Model Prediksi Hasil Keputusan Nasabah Berlangganan Deposito Berjangka Panjang EverBank Menggunakan Metode *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*

Kelly Mae
Information Systems
Multmimedia Nusantara University
Banten, Indonesia
kelly.mae@student.umn.ac.id

Abstract*—Classification merupakan salah satu Teknik *data mining* yang akan menugaskan suatu variabel dalam sekumpulan *variable* ke dalam variabel target guna memprediksi variabel target secara akurat untuk setiap situasi dalam suatu *dataset*. Jenis teknik *data mining* ini dapat digunakan sebagai salah satu implementasi *Data Science* pada penelitian dalam dunia perbankan (*Banking*). Subjek penelitian ini adalah nasabah EverBank yang menggunakan langganan deposito berjangka panjang (*long-term deposit*) yang telah berpartisipasi pada kampanye sebelumnya. Objek penelitian adalah hasil keputusan nasabah EverBank atas kelanjutan dalam penggunaan deposit berjangka panjangg. Adapun variabel target yang diteliti berupa *output* atau hasil keputusan nasabah EverBank yang diperoleh dari kampanye, sedangkan variabel respons berupa data nasabah EverBank, data yang berhubungan dengan kontak selama kampanye. Tujuan dari penelitian ini adalah membuat model prediksi hasil keputusan dengan menggunakan algoritma *Decision Tree, Random Forest, Logistic Regression,* dan *Gradient Boosting* berdasarkan data nasabah EverBank. Kegiatan *Data Science Implementation* dilakukan dengan menggunakan SAS yang akan diolah untuk mendapatkan statistika, *Exploratory Data Analysis* (EDA) hingga pembuatan model beserta analisis pada setiap algoritma yang digunakan untuk mengetahui apakah nasabah EverBank akan melanjutkan berlangganan deposito berjangka panjang.

Keywords*—Classification, Data Mining, Data Science Implementation, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, SAS, Long-term Deposit

I. LATAR BELAKANG & PEMAHAMAN BISNIS

A. Latar Belakang

Tujuan hidup merupakan suatu hal yang ingin dicapai oleh setiap manusia. Setiap generasi sudah jelas mempunyai tujuan hidup tersendiri yang disebabkan oleh perbedaan kebutuhan pada setiap generasinya [1]. Dalam hal ini, jenis tujuan hidup pada setiap orang beragam, seperti memperoleh kebahagiaan, harta dan kekayaan, impian dalam hal pekerjaan dan pendidikan, dan lain sebagainya. Dengan hal ini, tujuan hidup memegang peranan yang penting bagi keberlangsungan hidup manusia.

Terdapat berbagai usaha yang harus dilakukan dalam mencapai tujuan hidupnya. Salah satu usaha yang dapat dilakukan memiliki keterkaitan dengan ekonomi keuangan, yaitu menabung. Menabung merupakan salah satu kegiatan untuk menyisihkan serta menyimpan sisa uang yang dalam jangka waktu tertentu [2]. Dengan hal ini, maka hasil tabungan tersebut dapat digunakan untuk segala sesuatu yang membutuhkan biaya yang relatif tinggi maupun mampu bertahan dalam jangka waktu yang panjang.

Manusia pada umumnya menabung dan cenderung menyimpannya ke dalam bank untuk menjamin keselamatan keuangannya. Adapun tujuan lain dari penyimpanan tabungan ke dalam bank berupa pengaturan dan perencanaan keuangan, seperti membayar biaya kebutuhan hidup sehari-hari, cicilan, membuat tabungan khusus untuk kebutuhan darurat, dan lain sebagainya. Disinilah bank memiliki peranan penting bagi perekonomian negara sebagai salah satu faktor pendukung dalam pembangunan negara dikarenakan ketergantungannya pada dinamika perkembangan dan kontribusi yang diberikan bersifat nyata [3]. Dalam hal ini, adapun deposito berjangka sebagai salah satu produk bank yang membantu pengaturan dan perencanaan keuangan yang berlangsung dalam waktu yang lama.

Deposito berjangka dapat diartikan sebagai suatu deposito yang diterbitkan dengan jangka waktu yang sudah ditentukan [4]. Dengan kata lain, deposito berjangka merupakan tabungan jangka panjang dimana waktu penarikannya dapat dilakukan berdasarkan suatu kontrak atau perjanjian yang ditetapkan oleh bank. Perjanjian tersebut telah melibatkan para nasabah dimana terdapat beberapa ketentuan yang harus disetujui. Adapun batas waktu penarikan deposito berjangka yang berbeda pada setiap bank, baik dalam bentuk bulanan, tahunan, bahkan hingga puluhan tahun. Dengan demikian, EverBank telah menawarkan deposito berjangka panjang yang dapat membantu nasabah dalam menjaga keamanan keuangannya. Hal ini bertujuan agar nasabah pada akhirnya dapat

memanfaatkan keuangannya untuk mencapai kebutuhan hidup dan tujuan hidup yang diinginkan.

B. Pemahaman Bisnis

Adapun proses *business understanding* atau pemahaman bisnis dimana perusahaan harus mengetahui tujuan dari bisnis yang didirikan sehingga mengerti akan kebutuhan dan proses yang harus diselenggarakan. Terdapat suatu tujuan utama EverBank dalam keberlangsungan usahanya, yaitu mengidentifikasi dan mencari suatu solusi akan bagaimana cara meningkatkan kinerja dalam memasarkan produknya secara langsung ke nasabahnya terkait dengan langganan deposito berjangka panjang (*long-term deposit*) dengan memerhatikan hasil dari kampanye sebelumnya. Deposito berjangka merupakan salah satu bentuk investasi yang telah dilakukan oleh nasabah dalam bentuk uang tunai. Dengan cara ini, deposito berjangka memegang peranan penting dalam menghasilkan pendapatan yang berkontribusi pada pertumbuhan dan perkembangan bank.

Pada umumnya, terdapat beberapa metode bagi perusahaan perbankan dalam mempromosikan produknya demi memperoleh pendapatan. Akan tetapi, pemasaran secara langsung (*direct marketing*) merupakan salah satu metode yang paling banyak digunakan oleh bank dikarenakan efisiensinya dalam menjangkau nasabah, terutama melalui telepon. Namun, metode ini melibatkan penambahan biaya dalam hal Sumber Daya Manusia (SDM) dan fasilitas seperti *call center*. Tidak hanya itu saja, terdapat situasi tertentu dimana banyak nasabah sering merasa risih dengan kegiatan promosi yang dilakukan oleh pihak *customer service* bank sehingga hal ini dapat mengganggu privasi mereka. Dengan demikian, EverBank memastikan bahwa strategi mereka akan berjalan secara efisien dengan hanya melakukan kontak sebanyak sekali dengan nasabah yang pada akhirnya dapat mengetahui apakah nasabah tersebut akan memperpanjang langganan deposito berjangka panjang atau tidak.

II. TINJAUAN TEORITIS & METODOLOGI PENELITIAN

A. Tinjauan Teoritis

1. *Decision Tree*

Decision tree dapat didefinisikan sebagai pohon di mana pengujian atribut dan *node* daun ditunjukkan oleh setiap *node* untuk memberikan klasifikasi. Dalam hal ini, klasifikasi dari contoh pengujian dimulai pada *node* akar yang menguji nilai fitur per *node* dan mengurutkan ke cabang yang sesuai hingga mencapai klasifikasi melalui *node* daun [5].

Terdapat beberapa algoritma *decision tree* yang digunakan untuk klasifikasi, seperti *decision tree* C4.5, ID3, dan C5.0. Untuk memulainya, ID3 atau Iterative Dichotomiser 3 adalah algoritma *decision tree* sederhana yang diperkenalkan pada tahun 1986 oleh Quinlan Ross. Algoritma ID3 memiliki tujuan untuk membangun *decision tree* dengan menggunakan *top-down*, diikuti dengan mode pencarian serakah melalui set tes yang diberikan untuk atribut di setiap *node* pohon. Terdapat juga algoritma C4.5, perpanjangan dari algoritma ID3 Quinlan sebelumnya. Algoritma C4.5 digunakan untuk menghasilkan *decision tree*, menerapkan data kategorikal dan numerik untuk kebutuhan klasifikasi [6]. Terakhir, algoritma C5.0 merupakan perbaikan dari algoritma C4.5 yang menghasilkan pengklasifikasi yang terbukti sebagai *decision tree* atau kumpulan aturan dengan fitur yang ditingkatkan.

Terdapat beberapa penggunaan algoritma *decision tree*. *Decision tree* dapat membagi beberapa *dataset* ke dalam kelas-kelas yang terpisah. *Decision Tree* digunakan untuk semua jenis variabel target; Namun, ini terutama digunakan untuk yang berbentuk kategoris. *Decision Tree* melibatkan penggunaan metode *information gain* untuk melakukan pemisahan data dan metode penghitungan Entropy atau Gini Index dalam pemisahan data untuk menemukan homogenitas dalam kumpulan data [7]. Selain itu, *Decision Tree* dapat memprediksi hasil untuk laporan di masa mendatang. Akibatnya, *decision tree* adalah salah satu metode yang paling efektif untuk *data mining*.

Algoritma *Decision tree* memiliki kelebihan dalam mengklasifikasikan *record* yang tidak diketahui secara cepat. *Decision tree* sangat bagus dikarenakan adanya atribut yang berlebihan dan sedikit tegas dengan adanya *noise* jika metode *overfitting* diberikan. Namun, *Decision tree* memiliki kelemahan yaitu data yang tidak dapat diterapkan memiliki efek yang buruk dalam pembangunan *Decision Tree*. Pada kesempatan ini, setiap perubahan kecil pada data dapat mempengaruhi keseluruhan tampilan *decision tree*. Selain itu, sebuah sub-pohon dalam *decision tree* dapat diproduksi dalam jumlah yang banyak [8].

Berikut terdapat beberapa rumusan dari *Decision Tree*, antara lain:

a. Information Gain

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$$

Equation 1 Equation of Information Gain

where,

f = feature split on

D_p = dataset of the parent node

D_{left} = dataset of the left child node

D_{right} = dataset of the right child node

I = impurity criterion (Gini Index or Entropy)

N = total number of samples

N_{left} = number of samples at the left child node

N_{right} = number of samples at right child node

b. *Decision Tree with Gini Index*

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

Equation 2 Equation of Decision Tree with Gini Index

where,

p_j = proportion of the samples that belongs to class c for a particular node

c. *Deicision Tree with Entropy*

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

Equation 3 Equation of Decision Tree with Entropy

where,

p_j = proportion of the samples that belongs to class c for a particular node

*Ini adalah definisi entropi untuk semua kelas yang tidak kosong ($p \neq 0$). Entropi adalah 0 jika semua sampel pada node termasuk dalam kelas yang sama

2. *Random Forest*

Random forest adalah salah satu metode *ensemble* yang menggunakan seperangkat algoritma CART untuk prediksi. *Random forest* melibatkan pembangunan subset data *training* melalui proses penggantian dengan melakukan metode "*Bagging*". Dalam hal ini, data dengan nilai yang sama dapat dipilih selama beberapa kali, sedangkan data lain tidak dapat dipilih sama sekali [9].

Random forest dirumuskan oleh Leo Breiman dan Adele Cutler sedangkan *Random Forest* telah dinyatakan sebagai ciri khas mereka. Istilah algoritma *random forest* pertama kali dibentuk oleh Tin Kam Ho dari Bell Labs pada tahun 1995 yang disatukan dengan ide metode "*Bagging*" Breiman dan fitur seleksi acak yang dipresentasikan secara mandiri oleh Ho dengan bantuan dari Amit dan German. Hasilnya, teknik *random forest* mampu membangun sekelompok *decision tree* dengan beberapa perubahan yang terkontrol dengan baik [10].

Sebagai perbandingan, terdapat beberapa penggunaan algoritma *random forest*. *Random forest* memiliki metode "*Bagging*" yang lebih digeneralisasikan dan memiliki kemampuan untuk mengintegrasikan beragam tipe fitur. *Random forest* memiliki pendekatan *voting* mayoritas yang dapat meminimalkan adanya kesalahan klasifikasi secara efektif [11]. *Random forest* mampu menentukan peringkat signifikansi terkait untuk setiap prediktor yang didasarkan pada kesalahan prediksi regresi pada *Out-Of-Bag* atau OOB [12].

Algoritma *Random Forest* memiliki beberapa keuntungan dalam menangkap interaksi yang rumit antara fitur yang digunakan dengan memiliki pengetahuan tentang kombinasi nonlinier dari mereka. *Random forest* mampu beroperasi secara merata baik dengan nilai kontinu, diskrit, atau nilai hilang dengan sedikit atau tanpa modifikasi. Random Forest relatif kuat untuk *outlier*, seperti contoh *noise* dan label *noise*. Namun demikian, *Random Forest* memiliki kelemahan di mana skema yang dapat ditindaklanjuti secara tidak langsung terbukti dengan memberikan contoh. Akibatnya, hasil model secara tidak langsung dapat diinterpretasikan oleh manusia [13].

Berikut terdapat beberapa rumusan dari *Random Forest* antara lain:

a. *Gini Index: Gini(T)*

$$\boldsymbol{Gini(T) = 1 - \sum_{j=1}^{n} (p_j)^2}$$

Equation 4 Equation of Gini Index: Gini(T)

where,

p_j = relative frequency of class j in T

*Digunakan jika suatu *dataset* T berisi contoh dari n kelas

b. *Gini Index: $Gini_{split}(T)$*

$$\boldsymbol{Gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)}$$

Equation 5 Equation of Gini Index: Ginisplit(T)

Nilai atribut yang menyediakan SPLIT Gini (T) terkecil dipilih untuk membagi *node*

*Digunakan jika suatu *dataset* T dibagi menjadi dua subset T_1 dan T_2 dengan ukuran masing-masing, yaitu N_1 dan N_2, dan *gini index* dari *split* data tersebut berisi contoh dari n kelas

3. *Logistic Regression*

Logistic regression merupakan algoritma yang ditujukan untuk menyelesaikan klasifikasi biner [14]. Adapun variabel respons terdiri dari 2 hasil biner, diwakili oleh 0 atau 1. *Logistic regression* juga dapat didefinisikan sebagai teknik statistik untuk mengevaluasi hubungan antara variabel (baik variabel kategoris atau kontinu) dan hasil yang biner [15].

Fungsi *logistic* telah diciptakan pada abad ke-19 yang ditujukan untuk mendeskripsikan perkembangan populasi dan kursus reaksi kimia autokatalitik. Sebagai tambahan, fungsi *logistic* telah ditemukan oleh Raymond Pearl dan Lowell J. Reed pada tahun 1920 dalam studinya mengenai perkembangan populasi di *United States*. Namun, adapun penemuan awal dari Pierre-François Verhulst sebagai bentuk seugesti dalam 3 jenis karyanya. Adapun karya keduanya yang berjudul *Proceedings of the Belgian Royal Academy of*

1845 yang menyajikan sebuah *curve* yang dinamakan *logistic* yang di dalam diagram disebut *courbe logistique* telah digambarkan berdampingan dengan *courbe logarithmique* atau sekarang disebut dengan *exponential*. Dalam hal ini, Pearl dan Reed telah mengaplikasikan *curve* untuk perkembangan populasi tersebut sebagai bentuk penemuan kembali terhadap *logistic* dari Verhulst [16].

Adapun penggunaan *logistic regression* dalam memprediksi variabel dependen dengan dua atau lebih kategori. Dalam hal ini, *logistic regression* dapat digunakan untuk variabel dependen yang bersifat *dichotomous*. *Logistic regression* ditujukan untuk memperkenalkan proses analisis *logistic regression* yang bersifat biner yang mampu menggunakan data nyata. Dengan kata lain, *logistic regression* akan bermanfaat untuk digunakan untuk penelitian lebih lanjut dalam berbagai bidang di kehidupan nyata [17].

Model *logistic regression* memiliki suatu keunggulan dimana model ini lebih mudah diimplementasikan dan ditafsirkan. *Logistic regression* lebih sederhana daripada model lainnya dan memberikan akurasi tinggi untuk data sederhana. Adapun beberapa kelemahan yang dimiliki oleh *logistic regression*. Pertama, validitas model regresi tergantung pada jumlah dan kesesuaian variabel predictor independen yang diukur. Kedua, variabel harus memiliki besaran asosiasi yang konstan di seluruh rentang nilai untuk variabel tersebut. Ketiga, terdaoat banyak analisis *logistic regression* yang mengasumsikan bahwa efek dari satu prediktor tidak dipengaruhi oleh nilai prediktor lain. Interaksi semacam itu perlu secara eksplisit dimasukkan dalam analisis untuk memastikan perkiraan asosiasi yang valid [18].

Berikut terdapat beberapa rumusan dari *Logistic Regression*, antara lain:

a. Sigmoid Function as an Activation Function for Logistic Regression

$$f(x) = \frac{1}{1 + e^{-x}}$$

Equation 6 Equation of Sigmoid Function

where,

e = base of natural logarithms

value = numerical value one wishes to transform

b. Logistic Regression

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Equation 7 Equation of Logistic Regression

where,

x = input value

y = predicted output

b_0 = bias or intercept term

b_1 = coefficient for input (x)

4. *Gradient Boosting*

Gradient boosting merupakan istilah yang berasal dari 2 kata, yaitu *gradient* mengacu pada kesalahan atau residu, yang diperoleh setelah membangun suatu model. Sedangkan, *boosting* berarti meningkatkan. Ole karena itu, teknik ini dikenal sebagai *Gradient boosting Machine* (GBM). Dalam *gradient boosting*, setiap *decision tree* memprediksi kesalahan dari *decision tree* sebelumnya yang pada akhirnya dapat meningkatkan tingkat kesalahan atau *gradient*. Dengan demikian, *gradient boosting* adalah cara untuk secara bertahap meningkatkan yang bertujuan untuk mengurangi kesalahan [19].

Gradient boosting merupakan metode yang diambil dari AdaBoost (*Adaptive Boosting*) yang ditemukan oleh Freund dan Schapire pada tahun 1996. Selanjutnya, adapun jenis algoritma *boosting* dari Leo Breiman yang bernama "Arcing" dimana dia melakukan formalisasi atas AdaBoost sebagai optimalisasi penurunan *gradient* dalam ruang fungsi dimana gradien diperkirakan oleh prosedr nonparametric, seperti *CART tree*. Pada akhirnya, hasil penemuan Leo Breiman telah digeneralisasi oleh H. Jerome Friedman menjadi *gradient boosting* [20].

Teknik *machine learning* telah diterapkan pada data historis untuk membangun suatu model prediktif untuk memperkirakan dan merekomendasikan peristiwa pada waktu yang mendatang. Dalam hal ini, *gradient boosting* sebagai bagian dari *boosting* adalah suatu teknik *machine learning* ansambel berrupa gabungan dari beberapa model akurasi rendah guna membuat model akurasi tinggi. *Gradient boosting* dapat digunakan di berbagai bidang untuk meningkatkan prediksi, seperti kredit, asuransi, perilaku konsumen, diagnosis medis, penjualan, dan lain sebagainya [21].

Gradient boosting memiliki beberapa kelebihan, seperti mampu melatih model prediktif seperti metode *boosting* lainnya, seperti *random forest*. *Gradient boosting* mampu melakukan generalisasi pada model berdasarkan optimalisasi atas fungsi kerugian yang dapat dibedakan secara sewenang-wenang (*arbitrary differentiable loss function*). *Gradient boosting* lebih baik dalam menghindari kemungkinan *overfitting* yang biasanya terjadi pada *decision tree*. Hal ini dikarenakan *gradient boosting* memiliki kekuatan superior karena kemungkinan untuk dipengaruhi dengan skala set pelarihan lebih kecil, serta *outliers* dan fitur yang tidak relevan tersebut tidak dapat mengubah kinerja algoritma tersebut dengan mudah. Namun, *gradient boosting* melibatkan setiap putaran pelatihannya dapat mengurangi sisa dari putaran pelatihan sebelumnya, seperti melatih model baru dalam arah gradien untuk menurunkan nilai residual sebelumnya [22].

Berikut terdapat beberapa langkah serta rumusan dari *Gradient Boosting*, antara lain:

a. Fit a decision tree on data

$$[x = input, y = output]$$

Equation 8 Step 1 Gradient Boosting

b. Calculate error residuals by subtracting predicted target value from actual target value

$$[e_1 = y_{true} - y_{predicted1}]$$

Equation 9 Step 2 Gradient Boosting

c. Fit a new model on the error residuals as the target variables keeping the input variables same

$$[e_{predicted1}]$$

Equation 10 Step 3 Gradient Boosting

d. Add the predicted residuals to previous predictions

$$[y_{predicted2} = y_{predicted1} + e_{predicted1}]$$

Equation 11 Step 4 Gradient Boosting

e. Fit the next model on the remaining residuals

$$[e_2 = y_{true} + y_{predicted2}]$$

Equation 12 Step 5 Gradient Boosting

*Ulangi *Step* 2 hingga *Step* 5 hingga model *gradient boosting* mulai mengalami *overfitting* atau tidak mengalami perubahan dalam jumlah residual

B. Metode Penelitian

Adapun metode penelitian yang ditetapkan berupa metode penelitian kuantitatif yang melibatkan penggunaan angka dan akurasi [23]. Penelitian ini akan menjawab permasalahan dengan proses analisis data yang bersifat kuantitatif, yaitu mencari distribusi data dan kualitas hubungan antar variabel.

Data yang digunakan dalam penelitian ini berbentuk data kuantitatif, yaitu data nasabah EverBank. Dalam proses pengumpulan data, penulis menggunakan data yang bersumber sekunder dari situs web penghimpun data, yaitu Kaggle (https://www.kaggle.com/).

Data yang didapatkan adalah data yang bersifat valid dan cukup dikarenakan data tersebut terdiri atas data nasabah EverBank secara terperinci dan lengkap. Data tersusun cukup kompleks dan bervolume besar. Selain itu, penelitian kami menggunakan SAS Studio dan SAS Visual Analytics untuk melakukan pengolahan data dari Microsoft Excel, visualisasi data, serta pembuatan model yang dibutuhkan dengan menampilkan sejumlah grafik dari variabel yang tersedia.

Secara umum, penelitian ini akan diterapkan sesuai dengan alur penelitian yang diilustrasikan pada *figure* 1.

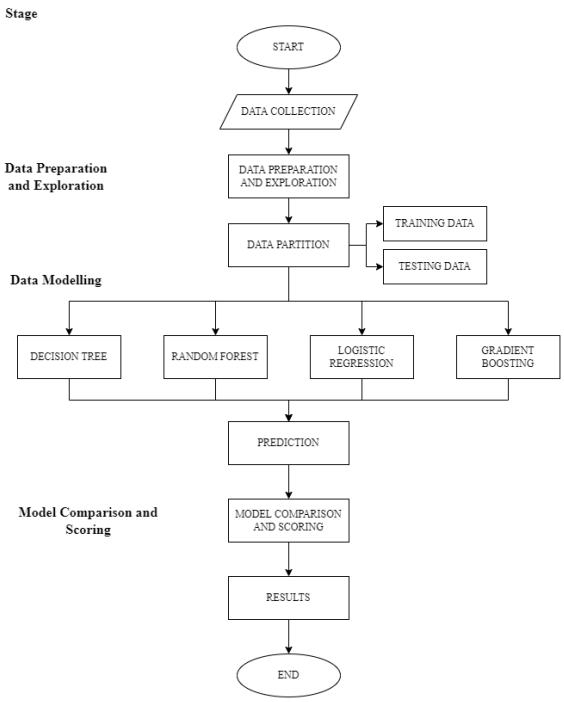


Figure 1 Framework Penelitian dengan Metode *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*

Adapun beberapa tahapan yang dilakukan EverBank dengan tujuan untuk pembuatan model prediksi pada hasil keputusan nasabah atas kelanjutan berlangganan deposito berjangka panjang dengan metode *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*, yaitu:

1) Melakukan pengumpulan data atas nasabah EverBank dari hasil kampanye. Adapun kegiatan lain berupa pembentukan latar belakang serta tujuan dari adanya penelitian ini.
2) Melakukan eksplorasi data analisis (*Exploratory Data Analytics*) dengan menampilkan struktur data, serta statistika secara sederhana, seperti nilai mean, median, min, max, dan lain sebagainya.
3) Melakukan visualisasi data sesuai variabel yang digunakan yang dilakukan secara terpisah.
4) Melakukan pemisahan data (*data splitting or data partitioning*) menjadi 2 jenis data, yaitu *testing data* dan *training data*.
5) Melakukan pembuatan berbagai model (*model building*) dengan menggunakan *training data*, yaitu model *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*.
6) Melakukan pemilihan dan penilaian model (*model selection and model scoring*) atas hasil model yang telah dibuat.
7) Melakukan penganalisisan model lebih lanjut serta membuat suatu kesimpulan dari penelitian yang terbukti oleh hasil olahan data dan model yang dimiliki.

C. Objek Penelitian

Penelitian ini menggunakan hasil keputusan nasabah EverBank sebagai objek penelitian. Adapun berbagai informasi yang berbeda pada setiap nasabahnya yang dapat dilihat dari usia, status perkawinan, tingkat pendidikan, dan lain sebagainya. Selain itu, adapun berbagai jenis pinjaman dan hasil kontak yang diperoleh setiap nasabah selama kampanye.

D. Pemahaman Data

Proses *data understanding* atau pemahaman data merupakan suatu proses dimana perusahaan menyelidiki data yang akan dipakai untuk tahap atau penelitian selanjutnya. Dalam hal ini, *dataset* EverBank berisi 20 atribut dengan 41.118 observasi yang muncul pada dataset ini, dimana variabel dependennya adalah "y" dengan nilai "ya" atau "tidak" yang menentukan apakah nasabah akan berlangganan deposito berjangka panjang. Sedangkan, variabel lainnya dapat dianggap sebagai variabel independen. Berikut terdapat rincian mengenai dataset yang digunakan, antara lain sebagai berikut:

Nama Variabel	Deskripsi Variabel	Tipe Variabel	Nilai Valid
Data Nasabah Bank			
id	ID Nasabah	Numerik	Numerikal
age	Usia	Numerik	Numerikal
job	Jenis Pekerjaan	Kategorikal	*Admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown*
marital	Status Perkawinan	Kategorikal	*Married, divorced, single, unknown*
education	Pendidikan	Kategorikal	*Basic.4y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown*
default	Pinjaman secara *default*	Kategorikal	*Yes, no, unknown*
housing	Pinjaman Rumah	Kategorikal	*Yes, no, unknown*
loan	Pinjaman Pribadi	Kategorikal	*Yes, no, unknown*
Atribut yang Berkaitan dengan Kontak dalam Kampanye			
contact	Tipe komunikasi	Kategorikal	*Cellular, telephone*
month	Kontak Terakhir dalam bentuk Bulan	Kategorikal	*January, February, March to December*
day	Kontak Terakhir dalam bentuk Harian	Kategorikal	*Monday, Tuesday, Wednesday, Thursday, Friday*
Atribut Kampanye Lainnya			
campaign	Kontak selama Kampanye	Numerik	Numerikal
pdays	Jumlah Hari yang Berlalu setelah Kontak Terakhir	Numerik	Numerikal
previous	Jumlah Kontak sebelum Kampanye	Numerik	Numerikal
poutcome	Hasil Kampanye Sebelumnya	Kategorikal	*Success, failure, nonexistent*
Atribut dengan Konteks Sosial dan Ekonomi			
emp.var.rate	Tingkat Variasi Pekerjaan per Kuartal	Numerik	Numerikal
cons.conf.idx	Indeks Keyakinan Konsumen per Bulan	Numerik	Numerikal
euribor3m	Tarif Euribor 3 Bulan per Harian (Hari)	Numerik	Numerikal

nr.employed	Jumlah Karyawan per Kuartal	Numerik	Numerikal
Variabel Output			
y	Nasabah telah Berlangganan Deposit Berjangka Panjang	*Binary*	*Yes, no*

Table 1 Tabel *Data Understanding*

III. PERSIAPAN & EKSPLORASI DATA

A. Persiapan Data

Setelah memahami *dataset* EverBank, adapun proses *data preparation* dimana data akan disiapkan terlebih dahulu sebelum dibawa menuju tahap pembuatan model. Terdapat juga *id* yang merupakan ID pada nasabah yang bersifat unik sehingga tidak terdapat nilai duplikat pada dataset EverBank.

Pada tahap ini akan dilakukan proses *standardize data* pada 11 kolom seperti kolom *job*, *marital status*, *education*, *default*, *housing*, *loan*, *contact*, *month*, *day_of_week*, *poutcome*, dan *y* guna memperoleh hasil nilai data yang konsisten. Adapun rincian proses *standardize data* dalam bentuk tabel, sebagai berikut:

No	Nama Kolom	Definisi Kolom
1	*job*	*Business Title*
2	*marital status*	*Address*
3	*education*	*Organization*
4	*default*	*Address*
5	*housing*	*Address*
6	*loan*	*Address*
7	*contact*	*Name*
8	*month*	*Name*
9	*day_of_week*	*Name*
10	*poutcome*	*Address*
11	*y*	*Address*

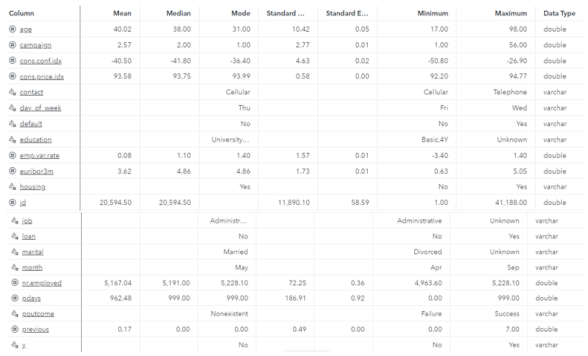
Table 2 Tabel Rincian *Standardize* Data EverBank

Oleh karena itu, *data preparation* akan membantu dalam memberikan jawaban yang benar terhadap suatu permasalahan bisnis yang dapat sekaligus meningkatkan tingkat akurasi dalam membuat keputusan bisnis. Hasil tabel dari proses *standardize data* adalah sebagai berikut:

Figure 2 Tabel Struktur Data *EverBank*

B. Eksplorasi Data

Eksplorasi data merupakan tahap yang menampilkan deskriptif statistik, seperti nilai *mean*, *median*, *mode*, *standard deviation* (SD), *standard error*, *minimum*, *maximum*, dan tipe data. Berikut deskriptif statistik pada data EverBank yang disajikan pada *figure* 3.



Column	Mean	Median	Mode	Standard ...	Standard E...	Minimum	Maximum	Data Type
age	40.02	38.00	31.00	10.42	0.05	17.00	98.00	double
campaign	2.57	2.00	1.00	2.77	0.01	1.00	56.00	double
cons.conf.idx	-40.50	-41.80	-36.40	4.63	0.02	-50.80	-26.90	double
cons.price.idx	93.58	93.75	93.99	0.58	0.00	92.20	94.77	double
contact			Cellular			Cellular	Telephone	varchar
day_of_week			Thu			Fri	Wed	varchar
default			No			No	Yes	varchar
education			University...			Basic.4Y	Unknown	varchar
emp.var.rate	0.08	1.10	1.40	1.57	0.01	-3.40	1.40	double
euribor3m	3.62	4.86	4.86	1.73	0.01	0.63	5.05	double
housing			Yes			No	Yes	varchar
id	20,594.50	20,594.50		11,890.10	58.59	1.00	41,188.00	double
job			Administr...			Administrative	Unknown	varchar
loan			No			No	Yes	varchar
marital			Married			Divorced	Unknown	varchar
month			May			Apr	Sep	varchar
nr.employed	5,167.04	5,191.00	5,228.10	72.25	0.36	4,963.60	5,228.10	double
pdays	962.48	999.00	999.00	186.91	0.92	0.00	999.00	double
poutcome			Nonexistent			Failure	Success	varchar
previous	0.17	0.00	0.00	0.49	0.00	0.00	7.00	double
y			No			No	Yes	varchar

Figure 3 Deskriptif Statistik pada Data EverBank

Adapun tampilan berbagai variabel dalam data EverBank yang divisualisasikan secara terpisah dalam bentuk *bar plot* guna mengetahui frekuensi nilai pada setiap datanya.

1. EverBank Customer Data

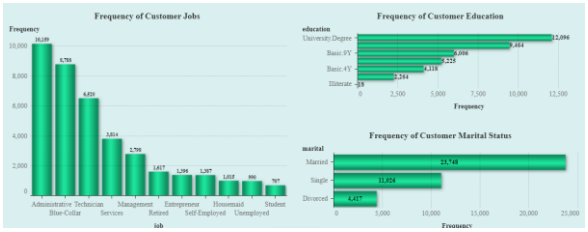


Figure 4 EverBank Customer Data

Terdapat *figure 4* menunjukkan visualisasi *bar plot* pada data nasabah (*Customer Data*) EverBank. Pada visualisasi ini, terdapat 3 pembagian *bar plot* yang menunjukkan berbagai jenis data nasabah EverBank.

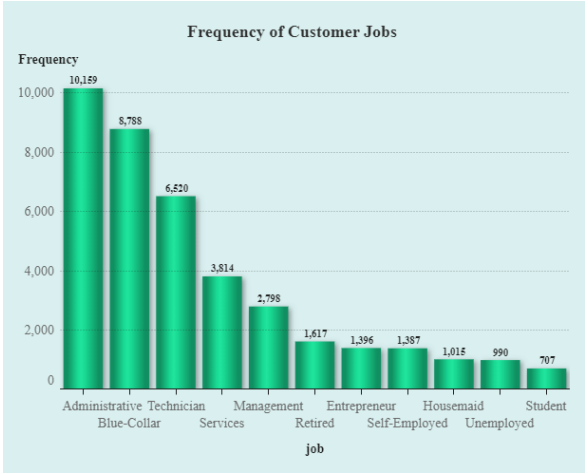


Figure 5 Frequency of Customer Jobs

Figure 5 menunjukkan *frequency of customer jobs*, yaitu frekuensi pekerjaan pada nasabah EverBank.

Berdasarkan *bar plot* di atas, frekuensi tertinggi berada pada nasabah yang bekerja sebagai *administrative* atau *admin* sebesar 10,159 orang. Sedangkan, frekuensi terendah berada pada nasabah yang memiliki pekerjaan sebagai *student* sebesar 707 orang.

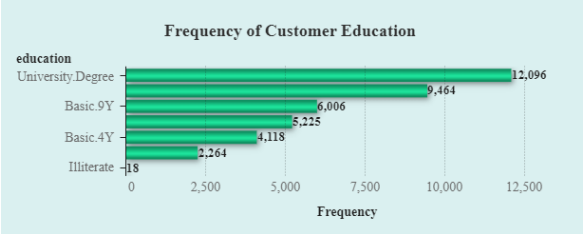


Figure 6 Frequency of Customer Education

Figure 6 menunjukkan *frequency of customer education*, yaitu frekuensi pendidikan yang dimiliki nasabah EverBank. Berdasarkan *bar plot* di atas, frekuensi tertinggi berada pada nasabah yang memiliki pendidikan terakhir di universitas sebesar 12,096 orang. Sedangkan, frekuensi terendah berada pada nasabah yang tidak memiliki pendidikan atau *illiterate* (buta huruf) sebesar 18 orang.

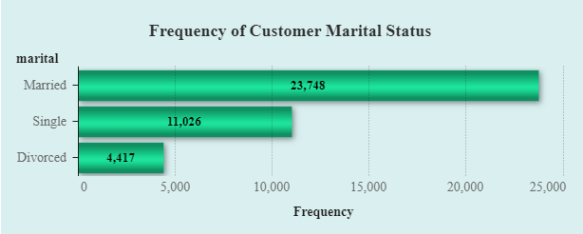


Figure 7 Frequency of Customer Marital Status

Figure 7 menunjukkan *frequency of customer marital status*, yaitu frekuensi status perkawinan yang dimiliki nasabah EverBank. Berdasarkan *bar plot* di atas, frekuensi tertinggi berada pada nasabah yang berstatus *married* (kawin) sebesar 22,748 orang. Sedangkan, frekuensi terendah berada pada nasabah yang berstatus *divorced* (cerai) sebesar 4,417 orang.

2. EverBank Customer Loan

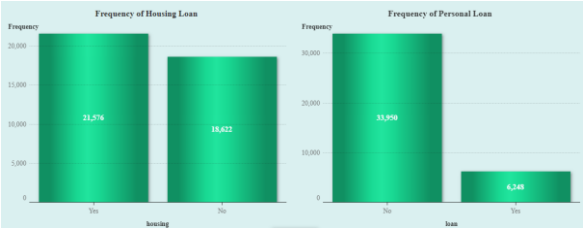


Figure 8 EverBank Customer Loan

Terdapat *figure 8* menunjukkan visualisasi *bar plot* pada data pinjaman nasabah (*Customer Loan*) EverBank. Pada visualisasi ini, terdapat 2 pembagian *bar plot* yang menunjukkan berbagai jenis data pinjman nasabah EverBank.

Figure 9 Frequency of Housing Loan

Figure 9 menunjukkan *frequency of housing loan*, yaitu frekuensi pinjaman untuk kebutuhan *housing* (perumahan) yang dimiliki nasabah EverBank. Berdasarkan *bar plot* di atas, nasabah yang memiliki pinjaman *housing* memiliki frekuensi tertinggi sebesar 21,576 orang. Sedangkan, nasabah yang tidak memiliki pinjaman *housing* memiliki frekuensi terendah sebesar 18,622 orang.

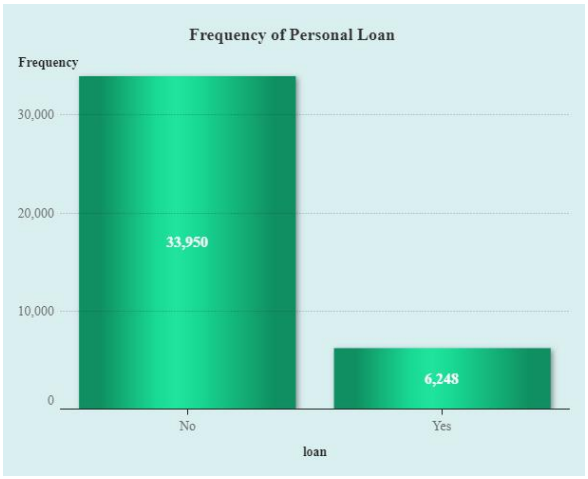


Figure 10 Frequency of Personal Loan

Figure 10 menunjukkan *frequency of personal loan*, yaitu frekuensi pinjaman untuk kebutuhan *personal* (pribadi) yang dimiliki nasabah EverBank. Berdasarkan *bar plot* di atas, nasabah yang tidak memiliki pinjaman *personal* memiliki frekuensi tertinggi sebesar 33,950 orang. Sedangkan, nasabah yang memiliki pinjaman *personal* memiliki frekuensi terendah sebesar 6,248 orang.

3. EverBank Previous Campaign Details

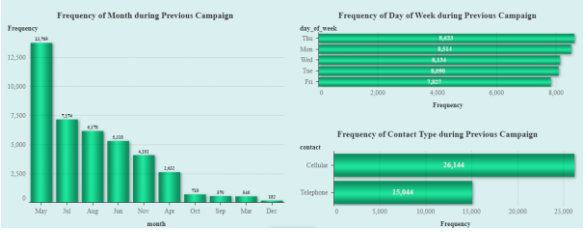


Figure 11 EverBank Previous Campaign Details

Terdapat *figure 11* menunjukkan visualisasi *bar plot* pada data kampanye, terutama durasi kampanye (*Campaign Duration*) EverBank. Pada visualisasi ini, terdapat 3 pembagian *bar plot*, yang disertai dengan 1 *bar plot* tambahan yang menunjukkan berbagai jenis data kampanye.

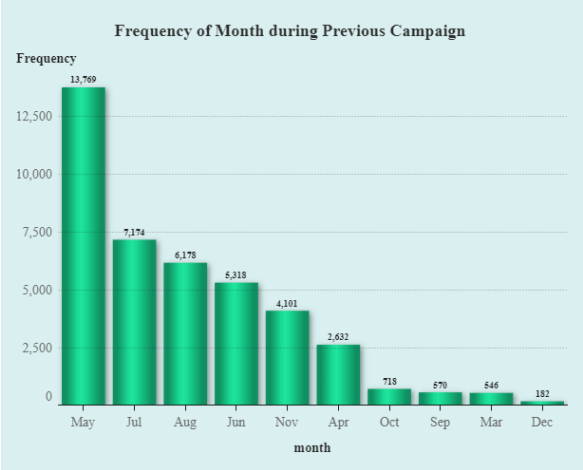


Figure 12 Frequency of Month during Previous Campaign

Figure 12 menunjukkan *frequency of month during campaign*, yaitu frekuensi kampanye sebelumnya yang diadakan dalam bentuk bulanan yang dimulai dari bulan bulan *Mar* (Maret) hingga *Dec* (Desember). Berdasarkan *bar plot* di atas, kampanye sebelumnya yang diadakan pada bulan *May* (Mei) memiliki frekuensi tertinggi yang ditujukan kepada 13,769 orang. Sedangkan, frekuensi terendah berada pada bulan *Dec* (Desember) dimana kampanye sebelumnya hanya diadakan untuk 182 orang.

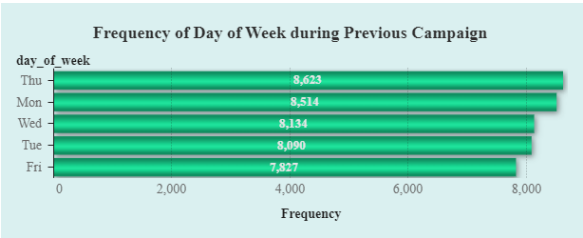


Figure 13 Frequency of Day of Week during Previous Campaign

Figure 13 menunjukkan *frequency of day of week during campaign*, yaitu frekuensi kampanye sebelumnya yang diadakan dalam bentuk harian yang dimulai pada hari *Mon* (Senin) hingga *Fri* (Jumat). Berdasarkan *bar plot* di atas, kampanye sebelumnya yang diadakan pada hari *Thu* (Kamis) memiliki frekuensi tertinggi yang ditujukan kepada 8,623 orang. Sedangkan, frekuensi terendah berada pada hari *Fri* (Jumat) dimana kampanye sebelumnya hanya diadakan untuk 7,827 orang.

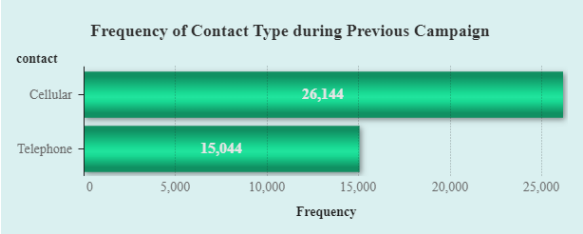


Figure 14 Frequency of Contact Type during Previous Campaign

Figure 14 menunjukkan *frequency of contact type during campaign*, yaitu frekuensi tipe kontak yang digunakan nasabah selama kampanye EverBank. Berdasarkan *bar plot* di atas, nasabah yang memakai *cellular* selama kampanye sebelumnya memiliki frekuensi tertinggi sebesar 26,144 orang. Sedangkan, nasabah yang memakai *telephone* selama kampanye sebelumnya memiliki frekuensi terendah sebesar 15,044 orang.

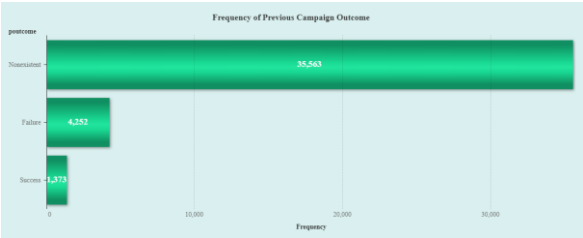


Figure 15 Frequency of Previous Campaign Outcome

Figure 15 menunjukkan *frequency of campaign outcome*, yaitu frekuensi hasil kampanye sebelumnya yang diadakan oleh EverBank. Berdasarkan *bar plot* di atas, hasil *nonexistent* memiliki frekuensi tertinggi sebesar 35,563 orang, dimana hal ini berarti sebanyak 35,563 nasabah tidak dapat dijangkau selama kampanye berlangsung. Sedangkan, nasabah yang sukses mengikuti kampanye sebelumnya memiliki frekuensi terendah sebesar 1,373 orang.

4. EverBank Customer Subscriptions

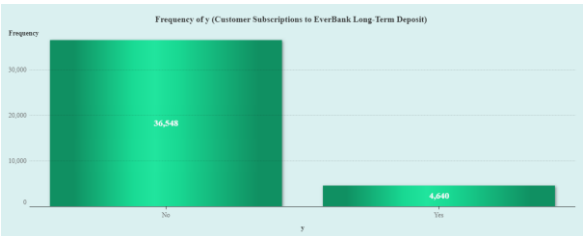


Figure 16 EverBank Customer Subscriptions

Figure 16 menunjukkan *frequency of y (customer subscriptions to EverBank Long Term Deposit)*, yaitu frekuensi nasabah berlangganan deposito berjangka panjang EverBank. Berdasarkan *bar plot* di atas, nasabah yang tidak berlangganan deposito berjangka panjang EverBank memiliki frekuensi tertinggi sebesar

36,548 orang. Sedangkan, nasabah yang berlangganan deposito berjangka panjang EverBank memiliki frekuensi terendah sebesar 4,640 orang.

Adapun pembagian data (*data partition*) dimana keseluruhan jumlah *dataset* EverBank akan dibagi menjadi 2. Selain itu, terdapat proporsi pada setiap *partitionnya* sebesar 70% untuk data *training* yang digunakan dalam pembuatan model data dan 30% untuk data *validation* yang akan digunakan dalam perbandingan model *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*. Adapun rincian tentang *data partition* pada *dataset* EverBank dalam bentuk tabel, yaitu sebagai berikut.

No	*Data Partition Percentage*	*Data Usage*
1	70%	*Training Data*
2	30%	*Validation Data*

Table 3 Table Data Partition Details

IV. PEMODELAN DATA

Setelah melalui proses persiapan data dan eksplorasi data, terdapat pembuatan model data dengan menggunakan data *training*. Pemodelan data akan dibuat berdasarkan model yang dipilih, yaitu model *Decision Tree*, *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*.

Selanjutnya, adapun proses perbandingan model (*model comparison*) yang dihasilkan melalui *confusion matrix* dengan data *validation*. Dalam hal ini, perbandingan dan penilaian model tersebut akan dilihat berdasarkan tingkat misklasifikasi (*misclassification rate*), dimana semakin kecil tingkat misklasifikasi, maka semakin baik model prediksi yang dimiliki. Adapun perhitungan akurasi yang didapatkan dari *misclassification rate*, yaitu sebagai berikut.

$$Accuracy = 1 - \text{Misclassification Rate}$$

Equation 13 Equation of Accuracy Rate

Pada akhirnya, model dengan tingkat misklasifikasi terendah dapat digunakan untuk memprediksi hasil keputusan nasabah EverBank terhadap berlangganan deposito berjangka panjang.

Pemodelan data ini melibatkan variabel *y* sebagai variabel respons, sedangkan variabel yang dijadikan variabel independent adalah variabel *job*, *education*, *default*, *housing*, *personal*, *day_of_week*, *month*, *contact*, dan *poutcome*. Berikut rincian penggunaan variabel dalam pemodelan data dalam bentuk tabel, yaitu sebagai berikut.

No	*Variable Name*	*Variable Role*
1	*y*	Respons
2	*job*	Prediktor
3	*education*	Prediktor
4	*default*	Prediktor
5	*housing*	Prediktor
6	*personal*	Prediktor
7	*day_of_week*	Prediktor
8	*month*	Prediktor
9	*contact*	Prediktor
10	*outcome*	Prediktor

Table 4 Tabel Penggunaan Variabel dalam Pemodelan Data

A. *Decision Tree Analysis*

Pertama, adapun pembuatan model *decision tree* yang dilakukan dengan membuat node beserta daunnya berdasarkan kondisi tertentu yang sesuai untuk klasifikasi.

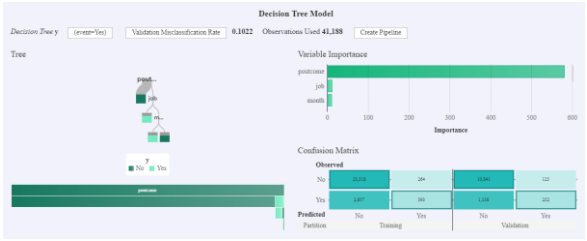


Figure 17 Decision Tree Model

Berikut terdapat analisis atas hasil pembuatan model *decision tree* berupa sebuah pohon yang berdasar pada variabel respons *y* dengan *node* dan daun yang dibuat berdasarkan kondisi tertentu. Pada *figure* 17 di atas, adapun beberapa kondisi yang memengaruhi *decision tree* tersebut, seperti variabel *poutcome*, *job*, dan *month.*

Adapun *variable importance plot* yang menunjukkan seberapa penting variabel tersebut terhadap prediksi model tersebut. Dalam hal ini, variabel *poutcome* memiliki tingkat *importance* yang tinggi terhadap model *decision tree* dari variabel *y*.

Terdapat *confusion matrix* yang dimiliki oleh model *decision tree*. Dari visualisasi *confusion matrix* tersebut, maka dapat dilihat bahwa nilai *actual* dari variabel 'No' dan 'Yes' positif dan diprediksi positif lebih banyak dibandingkan nilai *actual* 'No' dan 'Yes' positif dan diprediksi negatif. Selain itu, dapat dilihat juga bahwa nilai *misclassification rate* pada model *decision tree* sebesar 0.1019. Maka, tingkat akurasi pada model *decision tree* sebesar 89.81%.

B. *Random Forest Analysis*

Kedua, terdapat pembuatan model *random forest* yang dilakukan dengan membuat hutan yang terdiri atas beberapa pohon berdasarkan klasifikasi tertentu.

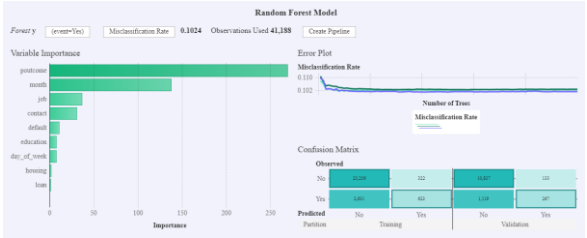


Figure 18 Random Forest Model

Berikut terdapat analisis atas hasil pembuatan model *random forest* yang dibuat berupa sebuah hutan yang berdasar pada variabel respons *y*. Pada *figure* 18 di atas, adapun *variable importance plot* yang menunjukkan seberapa penting variabel tersebut terhadap prediksi model tersebut. Apabila dilihat dari *variable importance plot*, variabel *poutcome* memiliki tingkat *importance* tertinggi, sedangkan variabel *loan* memiliki tingkat *importance* terendah terhadap model *random forest* dari variabel *y*.

Selanjutnya, adapun *error plot* yang menunjukkan *misclassification rate* pada model *random forest. Error plot* ini memerhatikan nilai *Out-of-Bag* (OOB) yang digunakan untuk memvalidasi model *random forest*. Adapun garis *error* dan nilai OOB *error* pada setiap hutan di *error plot*, dimana semakin rendah garis *error* dan nilai OOB *error* yang didapatkan, maka semakin baik model *random forest* yang dimiliki. Dari *error plot* pada *figure 18* di atas, dapat dilihat bahwa garis *error* pada setiap hutannya cenderung rendah sehingga nilai OOB *error* yang didapatkan cenderung rendah.

Terdapat *confusion matrix* yang dimiliki oleh model *random forest*. Dari visualisasi *confusion matrix* tersebut, maka dapat dilihat bahwa nilai *actual* dari variabel 'No' dan 'Yes' positif dan diprediksi positif lebih banyak dibandingkan nilai *actual* 'No' dan 'Yes' positif dan diprediksi negatif. Selain itu, dapat dilihat juga bahwa nilai *misclassification rate* pada model *random forest* sebesar 0.1024. Maka, tingkat akurasi pada model *random forest* sebesar 89.76%.

C. *Logistic Regression Analysis*

Ketiga, adapun pembuatan serta analisis model *logistic regression* yang dilakukan dengan melihat hubungan antara variabel respons dan predictor, serta variabel yang bersifat biner, yaitu bernilai 0 dan 1.

Figure 19 Logistic Regression Model

Berikut terdapat analisis atas hasil pembuatan model *logistic regression* hubungan variabel respons *y* dengan variabel predictor lainnya. Pada *figure* 19 di atas, adapun *fit summary* beserta *p-value* yang digunakan, yaitu 0.05 yang dapat menentukan signifikansi antara variabel *y* dengan variabel prediktor yang digunakan. Ketika nilai variabel prediktor lebih kecil daripada nilai *p-value*, maka terdapat bukti untuk mengatakan bahwa terdapat signifikasi statistik antara variabel respons dan prediktor, dan begitupun juga sebaliknya. Dari *fit summary plot* tersebut, adapun variabel *loan* dan *housing* yang memiliki hubungan yang signifikan terhadap variabel respons *y*. Sedangkan variabel lainnya yang tidak memiliki hubungan yang signifikan dengan variabel respons *y* adalah variabel *poutcome*, *month*, *contact*, *job*, *default*, *day_of_week*, dan *education*. Hal ini dikarenakan *p-value* dari variabel-variabel prediktor tersebut lebih besar daripada nilai *p-value*.

Figure 19 juga telah menunjukkan *residual plot* yang dapat menentukan tingkat akurasi model *logistic regression* yang dapat dilihat dari nilai residual dengan nilai probabilitas yang diprediksi. Dalam hal ini, setiap titik yang terdapat pada *residual plot* merupakan satu frekuensi dimana pengukuran seberapa bagus prediksi untuk nilai tersebut dapat dilihat dari jarak nilai tersebut dari garis *baseline* yang berada pada 0. Terdapat suatu kondisi apabila nilai residual yang didapatkan bernilai positif, maka prediksi terlalu rendah, nilai negatif menandakan bahwa prediksinya terlalu tinggi, dan nilai 0 menandakan bahwa prediksi tersebut sudah sesuai. Maka dari *residual plot* yang dimiliki, maka dapat disimpulkan bahwa kebanyakan nilai residual pada model *logistic regression* berada di dekat garis *baseline* sehingga tidak terdapat perbedaan yang signifikan pada prediksi yang diperoleh.

Terdapat *confusion matrix* yang dimiliki oleh model *logistic regression*. Dari visualisasi *confusion matrix* tersebut, maka dapat dilihat bahwa nilai *actual* dari variabel 'No' dan 'Yes' positif dan diprediksi positif lebih banyak dibandingkan nilai *actual* 'No' dan 'Yes' positif dan diprediksi negatif. Selain itu, dapat dilihat juga bahwa nilai *misclassification rate* pada model *logistic regression* sebesar 0.1031. Maka, tingkat akurasi pada model *logistic regression* sebesar 89.69%.

D. *Gradient Boosting Analysis*

Terakhir, adapun pembuatan dan analisis model *gradient boosting* yang dilakukan dengan melakukan pengulangan dalam pembuatan pohon guna memperkecil tingkat kesalahan atau *error*.

Figure 20 Gradient Boosting Model

Berikut terdapat analisis atas hasil pembuatan model *gradient boosting* yang berdasar atas hasil putaran pembuatan pohon atas *data training* pada variabel respons *y*. Pada *figure* 20 di atas, adapun *variable importance plot* yang menunjukkan seberapa penting variabel tersebut terhadap prediksi model tersebut. Apabila dilihat dari *variable importance plot*, variabel *poutcome* memiliki tingkat *importance* tertinggi, sedangkan variabel *loan* memiliki tingkat *importance* terendah terhadap model *gradient boosting* dari variabel *y*.

Berikut terdapat *iteration plot* dengan *misclassification rate* yang dilakukan atas pengulangan pada putaran dalam pembuatan pohon dengan data *training* yang mampu mengurangi sisa dari putaran sebelumnya. Adapun garis *iteration* pada setiap pohon yang dibuat, dimana semakin rendah garis *iteration* dan garis residual yang didapatkan, maka semakin baik model *gradient boosting* yang dimiliki. Maka, *iteration plot* pada *figure* 20 telah menunjukkan bahwa setiap pengulangan pembuatan pohon yang melibatkan penambahan pohon baru, nilai residual akan ikut menurun sehingga nilai *error* atau tingkt kesalahan yang didapatkan cenderung kecil.

Terdapat *confusion matrix* yang dimiliki oleh model *gradient boosting*. Dari visualisasi *confusion matrix* tersebut, maka dapat dilihat bahwa nilai *actual* dari variabel 'No' dan 'Yes' positif dan diprediksi positif lebih banyak dibandingkan nilai *actual* 'No' dan 'Yes' positif dan diprediksi negatif. Selain itu, dapat dilihat juga bahwa nilai *misclassification rate* pada model *gradient boosting* sebesar 0.0980. Maka, tingkat akurasi pada model *gradient boosting* sebesar 90.2%.

E. *Model Comparisons*

Setelah melakukan pembuatan dan analisis pada model yang digunakan, telah diperoleh 4 model prediksi dengan metode yang berbeda, yaitu model *decision tree*, *random forest*, *logistic regression*, dan *gradient boosting*. Keempat model akan menggunakan pendekatan yang beragam yang menghasilkan tingkat misklasifikasi dan tingkat akurasi yang berbeda.

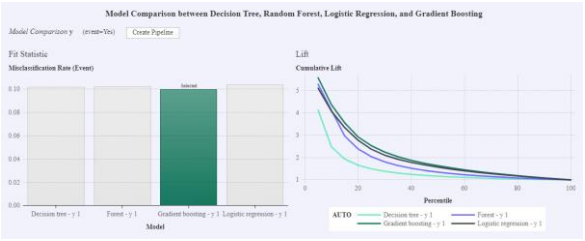


Figure 21 Model Comparisons

Berdasarkan *figure 21* di atas, terdapat penggunaan data *validation* untuk perbandingan keempat model berdasarkan tingkat misklasifikasi, diperoleh model *gradient boosting* dengan tingkat misklasifikasi terkecil, yaitu 0.0980 (90.2%). Sedangkan, model *decision tree* memegang tingkat misklasifikasi terkecil kedua, yaitu 0.1019 (89.81%), diikuti oleh model *random forest* dan *logistic regression* dengan tingkat misklasifikasi sebesar 0.1024 (89.76%) dan 0.1040 (89.69%) secara berurutan.

Selain itu, adapun metode perbandingan model lainnya yang dapat dilihat dari *lift chart* berdasarkan *cumulative lift gain* atas 4 model yang digunakan. Dalam hal ini, *lift chart* dan *cumulative lift gains* menentukan seberapa bagus model prediksi yang dimiliki dengan melihat jarak *lift curve* dengan *baseline* atau garis dasar, dimana semakin besar jarak antara *lift curve* dengan *baseline*, maka semakin bagus performa model yang didapatkan. Maka, *figure 21* telah menunjukkan bahwa model *gradient boosting* memiliki jarak *lift curve* dan *baseline* terbesar dari model lainnya. Pada akhirnya, model *gradient boosting* dapat digunakan untuk memprediksi di waktu mendatang.

V. KESIMPULAN

A. Kesimpulan

Setiap model *machine learning* akan menghasilkan tingkat akurasi yang berbeda, tergantung pada pendekatan untuk masing-masing metode. *Decision tree* menggunakan *node* untuk menggambarkan kondisi yang terbentuk, sementara *random forest* beroperasi dengan membangun banyak *decision tree* selama tahap *training* dan membawa mode kelas sebagai *output* oleh masing-masing pohon. Adapun *logistic regression* yang melibatkan penyelesaian klasifikasi biner dan evaluasi hubungan antarvariabel, sedangkan *gradient boosting* bertujuan untuk meningkatkan *gradient* yang secara bersamaan dapat mengurangi kesalahan.

Meskipun setiap model memiliki kelebihan dan kekurangannya masing-masing, penelitian ini memerlukan penggunaan *dataset* yang sama dengan tingkat akurasi yang berbeda. Untuk mendapatkan prediksi yang akurat, diperlukan model akurasi yang lebih besar untuk menghindari kesalahan dalam memprediksi. Penelitian ini memiliki suatu hasil atau *outcome* dimana nilai prediktif keputusan nasabah EverBank dalam berlangganan deposito berjangka panjang (*long-term deposit*) dengan model *gradient boosting* lebih akurat sehingga *dataset* dan variabel target yang dipilih lebih baik digunakan untuk prediksi di masa mendatang.

B. Saran

Untuk mengurangi kesalahan dalam mengadakan kampanye selanjutnya yang berdampak pada perilaku nasabah EverBank terhadap deposito berjangka panjang, diperlukan penggunaan analisis data. Dengan menggunakan *dataset* yang lengkap, hal ini dapat diamati variabel target untuk menentukan nilai keputusan nasabah EverBank dalam berlangganan deposito berjangka panjang. Dalam hal ini, model *gradient boosting* merupakan model yang tepat untuk digunakan dalam mendapatkan wawasan tentang perilaku nasabah terhadap keberlanjutan atau perpanjangan langganan deposito berjangka panjang.

C. Limitasi

Penelitian ini membutuhkan variabel target yang jelas dimana setiap kasus dapat menggunakan variabel target yang berbeda-beda. Akibatnya, berbagai jenis variabel target dapat sangat mempengaruhi model yang digunakan.

UCAPAN TERIMA KASIH

Kami selaku penulis mengucapkan terimakasih kepada Bapak Iwan Prasetiawan, S.Kom., M.M., selaku dosen pembimbing mata kuliah *Big Data Analytics*, program studi Sistem Informasi, Universitas Multimedia Nusantara karena telah bersedia meluangkan waktu untuk memberikan saran, arahan, serta bimbingan selama studi ini berlangsung.

DAFTAR PUSTAKA

[1] M. F. Aulia, A. M. Wahyu, P. G. Anugrah, T. Chusniyah and G. R. U. Hakim, "Tujuan Hidup sebagai Prediktor Kesejahteraan Psikologi pada Generasi Z," in *Seminar Nasional Psikologi dan Ilmu Humaniora (SENAPIH) 2021*, Malang, 2018.

[2] A. Murtani, "Sosialisasi Gerakan Menabung," in *Seminar Nasional Hasil Inovasi Pengabdian Kepada Masyarakat (SINDIMAS)*, Pontianak, 2019.

[3] M. R. Tambunan and I. G. S. Nasution, "Analisis Faktor-Faktor yang Mempengaruhi Keputusan Nasabah Menabung di Bank BCA Kota Medan (Studi Kasus Etnis Cina)," *Jurnal Ekonomi dan Keuangan*, vol. 1, no. 3, pp. 193-204, 2013.

[4] R. T. Rahayu, "Penerapan Strategi Pemasaran Produk Deposito BTN Ritel Rupiah di Bank Tabungan Negara Yogyakarta," Yogyakarta, 2020.

[5] S. Gavankar and S. Sawarkar , "A Novel EagerDT Complexity Approach to Deal with Missing Values in Decision," *International Journal of Simulation -- Systems, Science & Technology*, vol. 19, no. 6, pp. 1-5, December 2018.

[6] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094-2097, April 2016.

[7] P. Gulati, A. Sharma and M. Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review," *International Journal of Computer Applications*, vol. 141, no. 14, pp. 19-25, May 2016.

[8] M. Somvanshi, P. Chavan, S. Tambade and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, 2016.

[9] M. Belgiu and L. Drăguţ, "Random Forest in Remote Sensing: A Review of Applications and Future Directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24-31, 2016.

[10] Y. Li, C. Yan, W. Liu and M. Li, "A Principle Component Analysis-Based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification," *Applied Soft Computing*, vol. 70, pp. 1000-1009, September 2018.

[11] X. Mei, R. Wang, W. Yang, F. Qian, X. Ye, L. Zhu, Q. Chen, B. Han, T. Deyer, J. Zeng, X. Dong, W. Gao and W. Fang, "Predicting malignancy of pulmonary ground-glass nodules and their invasiveness by random forest," *Journal of thoracic disease*, vol. 10, no. 1, pp. 458-463, 2018.

[12] Y. Everingham, J. Sexton, D. Skocaj and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for sustainable development*, vol. 36, no. 27, 2016.

[13] M. Ibrahim, "An empirical comparison of random forest-based and other learning-to-rank algorithms," *Pattern Analysis and Applications*, vol. 23, pp. 1133-1155, 2020.

[14] C. Bonte and F. Vercauteren, "Privacy-preserving logistic regression training," *BMC Med Genomics*, vol. 11, no. 86, 2018.

[15] P. Ranganathan, C. S. Pramesh and R. Aggarwal, "Common pitfalls in statistical analysis: Logistic regression," *Perspectives in clinical research*, vol. 8, no. 3, pp. 148-151, 2017.

[16] J. S. Cramer, "The Origins of Logistic Regression," *Tinbergen Institute Working Paper*, vol. 119, no. 4, December 2002.

[17] Ö. Çokluk, "Logistic Regression: Concept and Application," *Educational Sciences: Theory and Practice*, vol. 10, no. 3, pp. 1397-1407, 2010.

[18] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA*, vol. 316, no. 5, pp. 533-534, August 2016.

[19] V. K. Ayyadevara, "Gradient Boosting Machine," *Pro Machine Learning Algorithms*, pp. 117-134, 2018.

[20] P. Bühlmann, "Remembrance of Leo Breiman," *The Annals of Applied Statistics*, vol. 4, no. 4, pp. 1638-1641, December 2010.

[21] P. Bahad and P. Saxena, "Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics," in *International Conference on Intelligent Computing and Smart Communication 2019*, Singapore, 2020.

[22] H. Cui, D. Huang, Y. Fang, L. Liu and C. Huang, "Webshell Detection Based on Random Forest-Gradient Boosting Decision Tree Algorithm," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, Guangzhou, 2018.

[23] S. Rutberg and C. Bouikidis, "Focusing on the Fundamentals: A Simplistic Differentiation Between Qualitative and Quantitative Research," *Nephrology nursing journal: journal of the American Nephrology Nurses' Association*, vol. 45, no. 2, pp. 209-212, March 2018.