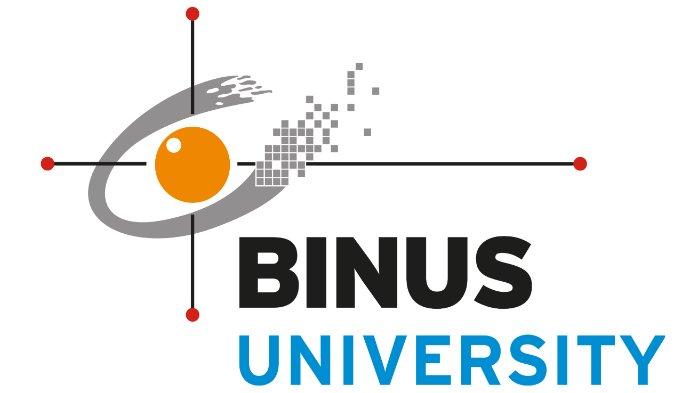
**Dokumentasi AOL Machine Learning**



**Anggota kelompok:**

**Agnes Calista - 2501980690**

**Olivia Intan Permata Dewi - 2540128650**

**Snow White Putri Safa - 2502020970**

**Jl. Kebon Jeruk Raya No. 27, Kebon Jeruk**

**Jakarta Barat 11530, Indonesia**

**2022**

# DAFTAR ISI

[**BAB I 3**](#_Toc125720422)

[**PENDAHULUAN 3**](#_Toc125720423)

[1.1 Latar Belakang 3](#_Toc125720424)

[1.2 Rumusan Masalah 4](#_Toc125720425)

[1.3 Tujuan Penelitian 4](#_Toc125720426)

[**BAB 2 5**](#_Toc125720427)

[**LANDASAN TEORI 5**](#_Toc125720428)

[2. 1 Algoritma Machine Learning 5](#_Toc125720429)

[2.1.1 Random Forest 5](#_Toc125720430)

[2.1.2 Decision Tree 5](#_Toc125720431)

[2.1.3 Logistic Regression 5](#_Toc125720432)

[2.1.4 SVM 5](#_Toc125720433)

[2.1.5 K-Nearest Neighbors 6](#_Toc125720434)

[**BAB 3 7**](#_Toc125720435)

[**METODE PENELITIAN 7**](#_Toc125720436)

[3.1 Exploratory Data Analysis 7](#_Toc125720437)

[3.2 Feature Engineering 7](#_Toc125720438)

[3.3 Train Machine Learning Models & Evaluate Machine Learning Models 7](#_Toc125720439)

[**BAB 4 8**](#_Toc125720440)

[**PEMBAHASAN 8**](#_Toc125720441)

[**BAB 5 13**](#_Toc125720442)

[**KESIMPULAN 13**](#_Toc125720443)

[**DAFTAR PUSTAKA 14**](#_Toc125720444)

# 

# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Di era ini, teknologi telah berkembang dengan pesat. Hampir semua masyarakat bergantung pada teknologi dalam kehidupan sehari-harinya. Teknologi memberikan kemudahan bagi penggunanya untuk mengakses informasi dan berkomunikasi. Banyak dampak positif yang bisa diberikan teknologi kepada penggunanya. Tetapi, selain dampak positif, teknologi juga memberikan dampak negatif, mudahnya pengguna untuk mengakses informasi membuat pengguna dapat dengan mudah menyebarkan informasi palsu (hoax). Pada mesin pencarian, dapat kita temui informasi mengenai syarat-syarat penting agar pengajuan pinjaman kepada bank dapat disetujui. Akan tetapi, tidak semua informasi bersifat akurat, karena informasi yang diberikan hanya sebatas berdasarkan pengalaman. Pada proyek ini, kami memberikan analisis untuk mengetahui variabel-variabel yang penting untuk persetujuan pinjaman berdasarkan data sehingga hasil analisis yang kita miliki dapat memberikan akurasi lebih dan bisa dipercaya.

Perusahaan ingin mengotomatiskan proses kelayakan pinjaman (real time) berdasarkan detail pelanggan yang diberikan saat mengisi formulir pada aplikasi online.

Dengan diadakannya pinjaman otomatis ini akan lebih mempermudah perusahaan dalam menganalisa tanpa membutuhkan waktu yang lama, selain itu untuk proses kemudahan dengan cepat. Agar proses ini dapat bersifat otomatis, mereka telah memberikan masalah ini untuk mengidentifikasi segmen pelanggan, yang berhak atas jumlah pinjaman sehingga mereka secara khusus dapat menargetkan pelanggan.

Pada proyek kali ini kami menggunakan dataset “Loan Prediction” yang memiliki atribut sebagai berikut:

* Loan\_ID = Loan ID
* Gender = Laki-laki atau perempuan
* Married = Status Pernikahan (Yes(Y) atau No(N))
* Dependants = Jumlah Tanggungan
* Education = Pendidikan (Graduate atau Under Graduate)
* Self\_Employed = Wiraswasta (Yes(Y) atau No(N))
* ApplicantIncome = Pendapatan
* CoapplicantIncome = Pendapatan Coapplicant
* LoanAmount = Jumlah Pinjaman Dalam Ribuan
* Loan\_Amount\_Term = Jangka waktu Pinjaman Dalam Sebulan
* Credit\_History = Riwayat Peminjaman
* Property\_Area = Daerah Tempat tinggal (Urban atau Semi Urban atau Rural)
* Loan\_Status = Peminjaman Disetujui(Yes(Y) atau No(N))

## 1.2 Rumusan Masalah

1. Proses peminjaman membutuhkan waktu yang lama untuk memprosesnya

2. Faktor apa saja yang mempengaruhi bank untuk memberikan pinjaman kepada nasabah?

3. Sulitnya untuk melakukan pinjaman ke bank

4. Metode yang paling cocok untuk klasifikasi “Loan-Prediction”

## 1.3 Tujuan Penelitian

1. Mengetahui faktor apa saja yang mempengaruhi bank untuk memberikan pinjaman uang kepada nasabah.
2. Memprediksi kelas atau kategori dari data baru berdasarkan karakteristik data yang ada.
3. Mengukur akurasi dengan beberapa algoritma, seperti Random Forest, Decision Tree, Logistic Regression, SVM, dan K-Nearest Neighbors.

# BAB 2

# LANDASAN TEORI

## 2. 1 Algoritma Machine Learning

Machine Learning merupakan mesin yang dikembangkan agar dapat belajar sendiri tanpa arahan. Algoritma machine learning didasari dengan ilmu matematika, statistika, dan lainnya. Selain itu Machine learning itu sendiri adalah salah satu cabang dari artificial intelligence.

Pada tugas ini kami menggunakan beberapa algoritma Machine Learning untuk mencari model terbaik dengan akurasi yang tepat. berikut beberapa algoritma yang kami pakai untuk modeling dari dataset Loan Prediction tersebut.

### 2.1.1 Random Forest

Random forest adalah metode klasifikasi yang berisi kumpulan dari pohon keputusan atau decision tree yang dijadikan ke dalam satu model. Random forest digunakan untuk kasus klasifikasi dengan dataset dalam jumlah besar.

### 2.1.2 Decision Tree

Decision tree adalah salah satu cara untuk memperoleh masa depan dengan cara membangun regresi ataupun klasifikasi modeling dalam struktur pohon.

### 2.1.3 Logistic Regression

Logistic Regression adalah teknik menganalisa menggunakan ilmu matematika agar dapat menemukan hubungan antara kedua faktor data. selain itu, untuk memprediksi nilai dari salah satu faktor tersebut berdasarkan faktor lainnya.

### 2.1.4 SVM

Algoritma support vector machine adalah metode untuk memisahkan dan mengelompokan data dengan melibatkan data training menggunakan klasifikasi. SVM bekerja untuk mencari hyperplane terbaik untuk memisahkan kedua buah kelas atau lebih pada ruang input. Hyperplane bisa berupa line pada dua dimensi dan berupa flat plane pada multiple plane.

### 2.1.5 K-Nearest Neighbors

Algoritma Nearest Neighbor merupakan algoritma klasifikasi paling sederhana, metode ini cukup mudah dipahami karena mengklasifikasikannya berdasarkan jarak terdekat dengan objek lain.

# BAB 3

# METODE PENELITIAN

## 3.1 Exploratory Data Analysis

* .shape() : mencari dimensi
* .info() : menampilkan informasi tentang dataset
* .describe() : menampilkan informasi deskriptif statistik
* .duplicated().sum() : menampilkan jumlah data yang terduplikasi
* .nunique() : menampilkan jumlah data yang bernilai unik.
* .drop() : menghapus column yang tidak terpakai
* .isna().sum() : menampilkan jumlah data yang memiliki nilai null.

## 3.2 Feature Engineering

1. *Imputation*

Mengganti missing value dengan nilai modus (nilai dengan frekuensi terbanyak)

1. *Label Encoding*

Mengubah kategorik menjadi numerik sehingga dapat dibaca oleh mesin.

1. *Find Value Count*

Melihat jumlah dari masing-masing variabel

1. *Creating Feature*

Membuat atribut baru yang berasal dari atribut yang sudah ada dengan metode agregasi sederhana (sum).

1. *Find Outliers*

Mengecek apakah terdapat outliers atau tidak dan apakah outliers tersebut harus dihilangkan atau tidak.

1. *Check Label*

Membuat klasifikasi menggunakan variabel target.

1. *Scaling*

Membuat data numerik pada dataset memiliki rentang nilai yang sama.

## 3.3 Train Machine Learning Models & Evaluate Machine Learning Models

1. Random Forest
2. Decision Tree
3. Logistic Regression
4. SVM
5. K-Nearest Neighbors

# BAB 4

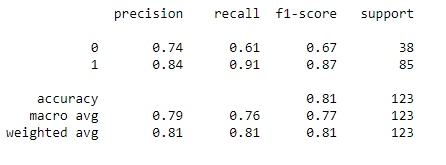
# PEMBAHASAN

Setelah kami melakukan proses *Exploratory Data Analysis* hingga proses training dataset, kami mendapatkan hasil sebagai berikut:

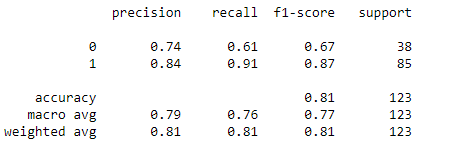
Karena semua model klasifikasi yang dibentuk memiliki data yang tidak seimbang, maka kita menggunakan nilai F1-score untuk mengukur akurasi. Kami memilih F1-score karena jika dibandingkan dengan Akurasi, F1-score dapat memberikan nilai yang lebih baik, karena tidak hanya menghitung jumlah error dari prediksi model yang kita bentuk, tetapi juga melihat tipe dari error yang ada(presisi ataupun sensitivitas).

1. Algoritma Random Forest

**(*with non-scaling attribute*)**



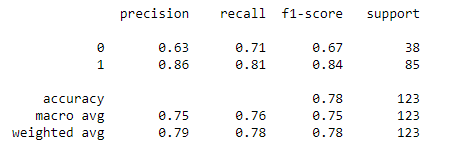
**(*with scaled attribute*)**



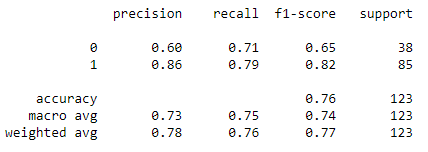
Pada algoritma Random Forest tidak memiliki perbedaan apapun sesudah ataupun sebelum menggunakan *feature engineering.*

1. Algoritma Decision Tree

**(*with non-scaling attribute*)**



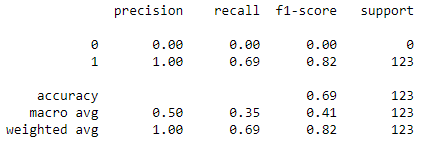
**(*with scaled attribute*)**



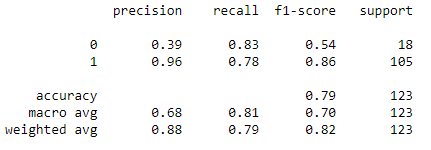
Pada algoritma Decision Tree, didapatkan nilai f1 sebelum di *scaling* sebesar 84%, tetapi setelah dilakukan *scaling* nilai f1 mengalami penurunan menjadi 82%.

1. Algoritma Logistic Regression

**(*with non-scaling attribute*)**



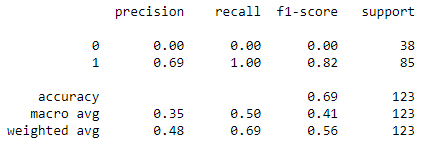
**(*with scaled attribute*)**



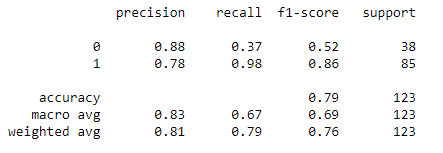
Pada algoritma Logistic Regression, didapatkan nilai f1 sebelum di *scaling* sebesar 82%, setelah dilakukan *scaling* nilai f1 mengalami kenaikan menjadi 86%.

1. Algoritma SVM

**(*with non-scaling attribute*)**



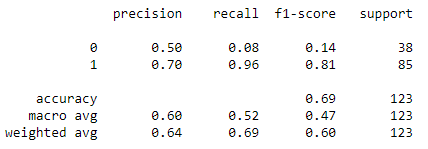
**(*with scaled attribute*)**



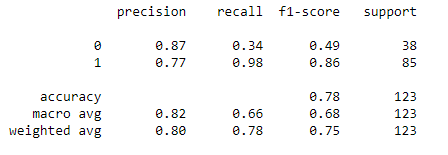
Pada algoritma SVM, didapatkan nilai f1 sebelum di *scaling* sebesar 82%, tetapi setelah dilakukan *scaling* nilai f1 mengalami kenaikan menjadi 86%.

1. Algoritma K-Nearest Neighbors

**(*with non-scaling attribute*)**



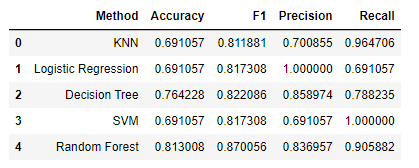
**(*with scaled attribute*)**



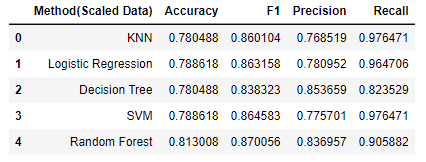
Pada algoritma K-Nearest Neighbors, didapatkan nilai f1 sebelum di *scaling* sebesar 81%, setelah dilakukan *scaling* nilai f1 mengalami kenaikan menjadi 86%.

Tabel Perbandingan

**(*with non-scaling attribute*)**

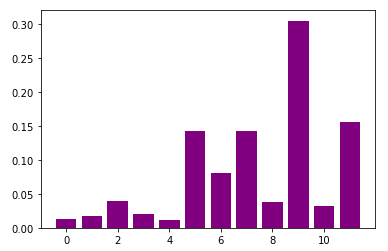


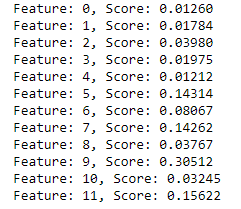
**(*with scaled attribute*)**

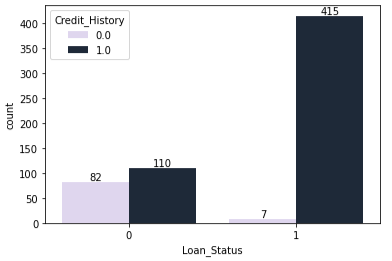


Model yang menggunakan scaling memiliki nilai metriks yang lebih tinggi, karena scaling membuat semua atribut berada dalam jarak yang sama, sehingga memberikan model peluang yang lebih besar untuk menemukan pola yang tepat. Tetapi tidak semua model akan memberikan nilai metriks yang lebih tinggi setelah dilakukan scaling. Dalam data diatas, didapatkan nilai F1 dari random forest tidak berubah dan nilai F1 dari model decision tree menurun setelah scaling, hal ini bisa disebabkan karena model yang dibuat, dimana random forest dan decision tree adalah model yang tidak membutuhkan *feature scaling*.

*Feature Importance*







Dengan menggunakan algoritma Random Forest kami mencari *Feature importance*, jika dilihat dari hasil diatas, dapat disimpulkan bahwa atributke-9 memiliki nilai yang paling besar yaitu sekitar 30% dari total semua atribut. Sehingga dapat disimpulkan bahwa Loan Status memiliki hubungan dengan atribut Credit History yang lebih besar dibanding dengan atribut lainnya.

Dari tabel perbandingan diatas, kita dapat menyimpulkan bahwa model algoritma klasifikasi terbaik yaitu random forest dengan nilai F1 sebesar 87,3%. Sehingga, dari semua model yang telah dibuat, kami mendapatkan random forest sebagai model *classifier* terbaik dalam dataset Loan Prediction”, serta fitur yang paling mempengaruhi Loan Status adalah Credit history, dengan korelasi sekitar 30%.

# BAB 5

# KESIMPULAN

Berdasarkan dari data yang sudah kami buat modelnya, dapat disimpulkan bahwa yang mempengaruhi bank untuk memberikan pinjaman uang kepada nasabah dipengaruhi oleh faktor riwayat kredit untuk status peminjaman. Riwayat kredit ini memang berperan penting dalam mengambil keputusan karena jika seorang peminjam tidak bayar tepat waktu, bagaimana perusahaan percaya bahwa ia akan membayar pinjaman tepat waktu. Kami telah membuat model dengan banyak algoritma seperti Random Forest, Decision Tree, Logistic Regression, SVM, dan K-Nearest Neighbors. Dari hasil yang kami dapatkan, Decision Tree memiliki nilai F1 paling rendah yaitu 83%, sedangkan Random Forest memiliki F1 yang paling tinggi yaitu 87%. Selisih kedua algoritma tersebut adalah 4%. Random Forest memiliki nilai F1 tertinggi dikarenakan menghindari overfitting pada sebuah dataset pada saat mencapai F1 yang maksimal dan hal ini telah kami buktikan pada model yang sudah kami buat. Sehingga Random forest merupakan model *classifier* terbaik dalam dataset Loan Prediction.

# DAFTAR PUSTAKA

Farokhah(2020). *Implementasi k-nearest neighbor untuk Klasifikasi Bunga*

*Dengan Ekstraksi Fitur Warna RGB,* 1129-1131. Diakses 14 Januari 2023, dari

<https://jtiik.ub.ac.id/index.php/jtiik/article/view/2608>

Fadilah, Laili(2018). *Klasifikasi Random Forest Pada Data Imbalanced*, 26-27. Diakses 20

Januari 2023, dari

<https://repository.uinjkt.ac.id/dspace/bitstream/123456789/55034/1/LAILI%20FADILAH-FST.pdf>

Hidayat, Anwar(2015). *Regresi Logistik*. Diakses 20 Januari 2023, dari

<https://www.statistikian.com/2015/02/regresi-logistik.html>

Kukuh, R.(2018). *Melakukan Feature Scaling Pada Dataset*. Diakses 20 Januari 2023, dari

<https://medium.com/machine-learning-id/melakukan-feature-scaling-pada-dataset-229531bb08de>

Putra, Edi Purnomo dan Kendrew Huang(2022). *Support Vector Machine Algorithm*. Diakses 21 Januari 2023, dari

<https://sis.binus.ac.id/2022/02/14/support-vector-machine-algorithm/>