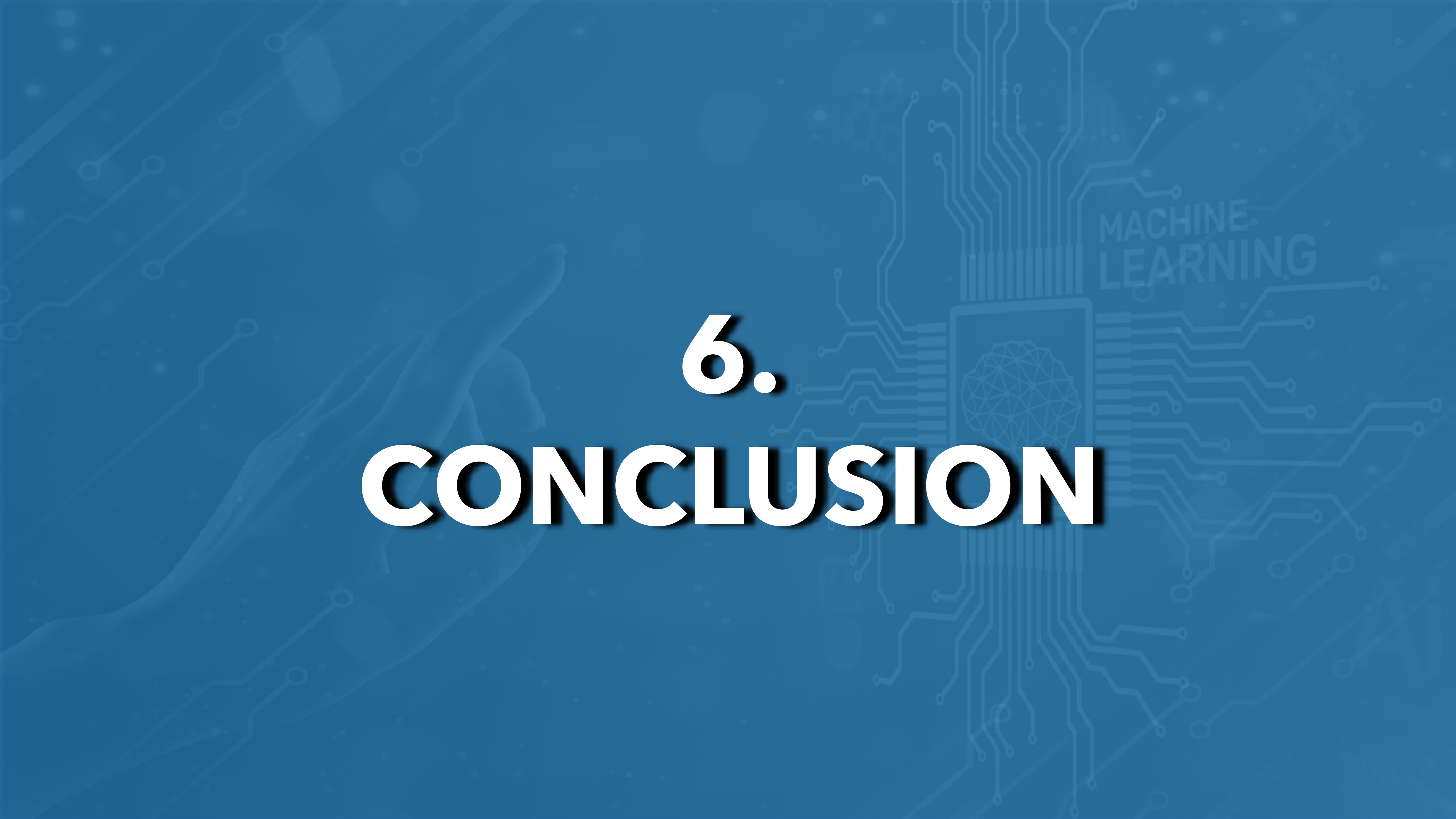
Data Testing (SMOTE)					
Model	Label Bad(1),Good(0)	PRECISION	RECALL	F1 SCORE	ACCURACY
Logistic Regression	**0**	0.86	0.66	0.75	0.65
	1	0.34	0.62	0.44	
Random Forest	**0**	0.80	0.98	0.88	0.78
	1	0.53	0.09	0.16	

EVALUATION

Data Train (TANPA SMOTE)					
Model	Label Bad(1),Good(0)	PRECISION	RECALL	F1 SCORE	ACCURACY
Logistic Regression	**0**	0.87	0.66	0.75	0.66
	1	0.34	0.65	0.45	
Random Forest	**0**	0.89	0.69	0.78	0.69
	1	0.37	0.68	0.48	

EVALUATION

Data Test (TANPA SMOTE)					
Model	Label Bad(1),Good(0)	PRECISION	RECALL	F1 SCORE	ACCURACY
Logistic Regression	**0**	0.87	0.67	0.75	0.66
	1	0.35	0.64	0.45	
Random Forest	**0**	0.87	0.68	0.76	0.67
	1	0.35	0.63	0.45	



6. CONCLUSION

CONCLUSION

This dataset has a lot of information and has a lot of large data and has a lot of missing values.

At the modeling, the oversampling technique with smote is not effective enough to overcome imbalance data because the model performance is only good on training data and its performance decreases when using test data. In modeling, the conclusion is that the random forest algorithm for modeling has better results than logistic regression in this case.

THANKS!

I WELCOME ANY CRITIQUE AND SUGGESTIONS FOR THIS PROJECT, I AM VERY OPEN TO RECEIVE THEM FOR THE IMPROVEMENT OF THIS PROJECT

Rakamin Academy X id/x partners