Evaluasi Perbandingan Algoritma Decision Tree, Random Forest, Naïve Bayes, KNN, dan SVM Untuk Penyelesaian Kasus Klasifikasi Ketertarikan Pelanggan Terhadap Kampanye Pemasaran Pada Bank X

Oleh :
Jericho Cristofel Siahaya
Universitas Multimedia Nusantara
Tangerang, Banten, Indonesia
jericho.cristofel@student.umn.ac.id

Abstract

Supervised classification has various modeling methods that can be implemented according to the specifications of the training data. These various methods have its own advantages and disadvantages, and these methods have been successfully implemented to solve various kinds of problem in various domains and fields, including economics issues. In this study, we compare several methods of supervised classification algorithms to solve classification problems on bank X marketing campaign data. We use some of the supervised classificatio algorithms such as decision tree, random forest, nave bayes, k-nearest neighbors, and support vector machine. The objective of this study is to analyze the evaluation between the algorithms and find the best accuracy performance which can be use to solve classification problem in marketing campaign data of bank X. There is no specific benchmark references to determine the perfect accuracy of each algorithm, but we use cross validation to find best estimated accuracy.

Keywords: classification, decision tree, random forest, naïve bayes, k-nearest neighbors, support vector machine, bank marketing

Abstrak

Supervised classification memiliki berbagai metode *modeling* yang dapat diimplementasikan sesuai dengan spesifikasi data yang akan di-*train*. Berbagai metode ini memiliki keunggulan serta kekurangannya tersendiri, dan metode-metode ini telah berhasil diimplementasikan untuk memecahkan berbagai macam kasus masalah di berbagai domain dan bidang, salah satunya adalah bidang ekonomi. Pada penelitian ini, peneliti mencoba untuk membandingkan beberapa metode ataupun algoritma *supervised classification* guna menyelesaikan permasalahan klasifikasi pada data kampanye pemasaran bank X. Beberapa algoritma yang akan dipakai adalah *decision tree, random forest, naïve bayes, k-nearest neighbors,* dan *support vector machine*. Tujuan dari penelitian untuk adalah untuk melihat evaluasi antara algoritma yang dipakai dan mencari performa dari algoritma terbaik untuk menyelesaikan kasus klasifikasi pada data kampanye pemasaran bank X. Tidak ada referensi tolak ukur khusus untuk menentukan performa tiap algoritma, namun peneliti menggunakan *cross validation* guna mencari estimasi akurasi terbaik dari setiap algoritma.

Keywords: *classification, decision tree, random forest, naïve bayes, k-nearest neighbors, support vector machine, bank marketing*

1. PENDAHULUAN

Dewasa ini, kampanye pemasaran (*marketing campaign*) telah menjadi salah satu bagian atau aktivitas yang melekat pada industri bank yang bertujuan untuk mempromosikan produk-produk finansial terbarunya kepada pelanggan mereka. Walaupun telah banyak riset serta penelitian yang membahas mengenai uji signifikansi keberhasilan kampanye pemasaran menggunakan berbagai macam metode uji statistik, namun hingga saat ini implementasi serta pemahaman terhadap *insight* dari teknik-teknik statistik tersebut masih menjadi tantangan yang cukup sulit untuk diselesaikan terutama oleh pelaku industri bank atau finansial.

Seiring dengan perkembangan zaman serta kemajuan teknologi yang semakin canggih, memori serta kemampuan komputasi komputer pun semakin meningkat secara signifikan, hal ini pun yang membuat berbagai macam perangkat lunak ataupun *tools* dapat dikembangkan guna membantu perusahaan melakukan analisa serta tindakan untuk kemajuan bisnisnya. Analisa terhadap data telah berkembang hingga ke bidang *machine learning*. *Machine learning* merupakan metode interdisipliner sains yang mana dengan bantuan konsep kecerdasan buatan, bertujuan untuk menciptakan sistem terotomatisasi yang dapat mempelajari informasi baru secara mandiri dan terus melakukan perbaikan guna menghasilkan *knowledge* baru. Metode *machine learning* telah banyak diimplementasikan di berbagai macam industri, seperti ritel, kesehatan, manufaktur, hingga pendidikan. Bergantung pada data yang diekstrak serta kasus yang ingin diselesaikan, *machine learning* dapat dibagi ke dalam 3 jenis pembelajaran (*learning*): *supervised*, *unsupervised*, serta *semi-supervised*.

Pada penelitian ini, kasus yang ingin diselesaikan adalah terkait *supervised learning* yaitu untuk membuat sebuah model klasifikasi. Peneliti akan menggunakan lima algoritma klasifikasi guna membandingkan dan mencari algoritma terbaik dengan melihat hasil akurasi dari tiap klasifikasi. Model klasifikasi yang dibangun dapat digunakan untuk memprediksi keputusan pelanggan apakah tertarik terhadap kampanye pemasaran oleh Bank X dengan cara melakukan deposit pada program yang ditawarkan. Jurnal ilmiah ini ditulis dengan struktur sebagai berikut: Tinjauan Teoritis (Bab 2), Metode Penelitian (Bab 3), *Algorithms* (Bab 3.1), *Prior Knowledge* (Bab 3.2), *Preparation* (Bab 3.2), Hasil dan Pembahasan (Bab 4), *Data Exploration* (Bab 4.1), *Data Preprocessing* (Bab 4.2), *Modeling* (Bab 4.3), *Modeling Improvement* (Bab 4.4), *Evaluation* (Bab 4.5), Penutup (Bab 5), Daftar Pustaka.

2. TINJAUAN TEORITIS

Pada penelitian oleh Grzonka (2016), implementasi dari algoritma *decision tree* yang dipilih untuk membuat model klasifikasi menghasilkan akurasi sebesar .84 sedangkan pada penelitian yang sama juga digunakan algoritma *random forest* yang menghasilkan akurasi sebesar .89. Selain itu, pada penelitian oleh Wisaeng (2013) menggunakan algoritma *support vector machine* mendapatkan hasil akurasi sebesar .87.

3. METODE PENELITIAN

Metode yang dipakai pada penelitian ini mengikuti siklus CRISP-DM dimulai dari *business understanding* dan *data understanding* yang akan dijelaskan lebih detail pada bagian *prior knowledge*, kemudian ada *data preparation* terkait eksplorasi serta *preprocessing* yang akan dibahas lebih lanjut pada Bab 4, pada bab yang sama juga akan dibahas mengenai *modeling* serta *evaluation*. *Deployment* atau implementasi model pada bisnis perusahaan akan dibahas pada bab terakhir (Bab 5) dari jurnal ini.

Algoritma-algoritma yang digunakan dan diuji coba pada penelitian mencangkup *decision tree, random forest, naïve bayes, knn,* dan *svm*. Algoritma-algoritma tersebut dipilih berdasarkan kepopuleran serta penggunaannya yang cukup mudah dengan bantuan beberapa *library open-source* yang telah tersedia. Definisi serta cara kerja singkat mengenai algoritma-algoritma tersebut akan lebih dijelaskan secara detail pada bagian *algorithms*.

Pada penelitian ini, peneliti mencoba menggunakan dua kali tahapan *modeling* dengan tujuan untuk melakukan pembuktian terhadap metode *modeling* tanpa *feature selection*, normalisasi, dan *cross-validation* serta menggunakan *feature selection*, normalisasi, serta *cross-validation*. Tidak terdapat batasan ataupun referensi (patokan) akurasi yang akan dipakai pada tahapan evaluasi, sehingga akurasi yang didapat akan dipilih hanya berdasarkan urutan terbesarnya.

1) *Algorithms*

Machine learning tidak lepas dari penerapan algoritma-algoritma yang pada dasarnya merupakan susunan formula logis dan sistematis yang digunakan untuk memecahkan suatu permasalahan atau menghasilkan sebuah informasi baru (Hill, 2016). Pada penelitian ini, peneliti menggunakan lima algoritma yang dipilih berdasarkan tingkat kepopuleran di kalangan praktisi data sains serta kemudahan penggunaannya yang dibantu dengan berbagai macam *library open-source* yang banyak tersedia di internet. Cara kerja dari tiap algoritma dijelaskan pada poin-poin di bawah ini:

- Decision Tree

  Decison Tree adalah sebuah diagram alir yang mirip dengan struktur pohon, dimana setiap internal node menotasikan atribut yang diuji, setiap cabangnya mepresentasikan hasil dari atribut tes tersebut dan leaf node mepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas. Istilah *decision tree* adalah proses menemukan kumpulan pola atau fungsi-fungsi yang mendeskripsikan dan memisahkan kelas data satu dengan lainnya, untuk dapat digunakan untuk memprediksi data yang belum memiliki kelas data tertentu (Han, 2006). Ada dua jenis algoritma *decision tree* yang terkenal, yaitu C4.5 dan *random forest.* Algoritma C4.5 merupakan algoritma yang dikembangkan dari algoritma ID3. C4.5 ini merupakan algoritma turunan dari algoritma ID3 dengan beragam peningkatan. Beberapa peningkatan ini diantaranya adalah, penanganan atribut-atribut numerik, missing value dan noise pada dataset, dan aturan-aturan yang dihasilkan dari model pohon yang terbentuk (Larasati & Sutrisno, 2018).
- Random Forest

  Random Forest (Classifier) merupakan salah satu model algoritma yang bersifat supervied learning dan berdasarkan pada ensembled learning. Ensembled Learning merupakan tipe yang bersifat mengkombinasikan algoritma-algoritma yang berbeda atau algoritma yang sama secara berkali-kali untuk membentuk model prediksi yang lebih kuat (Bahrawi, 2017).

  Algoritma Random Forest Classifier biasanya digunakan untuk menyelesaikan permasalahan-permasalahan di bidang klasifikasi maupun regresi, algoritma ini bersifat enslembed learning karena merupakan kombinasi dari kumpulan algoritma Decision Tree.

Dalam ensembled learning sendiri, terdapat 2 metode yaitu bagging dan boosting, yang dimana *random forest* ini menggunakan metode bagging, yaitu metode untuk melakukan training pada model secara paralel.

- Naïve Bayes

  Naive Bayes adalah salah satu algoritma *supervised learning* yang digunakan untuk melakukan klasifikasi secara statistikal dengan mengacu pada konsep probabilitas bersyarat pada Teorema Bayes (Parveen & Pandey, 2016; Tripathy et al., 2016). Metode yang dikembangkan dengan konsep dari Thomas Bayes ini memiliki asumsi bahwa terdapat independensi yang kuat antar *features* (Dey et al., 2016; Saritas & Yasar, 2019). Adapun persamaan umum dari Teorema Bayes yang digunakan dalam algoritma Naive Bayes adalah sebagai berikut:

  $$P(c \mid x) = P(x \mid c) \times P(c) / P(x)$$

  Dengan $P(c \mid x) = P(x1 \mid c)\ P(x2 \mid c) \ldots P(c)$. Dimana P(c | x) adalah posterior probability dari class (target), P(c) adalah prior probability dari class, P(x | c) adalah likelihood yakni probabilitas dari prediktor, dan P(x) adalah prior probability dari prediktor (Vembandasamy et al., 2015). Secara sederhana, Teorema Bayes juga dapat ditulis dengan rumus sebagai berikut:

  $$Posterior = Prior \times Likelihood / Evidence$$

  Pada Naive Bayes, nilai *evidence* selalu ditetapkan untuk setiap kelas pada satu sampel dan nilai posterior akan dibandingkan dengan nilai posterior dari kelas lainnya untuk mengklasifikasikan kelas (Marlina et al., 2016).

- KNN

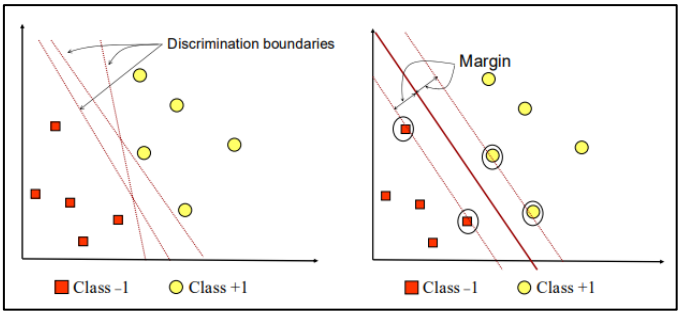
  Prinsip kerja K-Nearest Neighbor (KNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (neighbor) terdekatnya dalam data pelatihan [1]. Teknik ini termasuk dalam kelompok klasifikasi nonparametric. Di sini kita tidak memperhatikan distribusi dari data yang ingin kita kelompokkan. Teknik ini sangat sederhana dan mudah diimplementasikan. Mirip dengan teknik klastering, kita mengelompokkan suatu data baru berdasarkan jarak data baru itu ke beberapa data/tetangga (neighbor) terdekat (Santosa, 2007).

  Tujuan algoritma KNN adalah mengklasifikasikan obyek baru berdasarkan atribut dan training sample. Clasifier tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik query, akan ditemukan sejumlah k obyek atau (titik *training*) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari k obyek. Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru. Algoritma metode KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke training sample untuk menentukan KNN-nya.

- SVM

  Support Vector Machine, juga sering disingkat menjadi SVM, merupakan salah satu algoritma *supervised learning* yang didasari dengan pencarian fungsi pemisah (*hyperplane)*

terbaik ataupu`xn yang paling tepat untuk memisahkan beberapa class yang ada pada dataset (Nugroho et al., 2003). Metode yang dikembangkan oleh Boser, Guyon, Vapnik pada tahun 1992 ini menerapkan klasifikasi linear (*linear classification)* sebagai prinsip dasarnya (Srivastava & Bambhu, 2005).



Gambar 3.1 Support Vector Machine

Secara sederhana, konsep bekerja algoritma SVM adalah mencari *hyperplane* terbaik guna memisahkan dua buah class pada *input space*. Gambar x menunjukkan beberapa *data points* yang terbagi menjadi dua kelas, yaitu class -1 dan +1. Sementara, garis-garis berwarna merah menunjukkan berbagai alternatif garis pemisah (*discrimination boundaries)*. Garis berwarna merah tebal menandakan hyperplane terbaik yang berlaku sebagai pemisah antar kedua class (*separating hyperplane*).

2) *Prior Knowledge*

Prior Knowlegde merupakan informasi awal yang dimiliki oleh peneliti maupun *data scientist* (ilmuwan data) sebelum melakukan modeling ataupun penerapan algoritma yang akan menghasilkan informasi atau insight baru (*posterior knowledge*). Objektivitas dari tahapan *prior knowlegde* adalah untuk mendapatkan atau mengumpulkan informasi dari dua aspek, yaitu bisnis dan data. Pada penelitian ini, *prior knowledge* diperoleh dari dua aspek tahapan awal pada siklus CRISP-DM, yaitu *business understanding* dan *data understanding*. Penjabaran *prior knowledge* dari dua aspek tersebut, adalah sebagai berikut:

a. *Business Understanding*

i. *Business Objectives*

Bank X memiliki suatu program deposit di mana para pelanggan bank dapat membuka deposito atau simpanan yang hanya bisa diambil pada jangka waktu tertentu dengan tambahan beberapa persen bunga. Di sini goal dari bank X adalah ingin meningkatkan jumlah pelanggan yang membuka tabungan deposito. Beberapa pertanyaan bisnis (business related question) yang bisa diasumsikan:

- Apakah kualitas pelayanan bank mempengaruhi ketertarikan pelanggan membuka deposito?
- Apakah bunga deposito yang besar dapat meningkatkan ketertarikan pelanggan untuk membuka tabungan deposito?

ii. *Assess the Situation*

Setelah dilakukan pemahaman terhadap goal dari bisnis perusahaan, maka kemudian peneliti akan melakukan observasi terhadap aspek-aspek bisnis mana sajakah yang penting yang dapat dijadikan sebagai bahan untuk analisa permasalahan. Sekaligus, aspek-aspek bisnis ini akan menjadi sumber dari data yang kemudian dipakai dalam permodelan nanti. Beberapa aspek bisnis yang dipilih adalah sebagai berikut:

- Sales
- Marketing
- Customer service

Alasan dipilih ketiga aspek ini adalah karena aspek-aspek tersebut berhubungan langsung terhadap goal dari bisnis, yaitu melakukan promosi ataupun penawaran terhadap produk bank (simpanan deposito).

iii. *Data Mining Goals*

Pada tahapan ini, setelah goal dari bisnis dan aspek-aspek apa saja yang perlu dilihat sudah diobservasi maka peneliti kemudian akan merumuskan langkah solusi apa yang bisa diambil guna mengatasi masalah yang sedang terjadi. Di sini, peneliti merumuskan solusi dengan akan melakukan klasifikasi pada data pelanggan bank X. Klasifikasi ini bertujuan untuk melihat pola serta insight yang ada pada data pelanggan yang terbagi ke dalam dua kategori: yang mempunyai tabungan deposito dan yang tidak mempunyai tabungan deposito. Dari data mining goals ini, diharapkan perusahaan bank X dapat mengetahui tipe pelanggan seperti apa yang seharusnya mereka tawarkan produk.

iv. *Project Plan*

Secara umum perencanaan proyek data science (menggunakan kerangka CRISP-DM) dibagi ke dalam beberapa persen bagian, sebagai berikut:

i. 50% - 70% - Data Preparation
ii. 20% - 30% - Data Understanding
iii. 10% - 20% - Modeling, Evaluation, dan Business Understanding
iv. 5% - 10% - Deployment

b. *Data Understanding*

i. *Collect Data*

Pada tahapan ini, penelit melakukan ekstraksi terhadap data-data yang dibutuhkan. Umumnya, data bisa diambil dari beberapa sumber dan hal tersebut akan menyebabkan keterlembatan pada proses analisa. Di sini peneliti mengambil data kampanye pemasaran bank X yang telah diunduh dari situs Kaggle.

ii. *Describe the Data*

Data ini merupakan data direct marketing yang telah dilakukan oleh divisi Marketing Bank X. Pihak Marketing melakukan penawaran melalui saluran telepon, dan terkadang satu pelanggan bisa ditelpon lebih dari sekali untuk memastikan apakah si pelanggan berminat untuk membuka tabungan deposito atau tidak.

iii. *Explore the Data*

Data terdiri atas 20 atribut masukan (input) yang merupakan data-data pelanggan. Di mana atribut tersebut terbagi ke dalam 2 bagian, yaitu data pelanggan dan data riwayat penawaran (campaign).

- Data Pelanggan terdiri dari: umur, pekerjaan, status hubungan, pendidikan, status kredit, status pinjaman, status kredit rumah, rata-rata pendapatan pertahun
- Data Penawaran tediri dari: jenis kontak (hp atau telepon), kontak terakhir (bulan), kontak terakhir (hari), durasi percakapan, jumlah dikontak untuk penawaran deposito, interval hari setelah kontak terakhir dengan pelanggan, jumlah dikontak selain untuk menawarkan deposito, hasil dari marketing sebelumnya

iv. *Verify the Data*

Pada data kali ini peneliti tidak menemukan kejanggalan pada data, terutama untuk nilai null serta outliers (misalkan umur < 15 tahun lalu sudah memiliki kredit). Maka dapat disimpulkan kualitas data masih dalam keadaan baik, namun akan dilakukan eksplorasi statistika deskriptif lebih lanjut pada tahapan Data Preparation.

3) *Preparation (Tools & Programming Language)*

Tools ataupun perangkat lunak serta bahasa pemrograman yang dipakai pada penelitian ini mencangkup *Python 3* sebagai bahasa pemrograman yang mendukung pembuatan model, kalkulasi hingga melakukan komputasi-komputasi *machine learning*. *Jupyter Notebook* digunakan untuk melakukan *live code*, sedangkan terdapat pula *libraries* yang digunakan pada penelitian ini yaitu sebagai berikut:

- Numpy — untuk melakukan perhitungan dan kalkulasi aljabar/matematika
- Pandas — untuk melakukan manipulasi ataupun modifikasi terhadap *dataframe*
- Scikit-learn — untuk memanggil fungsi yang merupakan algoritma-algoritma *machine learning* yang digunakan pada penelitian ini serta melakukan *preprocessing* terhadap data
- Matplotlib dan Seaborn — untuk membuat visualisasi daripada data

4. HASIL DAN PEMBAHASAN

Setelah melakukan persiapan terhadap data (ekstrasi serta pemahaman) dan *tools*, maka peneliti akan melakukan tahapan selanjutnya yaitu *data preparation* untuk melihat lebih jauh mengenai deskripsi data dari segi teknis dan juga *preprocessing* untuk mentransformasi data sehingga dapat digunakan pada tahapan *modeling* nanti.

1) *Data Exploration*

- *Shape*

  Ukuran atau *volume* dari sebuah data memiliki pengaruh terhadap akurasi dari sebuah model klasifikasi. Data yang terlalu sedikit dapat menyebabkan model tidak bisa melakukan

generalisasi secara utuh terhadap *unseen data* dikarenakan terdapat bias pada kalkulasi. Di sini, peneliti menggunakan data kampanye pemasaran Bank X dengan ukuran data sebagai berikut:

Baris (Rows)	Kolom (Attributes)
45211	17

Tabel 4.1 *Shape of data*

- *In-Depth Details*

  Pengecekan masing-masing *attributes* dilakukan untuk memastikan tidak terdapat nilai kosong pada suatu baris yang dapat menyebabkan penurunan performa dari model klasifikasi. Hasil pengecekan yang telah dilakukan adalah sebagai berikut:

Column	Non-null	Count	Dtype
Age	45211	Non-null	Int64
Job	45211	Non-null	Object
Marital	45211	Non-null	Object
Education	45211	Non-null	Object
Default	45211	Non-null	Object
Balance	45211	Non-null	Int64
Housing	45211	Non-null	Object
Loan	45211	Non-null	Object
Contact	45211	Non-null	Object
Day	45211	Non-null	Int64
Month	45211	Non-null	Object
Duration	45211	Non-null	Int64
Campaign	45211	Non-null	Int64
Pdays	45211	Non-null	Int64
Previous	45211	Non-null	Int64
Poutcome	45211	Non-null	Object
y	45211	Non-null	Int64

Tabel 4. 2 *Indepth Details*

- *Descriptive Analytics*

  Analisis deskriptif merupakan suatu metode yang berfungsi untuk mendeskripsikan isi dari sebuah data secara singkat untuk memberi gambaran gambaran terhadap objek yang diteliti

	age	balance	day	duration	campaign	pdays	previous	y
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323	0.116985
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441	0.321406
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000	1.000000

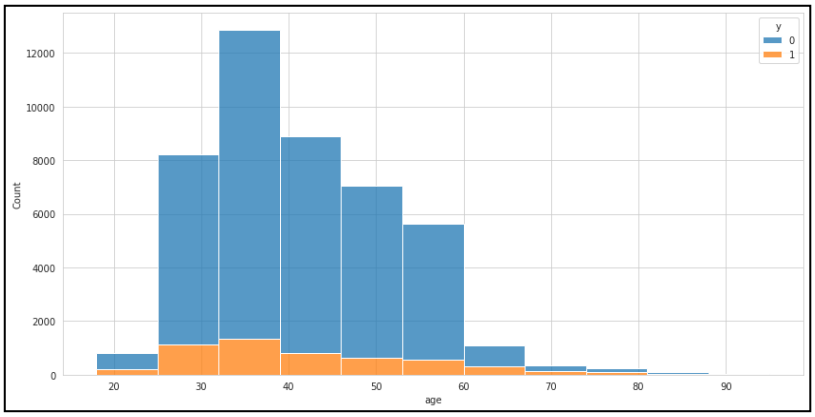
melalui data atau sampel yang telah terkumpul sebagaimana adanya (Lawless, 2010). Hasil dari analisis deskriptif pada data kampanye pemasaran bank X adalah sebagai berikut:

Gambar 4. 1 *Descriptive Analytics*

Dari hasil luaran deskriptif di atas, maka kita bisa menyimpulkan beberapa insight sebagai berikut:

- Usia termuda klien bank adalah 18, sedangkan usia tertua klien adalah 95. Rata-rata usia klien bank adalah 41
- Pendapatan rata-rata klien bank adalah 1528,54
- Durasi terpanjang dalam sekali penelponan adalah 3881 detik, sedangkan durasi terpendeknya adalah 2 detik, sepertinya klien yang ditelpon sangat tidak tertarik dan langsung menutup telepon tersebut
- Jumlah penelponan terbanyakan pada satu klien adalah 63 kali, sedangkan jumlah paling sedikitnya adalah 1

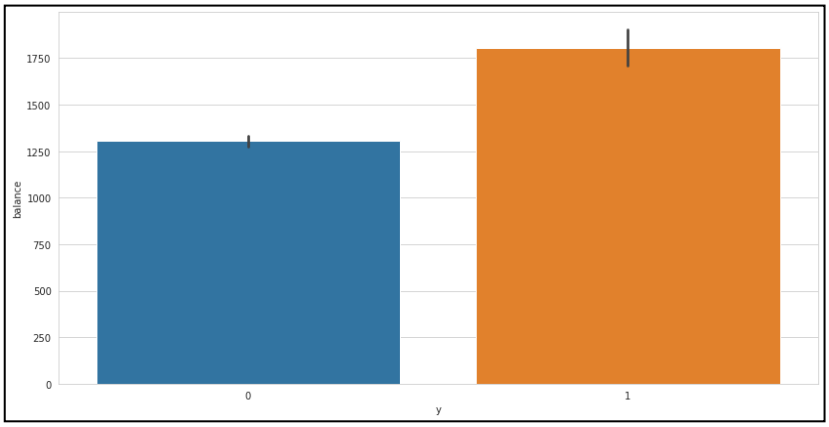
- *Visualizations*
  - *Age*



Gambar 4. 2 Proporsi persebaran umur

Dari visualisasi di atas dapat dilihat bahwa persebaran umur paling tinggi dari pelanggan bank X yaitu berkisar 30 hingga 40 tahun.

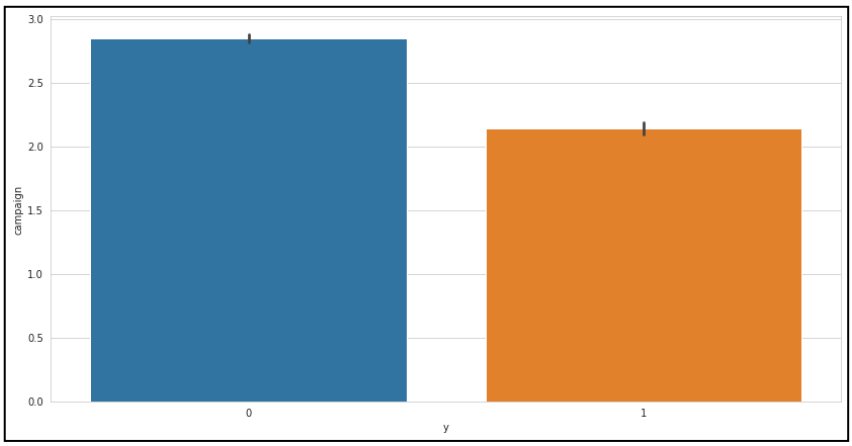
  - *Balance*



Gambar 4. 3 Pendapatan dan ketertarikan dengan *campaign*

Pada visualisasi di atas, dapat dilihat bahwa pelanggan bank yang memiliki pendapatan lebih dari 1,200 dolar lebih cenderung akan tertarik dengan *campaign* bank.

- *Campaign*

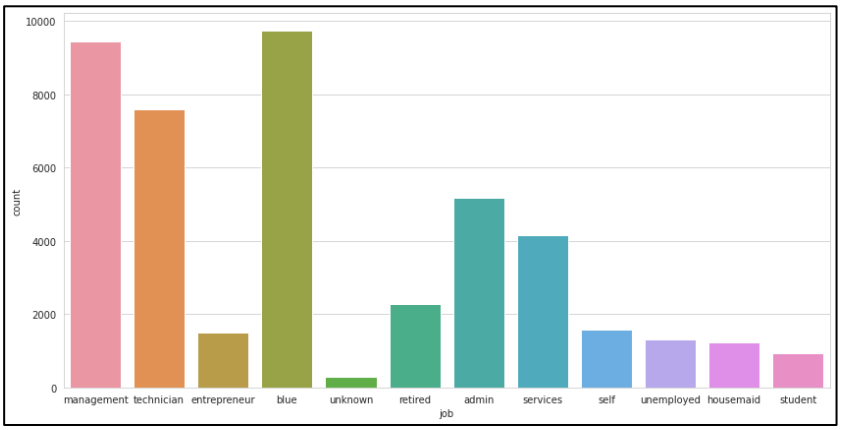


Gambar 4. 4 Jumlah riwayat penawaran *campaign*

Dari visualisasi di atas dapat dilihat bahwa pelanggan yang telah ditawarkan *campaign* sebanyak lebih dari 2 kali maka akan cenderung tidak tertarik dengan *campaign*. Asumsinya adalah pelanggan merasa terganggu dengan telemarketing yang dilakukan oleh bank sehingga malah menjadi tidak tertarik dengan program yang ditawarkan.

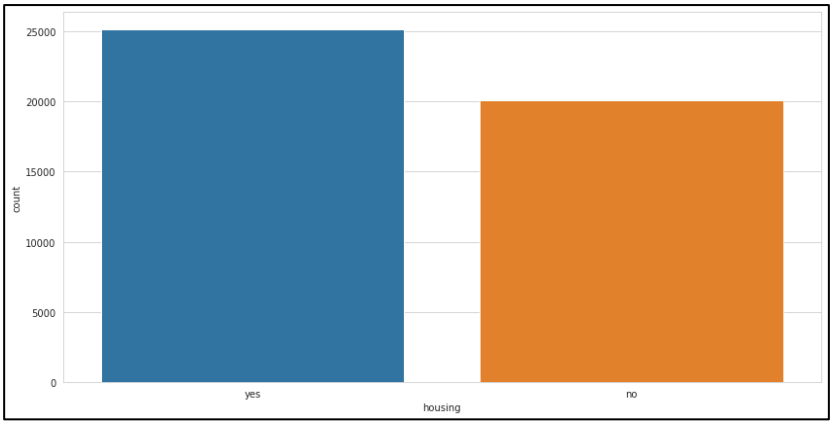
- *Categorical Variables*

Data yang digunakan pada penelitian ini memiliki 5 tipe atribut kategorikal, yaitu sebagai berikut:



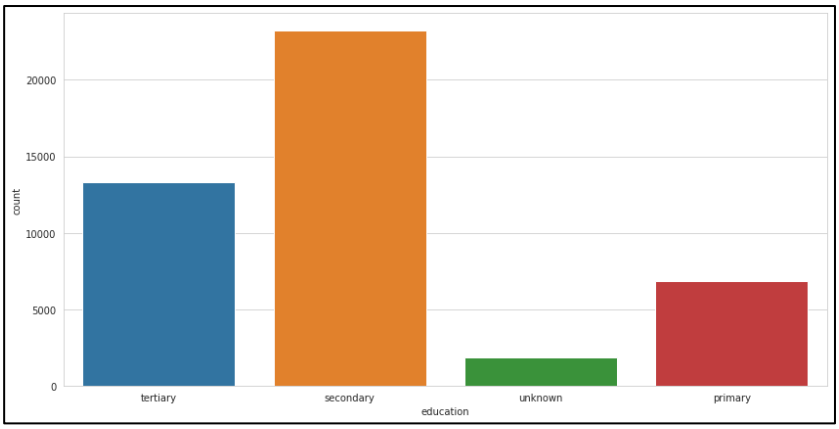
Gambar 4. 5 Proporsi Pekerjaan

Barplot di atas memperlihatkan persebaran proporsi dari jenis pekerjaan tiap pelanggan bank X. Dapat dilihat bahwa pelanggan bank X paling banyak berprofesi sebagai buruh (*blue collar*) dan yang kedua terbanyak adalah manajemen.



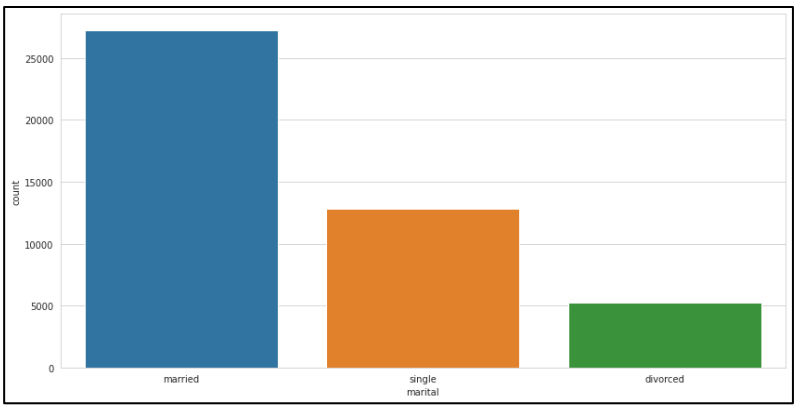
Gambar 4. 6 Memiliki Pinjaman Kredit Rumah

Barplot di atas memperlihatkan jumlah pelanggan yang memiliki pinjaman kredit rumah. Dapat dilihat bahwa pelanggan bank X banyak yang memiliki pinjaman kredit rumah ketimbang yang tidak memiliki pinjaman.



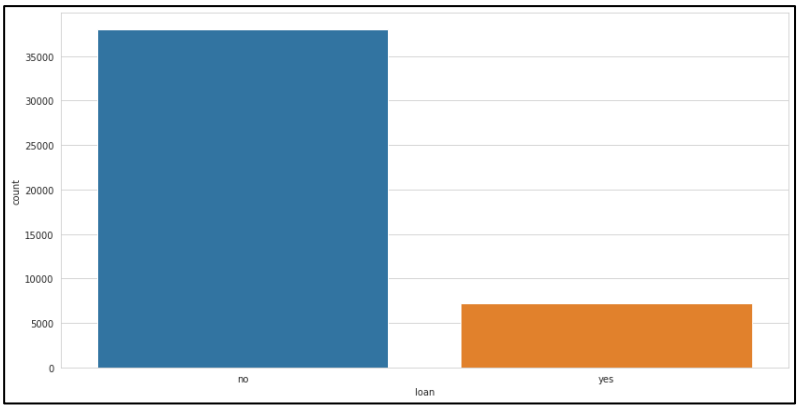
Gambar 4. 7 Pendidikan

Barplot di atas memperlihatkan persebaran proporsi dari jenis pendidikan tiap pelanggan bank X. Dapat dilihat bahwa pelanggan bank X paling banyak memiliki riwayat pendidikan SMA (secondary) dan yang kedua terbanyak adalah universitas atau sarjana/diploma.



Gambar 4. 8 Status

Barplot di atas memperlihatkan persebaran proporsi dari status hubungan pelanggan bank X. Dapat dilihat bahwa pelanggan bank X paling banyak berstatus menikah (*married*) dan yang kedua terbanyak adalah belum menikah (*single marital*).

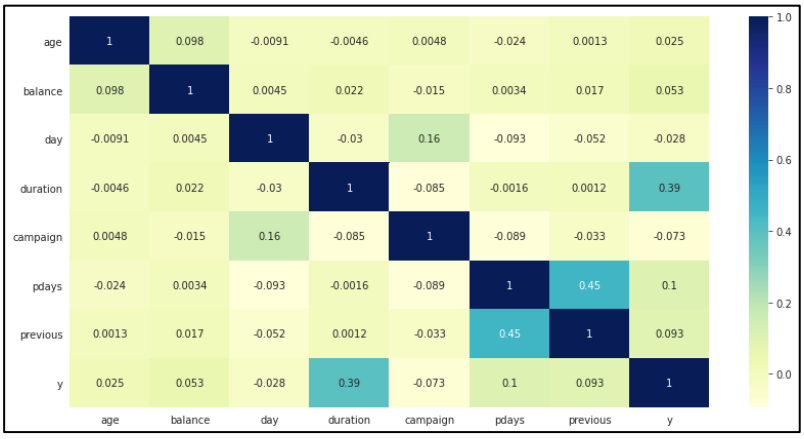


Gambar 4. 9 Proporsi Pekerjaan

Barplot di atas memperlihatkan jumlah pelanggan yang memiliki pinjaman di bank. Dapat dilihat bahwa pelanggan bank X banyak yang tidak memiliki pinjaman di bank ketimbang yang memiliki pinjaman. Namun, lebih banyak pelanggan yang memiliki pinjaman kredit rumah, asumsinya adalah dikarenakan lebih banyak pelanggan bank X yang berstatus sudah menikah maka besar kemungkinan pinjaman kredit rumah lebih diminati dibanding pinjaman bank.

- *Numerical Variables*

Data yang digunakan pada penelitian ini memiliki 8 tipe atribut numerikal, korelasi dari tiap atribut tersebut yaitu sebagai berikut:



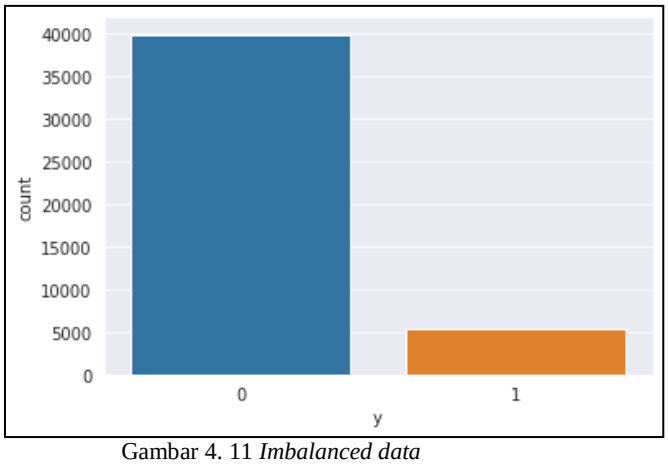
Gambar 4. 10 Korelasi Atribut Numerikal

Pada *heatmap* di atas dapat dilihat bahwa atribut yang berkorelasi paling tinggi dengan ketertarikan pelanggan terhadap kampanye pemasaran adalah *duration* yang mana asumsinya adalah semakin lama durasi telepon antara karyawan bank dan pelanggan maka jelas pelanggan mulai tertarik dengan program bank sedangkan jika durasinya cepat maka bisa dibilang pelanggan tidak tertarik dan ingin buru-buru langsung menyudahi percakapan telepon.

2) *Data Preprocessing*

- *Downsampling*

Pada data kampanye pemasaran bank X terdapat ketidakseimbangan (*imbalanced*) terhadap *class*, di mana *class 1* memiliki data sebanyak 5289 sedangkan *class 0* memiliki data sebanyak 39922.



Gambar 4. 11 *Imbalanced data*

Ketidakseimbangan data ini memiliki pengaruh terhadap performa model klasifikasi yang mana model akan lebih banyak mempelajari data dari salah satu kelas yang memiliki data lebih banyak sehingga dapat menyebabkan bias. Untuk mengatasi permasalahan ini maka dilakukan *downsampling* atau pengurangan data pada *majority class* yang memiliki jumlah data lebih banyak. Jumlah pengurangan data mengikuti proporsi dari *minority class* yaitu sebesar 5289.

- *Encode*

  Encoding bertujuan untuk mengubah data kategorikal huruf (*string*) menjadi bentuk ordinal sehingga nilai dapat dibaca oleh komputasi dan perhitungan kalkulasi dapat dilakukan oleh model. Pada penelitian ini terdapat 9 atribut yang memiliki tipe data kategorikal *string*. Peneliti melakukan *preprocessing* pada tiap atribut tersebut dengan mengubah tipe data menjadi bentuk ordinal dengan menggunakan *label encoder*.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
36	management	single	secondary	no	1511	yes	no	cellular	16	nov	270	1	-1	0	unknown	1
38	technician	married	secondary	no	557	yes	no	cellular	16	nov	1556	4	-1	0	unknown	1
53	management	married	tertiary	no	583	no	no	cellular	17	nov	226	1	184	4	success	1
34	admin	single	secondary	no	557	no	no	cellular	17	nov	224	1	-1	0	unknown	1
23	student	single	tertiary	no	113	no	no	cellular	17	nov	266	1	-1	0	unknown	1

Gambar 4. 12 Sebelum dilakukan *encoding*

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
37	4	1	2	0	413	0	0	0	16	10	138	2	-1	0	3	0
46	4	1	1	0	5127	0	0	0	6	1	472	1	-1	0	3	0
42	6	1	1	0	13242	0	0	2	9	6	149	1	-1	0	3	0
31	9	2	1	0	1	1	0	2	20	8	71	3	-1	0	3	0
32	9	2	2	0	696	0	1	0	30	0	270	2	293	2	0	0

Gambar 4. 13 Setelah dilakukan *encoding*

- *Splitting Dataset*

  Data yang telah dilakukan transformasi kemudian dibagi menjadi dua set, yaitu *training* dan *testing*. Peneliti membagi data dengan proporsi pembagian 80:20 yaitu, 80% untuk training dan 20% untuk testing.

3) *Modeling*

- ***Decision Tree***

  Pada permodelan dengan menggunakan *decision tree*, digunakan *max_depth* yaitu 10 dan *random_state* yaitu 32932. Model *decision tree* ini menghasilkan akurasi training sebesar 88% dan akurasi testing sebesar 81%

- ***Random Forest***

  Pada permodelan dengan menggunakan teknik *random forest*, digunakan *max_depth* yaitu 10 dengan *random_state* yaitu 32932 yang menghasilkan akurasi training sebesar 89% dan akurasi testing sebesar 84%

- ***Naïve Bayes***

  Pada permodelan dengan menggunakan *naïve bayes*, digunakan parameter default yang menghasilkan akurasi training sebesar 75% dan akurasi testing sebesar 75%

- ***KNN***

  Pada permodelan dengan menggunakan KNN, digunakan *n_neighbors* yaitu 5 yang menghasilkan akurasi training sebesar 82% dan akurasi testing sebesar 75%

- ***SVM***

  Pada permodelan dengan menggunakan SVM, digunakan *random_state* yaitu 32932 yang menghasilkan akurasi training sebesar 73% dan akurasi testing sebesar 73%

MODEL	AKURASI TRAINING	AKURASI TESTING
Decision Tree	.88	.81
Random Forest	.89	.84
Naïve Bayes	.75	.75
KNN	.82	.75
SVM	.73	.73

Tabel 4.14 Perbandingan akurasi antara algoritma

Pada tabel hasil akurasi dari tiap algoritma di atas dapat terlihat bahwa dua dari lima algoritma yang dicoba memiliki hasil akurasi yang sama antara training dan testing. Sedangkan, algoritma lain memiliki akurasi training yang lebih tinggi dibanding akurasi testing. Di sini terdapat *overfitting* terhadap model yang telah dibuat dan divalidasi, di mana model lebih merespon secara akurat terhadap data training dibanding data testing. *Overfitting* sendiri merupakan suatu kejadian di mana model dapat dengan akurat memprediksi data training namun tidak mampu memprediksi *unseen data* (testing) dengan akurat sehingga dapat menyebabkan suatu bias pada data hasil prediksi (Lever et al, 2016).

Overfitting pada dasarnya disebabkan oleh model yang terlalu kompleks sehingga menangkap bias dan varians yang terlalu tinggi. *Noise* pada data training dikalkulasikan terlalu detail, sehingga model mempelajari (*learning*) pola pada data training dengan sangat akurat hingga gagal menerjemahkan atau mengimplementasikan *knowledge* yang telah dipelajari pada data testing atau data baru. Untuk mengatasi masalah ini, maka peneliti akan melakukan *cross-validation* pada model. *Cross-validation* merupakan teknik *resampling* pada data yang berfungsi untuk melakukan estimasi terhadap kemampuan model dalam memprediksi serta untuk meminimalisir *overfitting* (Berrar, 2019). Walaupun tidak sepenuhnya dapat menghilangkan *overfitting* pada data, namun dengan

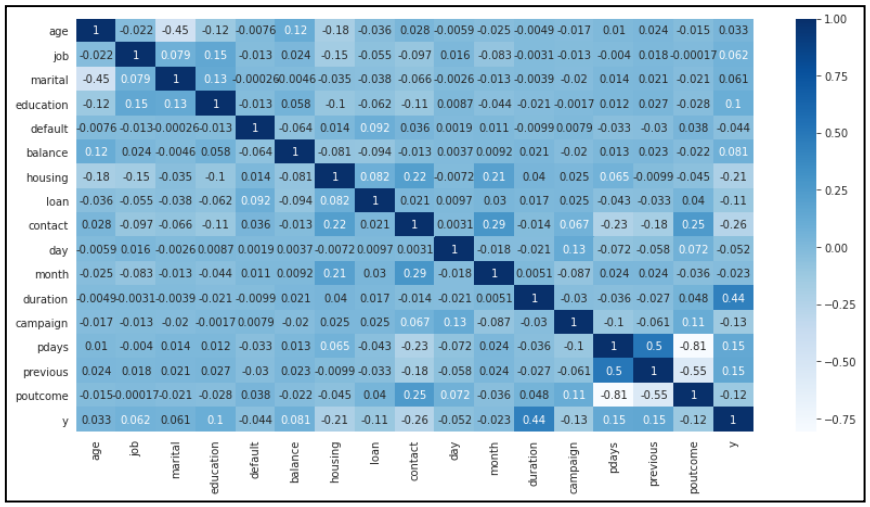
menggunakan *cross-validate* setidaknya dapat mengestimasi seberapa akurat model yang telah dibangun dengan cara melakukan *resampling* dan *subsetting* pada training data dan melakukan validasi di setiap iterasi atau "fold" sehingga dapat diketahui akurasi akhirnya dengan melihat rata-rata hasil akurasi dari tiap fold-nya. Akurasi akhir ini yang dipakai menjadi acuan dari model terbaik dengan *overfitting* paling minim.

Pada langkah selanjutnya, peneliti akan melakukan modeling sekali lagi untuk melakukan estimasi terhadap hasil akhir akurasi yang dirasa tidak bagus dikarenakan terdapat *overfitting* pada modeling yang pertama. Modeling kedua ini dilakukan dengan tahapan *improvement* atau perbaikan, yaitu *feature selection*, normalisasi (*normalize*), dan terakhir adalah *cross-validation*.

4) *Modeling (Improvement and Estimating)*

- ***Feature Selection***

  Feature selection merupakan tahapan perbaikan yang digunakan peneliti dengan tujuan untuk menyeleksi fitur-fitur ataupun atribut pada data yang tidak memiliki '*benefit*' terhadap hasil dari permodelan. *Feature selection* dilakukan dengan melihat nilai signifikansi *codependency* atau hubungan satu sisi negatif dari tiap variabel yang ada.



Gambar 4. 15 *Heatmap* korelasi antara variabel

Pada *heatmap* di atas terlihat terdapat tiga pasang variabel yang memiliki korelasi lebih dari sama dengan 0.5, yaitu *(pdays, previous), (pdays, outcome), (previous, outcome)*. Hal ini mengindikasikan terdapat multikolinearitas pada data yang dapat menyebabkan bias dikarenakan linearitasnya cukup kuat. Pada kasus klasifikasi, linearitas perlu dihindari untuk meminimalisir suatu bias pada data, pola-pola acak lebih diperuntukan agar model dapat mempelajari data dengan signifikan ketimbang hanya menyimpulkan *insight* secara linear.

Peneliti akan menghapus tiga varibel dari setiap pasangan tersebut, yaitu *pdays, previous,* dan *outcome*. Secara realita, ketiga variabel ini pun termasuk ke dalam data primer yang mana akan di-*generate* setelah dilakukannya suatu riset sedangkan klasifikasi di sini diperuntukkan untuk melakukan prediksi terhadap data sekunder dari riwayat informasi pelanggan.

- ***Normalize***

Normalisasi dilakukan dengan tujuan untuk menyamakan nilai setiap baris pada data dengan mengikuti nilai paling relatif (*relative scale*). Data sebelum dilakukan normalisasi:

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
37	4	1	2	0	413	0	0	0	16	10	138	2
46	4	1	1	0	5127	0	0	0	6	1	472	1
42	6	1	1	0	13242	0	0	2	9	6	149	1
31	9	2	1	0	1	1	0	2	20	8	71	3
32	9	2	2	0	696	0	1	0	30	0	270	2

Gambar 4. 16 Sebelum normalisasi

0	1	2	3	4	5	6	7	8	9	10	11	12
0.246753	0.363636	0.5	0.666667	0.0	0.041193	0.0	0.0	0.0	0.500000	0.909091	0.035558	0.020408
0.363636	0.363636	0.5	0.333333	0.0	0.097137	0.0	0.0	0.0	0.166667	0.090909	0.121618	0.000000
0.311688	0.545455	0.5	0.333333	0.0	0.193444	0.0	0.0	1.0	0.266667	0.545455	0.038392	0.000000
0.168831	0.818182	1.0	0.333333	0.0	0.036303	1.0	0.0	1.0	0.633333	0.727273	0.018294	0.040816
0.181818	0.818182	1.0	0.666667	0.0	0.044552	0.0	1.0	0.0	0.966667	0.000000	0.069570	0.020408

Gambar 4. 17 Setelah normalisasi

- ***Cross-Validation***

  Pada tahapan *cross-validation*, peneliti menggunakan 10 kali *folds* dengan metrik skor yang diambil adalah *accuracy*, *precision*, *recall*, dan *f1_score*. Kemudian diambil nilai rata-rata hasil metrik dari setiap sepuluh kali iterasi tersebut, yang menghasilkan hasil sebagai berikut:

	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	KNN	Best Score
Accuracy	0.713095	0.673865	0.748544	0.706383	0.633593	Random Forest
Precision	0.740993	0.704163	0.765234	0.671454	0.663533	Random Forest
Recall	0.640778	0.560609	0.707704	0.786369	0.501629	Gaussian Naive Bayes
F1 Score	0.683875	0.615537	0.731885	0.722897	0.564871	Random Forest

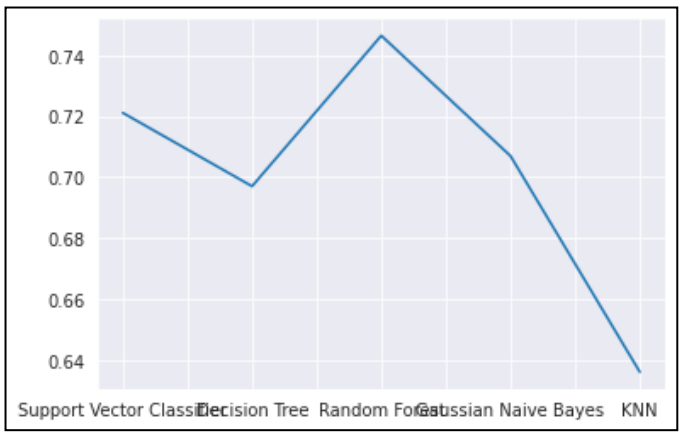
Gambar 4. 18 Hasil *cross-validation*

Dari gambar *dataframe* di atas dapat terlihat bahwa *random forest* memiliki tingkat rata-rata metrik yang lebih tinggi dibanding empat algoritma lainnya. Sedangkan, *naïve bayes* memiliki keunggulan 0.8% dibanding *random forest* pada metrik *recall*.

5) *Evaluation*

Akurasi yang dihasilkan dari modeling tahap awal mengindikasikan terdapat *overfitting* terhadap tiga model dari total lima model yang dicoba. *Cross-validation* dipakai bukan untuk menghilangkan *overfitting* namun untuk mengestimasi hasil akurasi dari n-folds yang telah dilakukan iterasi pada

sample dan subset data training yang berbeda. Peneliti telah melakukan *cross-validation* pada data kampanye pemasaran bank X dan mendapatkan hasil akurasi dari tiap model, sebagai berikut:



Gambar 4. 19 Hasil *mean* akurasi dari 10 *CV Folds*

Dari hasil akurasi *cross-validation* tersebut dapat dilihat bahwa *random forest* memiliki tingkat akurasi paling tinggi, yaitu 75%. Pada modeling tahap awal pun dapat dilihat *random forest* juga memiliki tingkat akurasi yang cukup tinggi pada data testing. Estimasi *cross-validation* ini dilakukan untuk melihat seberapa akurat model dalam memprediksi data dengan mempertimbangkan beberapa subset training sehingga menghindari terjadi bias pada prediksi data testing. Dari rata-rata akurasi 75% ini dapat disimpulkan bahwa model *random forest* telah bekerja cukup baik dalam melakukan prediksi pada *unseen data*. Walaupun hasil akurasi cukup jauh dibanding modeling tahap awal, namun estimasi *cross-validation* dapat dipakai sebagai acuan dari model yang memiliki tingkat *overfitting* cukup rendah sehingga meminimalisir terjadinya bias pada data baru yang nantinya akan dipakai dalam *deployment* model sebagai suatu perangkat lunak.

Dari hasil persiapan hingga modeling yang dilakukan dapat disimpulkan bahwa akurasi yang tinggi pada modeling tahap awal terjadi akibat model yang terlalu kompleks dan terlalu detail dalam mengkalkulasikan noise dan pola pada data. Perbaikan yang dilakukan dengan beberapa teknik seperti *feature selection*, normalisasi hingga *cross-validation* tidak membuktikan bahwa model telah terbebas dari masalah *overfitting* namun estimasi *cross-validation* setidaknya dapat menjadi 'acuan tengah' dari akurasi model dalam memprediksi *unseen data*.

5. PENUTUP

Pada penelitian ini, peneliti telah melakukan analisa hingga percobaan modeling terhadap data kampanye pemasaran menggunakan lima algoritma yang terpilih, yaitu *decision tree, random forest, k-nearest neighbors, naïve bayes*, dan *support vector machine*. Dari hasil analisa data, peneliti menemukan atribut yang memiliki relevansi tinggi serta faktor yang kuat terhadap *decision making* yaitu durasi (*duration*) dari telepon. Namun, pada kenyataannya di dunia nyata atribut ini baru akan diketahui ketika bank baru ingin menawari program kepada si pelanggan. Atribut-atribut sekunder seperti biodata pelanggan hingga riwayat kampanye sebelumnya menjadi atribut-atribut yang lebih efektif untuk mengukur tingkat ketertarikan pelanggan terhadap program yang akan ditawarkan.

Evaluasi dari permodelan yang telah dilakukan adalah *random forest* memiliki tingkat akurasi yang paling tinggi dalam mengklasifikasi data sedangkan yang paling rendah adalah KNN. Meskipun demikian, faktor-

faktor seperti *randomness sampling* hingga *hyperparatemer* yang tidak cocok dapat menjadi penyebab *misclassfication* yang tinggi pada beberapa algoritma lain. Penggunaan *grid search* sebaiknya diimplementasikan bersamaan dengan *cross-validation* sehingga model dapat melakukan *tuning* untuk mencari *parameter* mana yang lebih sesuai dengan data yang sedang di-*train*.

Akhir kata, kendati hasil akurasi yang sebenarnya tidak cukup untuk dikatakan sempurna dikarenakan di estimasi akurasi akhir hanya di bawah 80%, namun dari hasil penelitian serta evaluasi yang telah dilakukan dapat disimpulkan bahwa metode pohon (*tree-based*) bisa menjadi metode yang cukup efektif untuk mendukung perencanaan serta manajemen keputusan kampanye pemasaran yang nanti akan dilakukan oleh bank X ke depannya.

DAFTAR PUSTAKA

Bahrawi, N. (2019). Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based. Journal of Information Technology and Its Utilization, 2(2), 29. doi:10.30818/jitu.2.2.2695

Berrar, D. (2019). Cross-validation. *Encyclopedia of bioinformatics and computational biology*, *1*, 542-545.

Grzonka, D., Suchacka, G., & Borowik, B. (2016). Application of selected supervised classification methods to bank marketing campaign. *Information Systems in Management*, *5*(1), 36-48.
Han, J. & Kamber, M. 2006. Data Mining Concept and Tehniques. San Fransisco: Morgan Kauffman.

Hill, R. K. (2016). What an algorithm is. *Philosophy & Technology*, *29*(1), 35-59.

Larasati, D. A. H., & Sutrisno, T. (2018, October). Tourism Site Recommendation in Jakarta Using Decision Tree Method Based on Web Review. In International Conference on Information Technology, Engineering, Science & its Applications.

Lawless, H. T., & Heymann, H. (2010). Descriptive analysis. In *Sensory evaluation of food* (pp. 227-257). Springer, New York, NY.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting.

Marlina, L., Muslim, M., Siahaan, A. U., & Utama, P. (2016). Data Mining Classification Comparison (Naïve Bayes and C4. 5 Algorithms). Int. J. Eng. Trends Technol, 38(7), 380–383.

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). *Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika*. IlmuKomputer.

Parveen, H., & Pandey, S. (2016). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT), 416–419. https://doi.org/10.1109/ICATCCT.2016.7912034

Santosa, B., 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Ed.1, Graha Ilmu, Yogyakarta.

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88–91.

Srivastava, D., & Bambhu, L. (2005). Data Classification using Support Vector Machine. Journal of Theoretical and Applied Information Technology. 2.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), 4.

Wisaeng, K. (2013). A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE)*, *3*(4), 116-119.