PREDIKSI CALON NASABAH DEPOSITO BERBASIS CRISP-DM DAN *SMOTE XGBOOST CLASSIFICATION*

TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar

Magister Ilmu Komputer (M.Kom)

RIKI SUPRIYADI

14002371

Program Studi Ilmu Komputer (S2)

Fakultas Teknologi Informasi

Universitas Nusa Mandiri

2021

PREDIKSI CALON NASABAH DEPOSITO BERBASIS CRISP-DM DAN *SMOTE XGBOOST CLASSIFICATION*

TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar

Magister Ilmu Komputer (M.Kom)

RIKI SUPRIYADI

14002371

Program Studi Ilmu Komputer (S2)

Fakultas Teknologi Informasi

Universitas Nusa Mandiri

2021

SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME

Yang bertanda tangan di bawah ini:

Nama : Riki Supriyadi
NIM : 14002371
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : Data Mining

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: "Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*" adalah hasil karya sendiri, dan semua sumber baik yang kutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri dicabut/dibatalkan.

Jakarta, 26 Juli 2021
Yang menyatakan,

Riki Supriyadi

HALAMAN PERSETUJUAN DAN PENGESAHAN TESIS

Tesis ini diajukan oleh:

Nama	:	Riki Supriyadi
NIM	:	14002371
Program Studi	:	Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Data Mining*
Judul Tesis	:	Prediksi Calon Nasabah Deposito Berbasis CRISP-DM Dan SMOTE XGBoost Classification

Telah dipertahankan pada periode 2021-1 dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri.

Jakarta, 19 Agustus 2021

PEMBIMBING TESIS

Pembimbing I	:	Dr. Didi Rosiyadi, M.Kom	
Pembimbing II	:	Eni Heni Hermaliani, M.M, M.Kom	

D E W A N P E N G U J I

Penguji I	:	Dr. Hilman Ferdinandus Pardede, S.T, M.EICT	
Penguji II	:	Dr. Agus Subekti, M.T	
Penguji III / Pembimbing I	:	Dr. Didi Rosiyadi, M.Kom	

	LEMBAR BIMBINGAN TESIS
	UNIVERSITAS NUSA MANDIRI

NIM : 14002371
Nama : Riki Supriyadi
Pembimbing 1 : Dr. Didi Rosiyadi, M.Kom
Judul Tesis : Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*

NO	Tanggal Bimbingan	Pokok Bahasan	Paraf Dosen Pembimbing 1
1	8 April 2021	Pengajuan Judul	
2	15 April 2021	Pengajuan BAB I	
3	22 April 2021	Revisi BAB 1 & Pengajuan BAB II	
4	16 Juli 2021	Revisi BAB II	
5	25 Juli 2021	Pengajuan BAB III	
6	30 Juli 2021	Revisi BAB III dan Pengajuan BAB IV	
7	31 Juli 2021	Revisi BAB IV dan Pengajuan BAB V	
8	2 Agustus 2021	Acc Keseluruhan	

Catatan untuk dosen pembimbing

Bimbingan Tesis
- Dimulai pada tanggal: 08 April 2021
- Diakhiri pada tanggal: 02 Agustus 2021
- Jumlah pertemuan bimbingan: 8 (Delapan) kali bimbingan

Disetujui oleh,
Dosen Pembimbing

(Dr. Didi Rosiyadi, M.Kom**)**

	LEMBAR BIMBINGAN TESIS
	UNIVERSITAS NUSA MANDIRI

NIM : 14002371
Nama : Riki Supriyadi
Pembimbing 2 : Eni Heni Hermaliani, MM, M.Kom
Judul Tesis : Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*

NO	Tanggal Bimbingan	Pokok Bahasan	Paraf Dosen Pembimbing 2
1	8 April 2021	Pengajuan Judul	
2	15 April 2021	Pengajuan BAB I	
3	22 April 2021	Revisi BAB 1 & Pengajuan BAB II	
4	16 Juli 2021	Revisi BAB II	
5	25 Juli 2021	Pengajuan BAB III	
6	30 Juli 2021	Revisi BAB III dan Pengajuan BAB IV	
7	31 Juli 2021	Revisi BAB IV dan Pengajuan BAB V	
8	2 Agustus 2021	Acc Keseluruhan	

Catatan untuk dosen pembimbing

Bimbingan Tesis
- Dimulai pada tanggal: 08 April 2021
- Diakhiri pada tanggal: 02 Agustus 2021
- Jumlah pertemuan bimbingan: 8 (Delapan) kali bimbingan

Disetujui oleh,
Dosen Pembimbing 2

(Eni Heni Hermaliani, MM, M.Kom**)**

KATA PENGANTAR

Puji syukur Alhamdulillah, penulis panjatkan kehadirat Allah, SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan laporan tesis ini tepat pada waktunya. Dimana laporan tesis ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul tesis, yang penulis ambil sebagai berikut "Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*" Tujuan penulisan laporan tesis ini dibuat sebagai salah satu syarat untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) pada Program Studi Ilmu Komputer (S2) Universitas Nusa Mandiri.

Laporan Tesis ini diambil berdasarkan dataset publik yang bersumber dari UCI Repository. Penulis juga mencari dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, *internet*, dll yang terkait dengan pembahasan pada laporan tesis ini.

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak dalam pembuatan laporan tesis ini, maka penulis tidak dapat menyelesaikan laporan tesis ini tepat pada waktunya. Untuk itu izinkanlah penulis dalam kesempatan ini untuk mengucapkan terima kasih kepada:

1. Rektor Universitas Nusa Mandiri
2. Wakil Rektor I Bidang Akademik Universitas Nusa Mandiri
2. Dekan Fakultas Fakultas Teknologi Informasi Universitas Nusa Mandiri
3. Ketua Program Studi Ilmu Komputer
4. Bapak Dr. Didi Rosiyadi, M.Kom selaku pembimbing pertama yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
5. Eni Heni Hermaliani, MM, M.Kom selaku pembimbing kedua yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
6. Orang tua tercinta yang telah memberikan dukungan material dan moral kepada penulis.

7. Seluruh teman-teman mahasiswa Ilmu Komputer (S2) yang selalu membantu dan memberi dukungan satu sama lain.
8. Seluruh Dosen Program Studi Ilmu Komputer (S2) Universitas Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
9. Seluruh staf di lingkungan Universitas Nusa Mandiri yang telah melayani penulis dengan baik selama kuliah.
10. dan lain-lain.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan laporan tesis ini. Penulis menyadari bahwa penulisan laporan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan untuk yang akan datang.

Akhir kata semoga laporan tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 26 Juli 2021

Riki Supriyadi

Penulis

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya:

Nama : Riki Supriyadi
NIM : 14002371
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : Data Mining
Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri **Hak Bebas Royalti Non-Eksklusif *(Non-exclusive Royalty-Free Right)*** atas karya ilmiah kami yang berjudul: "Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*" beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak Universitas Nusa Mandiri berhak menyimpan, mengalih-media atau bentuk-kan, mengelolanya dalam pangkalan data (*database*), mendistribusikannya dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari kami selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 26 Juli 2021
Yang menyatakan,

Riki Supriyadi

ABSTRAK

Nama : Riki Supriyadi
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : Data Mining
Judul Tesis : Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*

Deposito merupakan simpanan yang hanya dapat dibayarkan pada waktu tertentu dan memiliki tingkat suku bunga lebih tinggi dari cek atau tabungan biasa. Namun jumlah nasabah yang menabung deposito masih tergolong minim. Penelitian ini dilakukan untuk memprediksi calon nasabah yang menabung deposito dengan menerapkan metode CRISP-DM. Serta menentukan hasil akurasi analisa dengan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression.* Dalam mengatasi data tidak seimbang diterapkan *Synthetic Minority Over-sampling Technique* (*SMOTE*). Algoritma *Extreme Gradient Boosting (XGBoost)-SMOTE* dengan penerapan *hyperparameter* menggunakan GridsearchCV dalam menentukan parameter yang akan digunakan mendapatkan hasil terbaik dibanding algoritma lain pada penelitian ini, akurasi yang didapat 97% dan AUC 0.99 yang termasuk kategori *Excellent Classification*. Dengan demikian hasil penelitian menunjukan algoritma *XGBoost-SMOTE* lebih baik dalam memprediksi calon nasabah deposito, serta penerapan *SHAP value* sebagai *feature important* diketahui bahwa *duration* merupakan *feature* yang paling berkontribusi terhadap model yang dibangun.

Kata kunci:
Deposito, *XGBoost*, *SMOTE*, SHAP *Value*

ABSTRACT

Nama : Riki Supriyadi
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : Data Mining
Judul Tesis : Prediksi Calon Nasabah Deposito Berbasis CRISP-DM dan *SMOTE XGBoost Classification*

Deposits are deposits that can only be paid at certain times and have an interest rate higher than ordinary checks or savings. But the number of customers who save deposits is still relatively minimal. This research was conducted to predict prospective customers who save deposits by applying the CRISP-DM method. As well as determining the results of analysis accuracy with an extreme gradient boosting (XGBoost) algorithm that is compared with other algorithms such as Decision Tree, K-Nearest Neighbor and Logistic Regression. In overcoming unbalanced data applied Synthetic Minority Over-sampling Technique (SMOTE). Extreme Gradient Boosting (XGBoost)-SMOTE algorithm with the application of hyperparameter using GridsearchCV in determining the parameters to be used get the best results compared to other algorithms in this study, accuracy obtained 97% and AUC 0.99 which belongs to the category of Excellent Classification. Thus the results of the study showed the XGBoost-SMOTE algorithm is better at predicting prospective deposit customers, as well as the application of SHAP value as a feature important, it is known that duration is the feature that most contributes to the model built.

Keywords:
Deposits, XGBoost, SMOTE, SHAP value

DAFTAR ISI

Halaman

HALAMAN SAMPUL i
HALAMAN JUDUL ii
SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME .. iii
HALAMAN PERSETUJUAN DAN PENGESAHAN TESIS..... iv
LEMBAR BIMBINGAN TESIS v
KATA PENGANTAR vii
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS ix
ABSTRAK x
ABSTRACT xi
DAFTAR ISI xii
DAFTAR TABEL xiv
DAFTAR GAMBAR xv
DAFTAR LAMPIRAN xvi
BAB 1 PENDAHULUAN 1
1.1. Latar Belakang Penulisan 1
1.2. Identifikasi Masalah 3
1.3. Tujuan Penelitian 3
1.4. Ruang Lingkup Penelitian 4
1.5. Hipotesis 4
1.6. Sistematika Penulisan 5
BAB 2 LANDASAN /KERANGKA PEMIKIRAN 6
2.1. Tinjauan Pustaka 6
2.1.1. Data Mining 6
2.1.2. *Extreme Gradient Boosting* (*XGBoost)* 7
2.1.3. *Decision Tree* 9
2.1.4. *K-Nearest Neighbor (KNN)* 10
2.1.5. *Logistic Regression* 11
2.1.6. *Synthetic Minority Over-sampling Technique* (*SMOTE)* 11
2.1.7. *Python* 12
2.1.8. Deposito 13
2.1.9. CRISP-DM 15
2.2. Tinjauan Studi 17
2.3. Tinjauan Organisasi/Objek Penelitian 24
BAB 3 METODOLOGI PENELITIAN 25
3.1. Tahapan Penelitian 25
BAB 4 HASIL PENELITIAN DAN PEMBAHASAN 30
4.1. *Business Understanding* 30
4.2. *Data Understanding* 30
4.3. *Data Preparation* 35
4.3.1. *Preprocessing* 35
4.4. *Modelling* 38
4.4.1. *Extreme Gradient Boosting* (*XGBoost)* 38
4.4.2. *Decision Tree* 39
4.4.3. *K-Nearest Neighbor (KNN)* 39

4.4.4. *Logistic Regression* 39
4.5. *Evaluation* 40
4.5.1. *Extreme Gradient Boosting* (*XGBoost)* 40
4.5.2. *Decision Tree* 42
4.5.3. *K-Nearest Neighbor* (KNN) 44
4.5.4. *Logistic Regression* 46
4.5.5. *XGBoost-SMOTE* 48
4.5.6. *Decision Tree-SMOTE* 50
4.5.7. *KNN – SMOTE* 52
4.5.8. *Logistic Regression – SMOTE* 54
4.5.9. Rangkuman Hasil 56
4.5.10. *Feature Important* dengan SHAP *Value* 58
BAB 5 PENUTUP 60
5.1. Kesimpulan 60
5.2. Saran 61
DAFTAR REFERENSI 62
DAFTAR RIWAYAT HIDUP 68
LAMPIRAN 69

DAFTAR TABEL

Tabel 3.1. *Confusion Matrix* 28
Tabel 3.2. *Performance* Keakurasian AUC 29
Tabel 4.1 *Dataset Bank marketing* 31
Tabel 4.2 Cek *Missing Valu*e 35
Tabel 4.3 *Label Encoding* 36
Tabel 4.4. Hasil Evaluasi *XGBoost* 41
Tabel 4.5. Hasil Evaluasi *Decision Tree* 43
Tabel 4.6. Hasil Evaluasi *KNN* 45
Tabel 4.7. Hasil Evaluasi *Logistic Regression* 47
Tabel 4.8. Hasil Evaluasi *XGBoost- SMOTE* 49
Tabel 4.9. Hasil Evaluasi *Decision Tree-SMOTE* 51
Tabel 4.10. Hasil Evaluasi *XGBoost* 53
Tabel 4.11. Hasil Evaluasi *Logistic Regression-SMOTE* 55
Tabel 4.12 Perbandingan hasil penelitian 56

DAFTAR GAMBAR

Gambar 2.1. Perkembangan suku bunga rata-rata 14
Gambar 2.2. Tahapan CRISP-DM 15
Gambar 3.1. Model Penelitian 25
Gambar 4.1. Atribut tipe Kategori 32
Gambar 4.2. Hubungan variabel kategori dengan target 33
Gambar 4.3 Atribut tipe numerik 34
Gambar 4.4 Data target 34
Gambar 4.5 Mengatasi *Imbalance* data 38
Gambar 4.6 *Confusion* matrix *XGBoost* 40
Gambar 4.7 ROC *XGBoost* 41
Gambar 4.8 *Confusion matrix Decision Tree* 42
Gambar 4.9 ROC *Decision Tree* 43
Gambar 4.10 *Confusion matrix KNN* 44
Gambar 4.11 ROC *KNN* 45
Gambar 4.12 *Confusion matrix Logistic Regression* 46
Gambar 4.13. ROC *Logistic Regression* 47
Gambar 4.14. *Confusion matrix XGBoost- SMOTE* 48
Gambar 4.15. ROC *XGBoost-SMOTE* 49
Gambar 4.16. *Confusion matrix Decision Tree-SMOTE* 50
Gambar 4.17. *ROC Decision Tree-SMOTE* 51
Gambar 4.18 *Confusion matrix KNN-SMOTE* 52
Gambar 4.19 ROC *KNN-SMOTE* 53
Gambar 4.20. *Confusion matrix Logistic Regression-SMOTE* 54
Gambar 4.21 ROC *Logistic Regression-SMOTE* 55
Gambar 4.22 Gambar Hasil perbandingan 57
Gambar 4.23 *Feature Important SHAP Value* 58

DAFTAR LAMPIRAN

Lampiran 1. Sampel Dataset 69
Lampiran 2. Hasil Penelitian 70

BAB I
PENDAHULUAN

1.1. Latar Belakang Penulisan

Bank menjadi pilihan untuk menyimpan atau menabung uang, hal ini sudah menjadi kebiasaan di kalangan masyarakat dari dulu hingga sekarang. Pertumbuhan bank terutama di Indonesia sudah sangat cepat, banyak bank baru yang bermunculan dengan membawa beberapa inovasi dalam pelayanan yang diberikan, lebih mudah dalam proses pembukaan rekening ataupun saat bertransaksi. Hal ini tentu dipengaruhi dengan perkembangan teknologi yang semakin canggih.

Dalam perbankan salah satu produk yang sering ditawarkan kepada nasabah adalah deposito. Deposito adalah simpanan yang hanya dapat ditarik pada saat jatuh tempo [1]. Berdasarkan Undang-Undang Nomor 10 Tahun 1998, terdapat pengertian deposito itu sendiri, yaitu simpanan yang hanya dapat dibayarkan pada waktu tertentu melalui kesepakatan antara penyimpan dan bank [2]. Dan deposito memiliki tingkat bunga tahunan lebih tinggi dari cek atau tingkat bunga tabungan biasa. Namun jumlah nasabah yang menabung deposito masih tergolong minim karena meskipun tingkat bunga deposito lebih tinggi dari tabungan biasa, deposito hanya dapat dibayarkan pada waktu tertentu, hal ini membuat deposito dapat digunakan sebagai alternatif bentuk investasi [3]. Untuk itu, bank harus memanfaatkan peluang ini untuk menerapkan strategi pemasaran dan kegiatan promosi yang efektif.

Bank biasanya menggunakan pemasaran langsung untuk menargetkan basis pelanggannya dan menjangkau pelanggan untuk mencapai tujuan. Pelanggan yang dipilih oleh bank biasanya dihubungi secara pribadi melalui informasi kontak, seperti telepon seluler atau email, sehingga bank dapat menerima informasi langsung dari pelanggan, apakah pelanggan telah berlangganan produk yang disediakan oleh bank atau belum [4]. Dalam pemasaran langsung, dapat menggunakan salah satu metode untuk memprediksi pelanggan yang membuka simpanan di bank menggunakan prediksi berdasarkan data pelanggan yang ada, kemudian menggunakan teknik pengenalan pola (seperti statistik dan matematika) untuk memeriksa kumpulan data besar yang disimpan dan memprosesnya untuk

menemukan hubungan, pola, dan tren yang bermakna. proses tersebut dikenal dengan *data mining*.

Data mining adalah proses menemukan informasi dan pola yang berguna dari data yang sangat besar. *Data mining* juga dikenal sebagai *knowledge Discovery, knowledge extraction, data/ pattern analysis information, harvesting* dan lain lain [5]. *Data mining* menggunakan analisis matematis untuk mendapatkan atau menemukan pola dan tren dari data. Pada umumnya pola-pola tersebut sulit ditemukan dengan eksplorasi data biasa/tradisional, hal ini disebabkan hubungan antar data yang terlalu rumit, atau bisa juga disebabkan oleh data yang begitu besar.

Data mining bertujuan untuk menemukan pola yang sebelumnya tidak diketahui. Setelah diperoleh, pola-pola tersebut dapat digunakan untuk menyelesaikan berbagai macam masalah. *Data mining* juga telah menjadi teknologi baru yang kuat dengan potensi besar untuk membantu bisnis fokus pada hal yang paling penting [6]. Informasi dalam data yang dikumpulkan tentang perilaku pelanggan dan calon pelanggan. *Data mining* memungkinkan perusahaan untuk menemukan informasi dalam jumlah data yang begitu besar melalui pengolahan yang tepat dan efektif menggunakan berbagai metode yang ada pada *data mining* sehingga *data mining* dapat dengan mudah digambarkan. sebagai pola atau model atau aturan atau temuan dari *data mining*, Selain itu, *data mining* dapat membantu menjadi tren penjualan, mengembangkan kampanye pemasaran yang lebih cerdas, dan secara akurat memprediksi loyalitas pelanggan. [5]

Minat nasabah dalam pembukaan tabungan deposito pada perbankan tidak terlalu tinggi, hal ini membuat marketing pemasaran harus bisa mengetahui kondisi atau hal yang bisa menarik nasabah untuk menabung deposito. Berdasarkan hal tersebut perlu dilakukan penelitian untuk menentukan aspek kecenderungan dan juga algoritma apa yang paling sesuai dan efektif dalam melakukan prediksi terhadap nasabah yang melakukan pembukaan tabungan deposito di perbankan.

Penelitian dilakukan dengan data publik, sumber data yang digunakan bisa diperoleh dari berbagai sumber salah satunya UCI *Repository*, berdasarkan dataset *Bank Marketing* yang diperoleh dari UCI *Repository* [7] pada penelitian ini akan dilakukan beberapa eksperimen dalam memprediksi calon nasabah yang menabung

deposito di bank untuk mengetahui kecenderungan apa yang dominan terhadap nasabah yang menabung deposito dengan beberapa algoritma *data mining*.

Adapun manfaat dari penelitian ini adalah untuk mengetahui algoritma apa yang dapat memprediksi calon nasabah deposito dengan baik berdasarkan hasil akurasi yang diperoleh dan membantu manajemen perbankan mengenai minat nasabah terhadap pembukaan tabungan deposito berdasarkan data nasabah yang diperoleh dari pemasaran langsung, serta dapat memberikan bukti secara empiris untuk teori yang berkaitan sehingga dapat dijadikan sumbangan pemikiran untuk pengembangan teori berikutnya.

1.2. Identifikasi Masalah

Berdasarkan latar belakang penelitian maka identifikasi masalah dalam penelitian ini meliputi:

1. Data nasabah pada *Bank marketing* yang diperoleh berdasarkan pemasaran langsung memiliki jumlah yang besar, sehingga perlu diolah menggunakan *data mining*.
2. Perlunya melakukan eksperimen dalam mencari algoritma yang tepat untuk memprediksi nasabah yang melakukan pembukaan tabungan deposito di bank.
3. Perlunya mengetahui faktor yang mempengaruhi nasabah dalam menabung deposito berdasarkan aspek dominan pada hasil eksperimen yang dilakukan untuk mempermudah strategi pemasaran bagi perbankan.

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu:

1. Memperoleh model dari beberapa algoritma yang memiliki akurasi tertinggi bertujuan untuk memprediksi calon nasabah yang menabung deposito.
2. Penelitian ini dapat menganalisa perilaku nasabah yang melakukan pembukaan tabungan deposito di perbankan.
3. Mendapatkan bahan evaluasi untuk pihak marketing bank dalam meningkatkan promosi dan kualitas layanan terhadap nasabah.

1.4. Ruang Lingkup Penelitian

Ruang lingkup dalam penelitian ini yaitu melakukan penelitian menggunakan dataset *Bank marketing* yang diperoleh dari UCI *Repository*. Data diolah dengan melakukan eksperimen menggunakan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression*. Dengan melakukan perbandingan dua eksperimen yaitu data yang tidak seimbang (*imbalance data*) dan data yang sudah seimbang (*balance*) melalui penerapan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) menggunakan bahasa pemrograman *Python*. Kemudian hasil keseluruhan eksperimen dikomparasi untuk menentukan nilai akurasi terbaik dalam melakukan prediksi serta menentukan atribut yang paling berpengaruh terhadap nilai yang dihasilkan.

1.5. Hipotesis

Hipotesis dari penelitian ini yaitu:

1. Hipotesis Umum:

Penerapan algoritma klasifikasi *XGBoost* pada *imbalance data* yang diatasi dengan teknik *SMOTE* dapat meningkatkan akurasi pada model yang dibuat dan memiliki hasil akurasi yang lebih baik dari algoritma data mining lainya seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression.*

2. Hipotesis Khusus:

Hasil akurasi yang dihasilkan dapat diterapkan oleh bank sebagai langkah awal dalam melakukan pemasaran secara langsung terhadap nasabah yang akan membuka tabungan deposito.

1.6. Sistematika Penulisan

Penelitian ini mempunyai susunan dalam penulisan agar tersusun secara sistematis dan memudahkan pembaca untuk mempelajarinya. Berikut sistematika penulisan Tesis ini:

BAB I PENDAHULUAN

Pada bab ini menjelaskan gambaran umum dari penulisan Tesis yang terdiri atas latar belakang penulisan, identifikasi masalah, tujuan penelitian, ruang lingkup penelitian, hipotesis dan sistematika penulisan

BAB II LANDASAN/KERANGKA PEMIKIRAN

Pada bab ini dijelaskan beberapa teori dasar yang mendukung dalam pengerjaan Tesis ini seperti *Extreme Gradient Boosting (XGBoost), Decision Tree, K-Nearest Neighbor* dan *Logistic Regression* dan meliputi penelitian terdahulu serta landasan teori lainnya untuk mendukung penelitian ini.

BAB III METODOLOGI PENELITIAN

Pada bab ini menjelaskan tahap-tahap pengerjaan dalam menyelesaikan penelitian ini sehingga dapat dirancang secara sistematis dan diatur dengan sebaik-baiknya.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Pada bab ini menjelaskan tentang analisa dan pembahasan yang dilakukan mulai dari pengambilan data, pengolahan data dengan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression* serta membandingkan data tidak seimbang (*imbalance*) dengan data seimbang (*balance*) dengan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) untuk mendapatkan akurasi dari masing-masing algoritma yang digunakan.

BAB V PENUTUP

Pada bab ini berisi kesimpulan yang diperoleh analisa dan pembahasan dari bab-bab sebelumnya serta saran untuk pengembangan penelitian selanjutnya**.**

BAB II
LANDASAN/KERANGKA PEMIKIRAN

2.1. Tinjauan Pustaka

Tinjauan pustaka yang digunakan pada penelitian ini merupakan teori-teori yang menjadi landasan dalam penelitian ini yang didapat melalui referensi dari buku, website dan jurnal-jurnal penelitian baik yang internasional maupun nasional. Berikut tinjauan pustaka yang mendukung teoritis dari penelitian ini adalah sebagai berikut:

2.1.1. *Data mining*

Data mining adalah proses penggalian data (sebelumnya tidak diketahui, implisit, dan dianggap tidak berguna) dari sejumlah besar data menjadi informasi, pengetahuan, atau pola [8]. *Data mining* merupakan proses yang menggunakan berbagai teknik dan alat analisis data untuk menemukan hubungan dan pola yang tersembunyi. Pendekatan dasar dalam *Data mining* adalah untuk meringkas data dan untuk mengekstrak informasi berguna yang masuk akal dan sebelumnya tidak diketahui [9].

Secara umum, *data mining* merupakan proses menemukan pola yang menarik dan tersembunyi dalam kumpulan data besar yang disimpan dalam basis data, gudang data, atau area penyimpanan data lainnya. Kunci *data mining* adalah data bisnis, informasi dan solusi. Mengubah data menjadi informasi dan tujuan akhir dari *data mining* adalah menggunakan informasi untuk membuat keputusan bisnis yang lebih baik dan keputusan yang lebih baik. [10]

Data mining dapat menemukan tren dan pola tersembunyi yang tidak muncul dalam analisis kueri sederhana sehingga dapat memiliki bagian penting dalam hal menemukan pengetahuan dan membuat keputusan. Tugas-tugas semacam itu dapat bersifat prediksi seperti klasifikasi dan regresi atau deskriptif seperti *clustering* dan asosiasi [11].

Data mining biasanya dilakukan melalui tiga langkah [5]:

1. Eksplorasi: Hal yang harus dilakukan adalah mempersiapkan data dalam jumlah yang besar, selanjutnya data dibersihkan atau direduksi sesuai dengan yang diperlukan, dan menghilangkan missing value sehingga data yang tersisa adalah data yang benar-benar dapat digunakan.
2. Pemodelan atau identifikasi pola: Membangun model statistik dengan tujuan mengevaluasi sehingga dapat menghasilkan prediksi mana yang terbaik dan paling akurat. Proses ini boleh jadi memakan waktu lama karena model yang diterapkan berbeda pada kumpulan data yang sama dan dilakukan berulang-ulang untuk kemudian dibandingkan hasilnya.
3. Penerapan: Pada tahap akhir ini dilakukan pengujian model yang digunakan terhadap data latih dan data uji untuk menghasilkan prediksi atau perkiraan hasil yang sesuai dengan apa yang diharapkan.

Peranan *data mining* terdiri dari dua [12], diantaranya yaitu:

1. Metode Prediksi: Metode prediksi dengan menggunakan beberapa variabel untuk memprediksi nilai yang tidak diketahui atau nilai di masa yang akan datang dari variabel lain.
2. Metode Deskripsi: Menemukan bentuk yang mampu diartikan manusia (*human-interpretable patterns*) yang dapat menjelaskan *data* tertentu.

2.1.2. *Extreme Gradient Boosting* (*XGBoost)*

Extreme Gradient Boosting (*XGBoost)* diusulkan oleh Chen Tianqi pada tahun 2016, menghadirkan kompleksitas komputasi yang rendah, kecepatan *running* yang cepat dan akurasi yang tinggi [13]. Mirip dengan *gradient boosting*, *XGBoost* menggabungkan pengklasifikasi dasar yang lemah menjadi pengklasifikasi yang lebih kuat. Pada setiap iterasi dari proses pelatihan, sisa dari pengklasifikasi dasar digunakan di berikutnya classifier untuk mengoptimalkan fungsi tujuan [14].

Extreme Gradient boosting (*XGBoost)* digunakan untuk masalah *supervised learning*. Skalabilitas yang bagus dan kecepatan running yang tinggi dari *XGBoost* membuatnya menjadi metode pembelajaran mesin yang sukses [15]. *XGBoost* memanfaatkan *gradient boosting* untuk menemukan nilai objektif yang optimal

[16]. Inti dari algoritma *XGBoost* adalah belajar fungsi baru f(x) agar sesuai dengan residual yang diprediksi terakhir [17].

XGBoost menerapkan ekspansi *second-order Taylor* ke *loss function* untuk menggantikan turunan pertama tidak seperti *Gradient boosting* konvensional, seperti persamaan berikut: [18].

$$L = \sum_i l(y, O(x_i) + \sum_k \Omega(G_k) \quad \text{.....(2.1)}$$

Di mana, l adalah *loss function* dari *training* dan L mendefinisikan *loss function* asli untuk algoritma *XGBoost.* Sisanya dari notasi konstan sama dengan metode boosting. G didefinisikan sebagai *weak estimator* untuk pohon keputusan. Selain itu, kompleksitas pohon keputusan, Ω (G_m) adalah diagregasikan dengan suku pertama untuk membentuk fungsi tujuan. Definisi istilah reguler, Ω (G_m), dihitung sebagai:

$$\Omega(G) = \omega T + \frac{1}{2} a \sum_{j=1}^{T} s_j^2 \quad \text{.....(2.2)}$$

di mana, T menunjukkan jumlah *leaf* pohon keputusan. Sementara, ω_j^2 menunjukkan norma skor L2 untuk masing-masing leaf. γ adalah ambang batas kontrol untuk membagi *node*, dan λ adalah koefisien untuk mengurangi masalah *overfitting*.

Persamaan akhir dapat dibentuk sebagai:

$$L^m = \sum_{i=1}^{N} l(y_i, O_i^{m-1} + G_m(x_i)) + \Omega(G_m) \quad \text{.....(2.3)}$$

$$\approx \sum_{i=1}^{N} [l(y_i, O_i^{m-1}) + g_i\, G_m(x_i) + \frac{1}{2} o_i G_m(x_i)] + \Omega(G_m) \quad \text{.....(2.4)}$$

Akhirnya, dalam persamaan sebelumnya, dua variabel mendefinisikan turunan pertama dan turunan kedua dari *loss function* adalah: *gi* = *∂Om*-1*l*(*yi, Om i* -1) and *oi* = *∂*2*Om*-1*l*(*yi, O*$_1^{m-1}$), masing-masing.

2.1.3. *Decision Tree*

Decision Tree merupakan teknik yang paling banyak digunakan untuk pembelajaran mesin, pengenalan pola, dan tugas klasifikasi [19]. *Decision Tree* adalah pohon keputusan dimana setiap node menunjukkan *fiture* (atribut), setiap tautan (cabang) menunjukkan keputusan (*rule*) dan setiap daun menunjukkan hasil (kategoris atau nilai). Karena pohon keputusan meniru pemikiran tingkat manusia jadi sangat mudah untuk mengambil data dan membuat interpretasi yang bagus [20].

Proses dalam *Decision Tree* yaitu mengubah bentuk data (tabel) menjadi model pohon (*tree*) kemudian mengubah model pohon tersebut menjadi aturan (*rule*) [21].

Keuntungan utama dari algoritma *Decision Tree* adalah akurasi klasifikasi tinggi dan ketahanan yang kuat. kekurangannya adalah bahwa ia cenderung terlalu pas, *Decision Tree* mudah dipengaruhi oleh sampel dan *subtree* dapat diulang berkali-kali dalam *Decision Tree* [19].

Algoritma yang digunakan adalah algoritma ID3. Algoritma ID3 menggunakan konsep dari *Entropy* dan *Information Gain* [21].
Untuk nilai *Entropy* dapat ditemukan dengan menggunakan rumus:

$$Entropy\ (\mathrm{S}) = \sum_{i=1}^{k} -p_j\ log_2 p_j \quad \text{.....(2.5)}$$

Keterangan:

S : ruang (data) sampel yang digunakan untuk *training*

K : banyaknya partisi pada S

pj : probabilitas yang didapat dari *Sum* (Ya) dibagi dengan total sampel

$$\mathrm{Gain}\ (\mathrm{A}) = Entropy\ (S) - \sum_{i=1}^{k} \frac{|Si|}{|S|} Entropy\ (S_{i)} \quad \text{.....(2.6)}$$

Keterangan:

S : ruang (data) sampel yang digunakan untuk *training*

A : atribut

$|Si|$: jumlah sampel untuk nilai V

$|S|$: jumlah seluruh sampel data

$Entropy\ (S_{i)}$: *entropy* untuk sampel – sampel yang memiliki nilai i

2.1.4. *K-Nearest Neighbor* (KNN)

K-Nearest Neighbor (KNN) merupakan salah satu algoritma yang digunakan dalam masalah klasifikasi. cara kerja KNN adalah mencari jarak terdekat antara data yang akan diestimasi dengan tetangga terdekat pada data latih [22].

Tapi KNN adalah metode pembelajaran yang malas. Jumlah perhitungan besar untuk sampel uji diperlukan, dan biaya memori juga tinggi. Sementara itu, pemilihan rumus jarak dalam algoritma KNN juga mempengaruhi final akurasi klasifikasi. Dua faktor utama menyebabkan lebih rendah kinerja KNN. Yang pertama adalah pengaruh tinggi dimensi pada pengukuran jarak, yang mencerminkan bahwa dengan jumlah variabel meningkat dan kemampuan diskriminasi jarak *Euclidean* menjadi lebih buruk. Kedua, semakin besar jangkauan variabel, semakin dominan terjadi pada jarak perhitungan [23].

Algoritma KNN didasarkan pada matrik jarak yang berfungsi untuk menghitung jarak antara sampel yang akan diklasifikasikan x dan setiap sampel dalam set pelatihan, urutkan jarak yang dihitung, dan pilih k sampel pelatihan yang paling dekat ke sampel untuk diklasifikasikan sebagai k nearest neighbors dari x. Jika sampel milik kelas tertentu dari *k nearest neighbors* adalah mayoritas, klasifikasi representatif sampel x diklasifikasikan ke dalam kategori [23].

Algoritma *K-Nearest Neighbor* (KNN) adalah salah satu algoritma paling sederhana yang digunakan untuk menyelesaikan masalah klasifikasi dan seringkali memberikan hasil yang kompetitif dan penting [24].

Sebagian besar peneliti telah menerapkan ukuran *Euclidean* untuk menghitung jarak minimum seperti yang ditunjukkan di bawah ini:

$$Eucd\ (X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad \text{..... (2.7)}$$

Keterangan:

X : Sampel data *training*

Y : Sampel data *testing*

n : *input parameter*

2.1.5. *Logistic Regression*

Logistic Regression adalah salah satu teknik statistik yang paling menguntungkan untuk pembentukan model probabilitas yang baik untuk mengklasifikasikan variabel dependen dikotomis dengan kategoris atau campuran dengan faktor pembeda non-kategoris. Model *Logit* mencoba membentuk model regresi dengan metode *maximum-likelihood* untuk memilih keanggotaan kelas yang terbaik [25].

Logistic Regression adalah teknik analisis statistik untuk menggambarkan hubungan antara satu variabel biner dependen (juga disebut sebagai variabel dikotomis) dan satu atau lebih variabel bebas dalam data [26]. Ia bekerja dengan menemukan yang paling cocok model untuk memilih parameter yang memaksimalkan probabilitas kasus yang diamati. Variabel independen mengacu pada *input* dalam kumpulan data, dan variabel dependen berhubungan dengan parameter keluaran model [27].

Secara matematis, fungsi *Logistic Regression* didefinisikan sebagai:

$$O = log = \frac{prob}{1 - prob} = c_0 + c_1 a_2 + c_2 a_2 + \cdots + c_n a_n \quad \text{.....(2.8)}$$

Keterangan:

O : nilai *respons* yang akan diprediksi,

Prob : mencerminkan probabilitas kehadiran sifat yang diinginkan

1-prob : kemungkinan tidak adanya sifat minat

c1cn : nilai koefisien yang akan ditentukan selama pelatihan

a1..an): *parameter input* dalam data

2.1.6. *Synthetic Minority Over-sampling Technique* (*SMOTE*)

Synthetic Minority Over-sampling Technique (*SMOTE*) merupakan salah satu teknik dalam mengatasi data tidak seimbang atau *imbalance data* didasarkan pada pengambilan sampel data dari kelas minoritas dengan menghasilkan titik data pada segmen utama. Pendekatan ini sangat sederhana, dan sangat berhasil dalam praktek [28].

SMOTE diusulkan oleh Chawla pada tahun 2002, hal ini didasarkan pada prinsip *oversampling* kelas minoritas dan *undersampling* kelas mayoritas untuk menangani masalah ketidak seimbangan sampel. Kelas dengan jumlah sampel besar disebut kelas mayoritas, dan kelas dengan jumlah sampel sedikit sampel disebut kelas minoritas [29]. Misalkan diberikan data dengan jumlah variabel p maka jarak antara $x^T = [x_1, x_2, \ldots, x_p]$ dan $z^T = [z_1, z_2, \ldots, z_p]$ adalah $d(x, y) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2}$. Untuk membangkitkan data dengan metode *SMOTE* maka digunakan persamaan sebagai berikut:

$$x_{syn} = x_i + (x_{knn} - x) \quad \ldots\ldots(2.9)$$

x_{syn} merupakan pengamatan baru hasil pembangkitan, x_{i} adalah pengamatan ke-i, x_{knn} merupakan x terdekat dari x_{i}, serta gamma adalah bilangan acak dari 0 hingga 1. Untuk data nominal, mengisi sebagian besar nilai tetangga terdekat. Untuk perhitungan jarak di *SMOTE*, jika ada variabel kategori maka akan diganti dengan kuadrat median standar deviasi variabel kontinyu kelas minoritas jika nilai kategorik pada pengamatan ke-i dan j berbeda.

2.1.7. *Python*

Python merupakan bahasa pemrograman dengan lintas *platform*, gratis dan *open source*, serta dapat digunakan untuk mengembangkan aplikasi. *Python* memiliki standar *library* yang sangat luas (*Python Standard Library*) yang dapat digunakan untuk menyelesaikan berbagai masalah pemrograman di dunia nyata [30]. *Python* dikembangkan pada tahun 1991 oleh programmer kelahiran Belanda Guido van Rossum di CWI Amsterdam sebagai pepengembangan dari bahasa pemrograman ABC. Karena kecintaan Guido pada acara TV *Monty Python's Flying Circus*, Guido memilih *Python* sebagai nama bahasa yang ia ciptakan [31].

Berbeda dengan bahasa pemrograman R yang terutama ditujukan untuk analisis data statistik, *Python* muncul di berbagai aplikasi yang jauh lebih luas seperti Internet dan pengembangan situs *web*, akses basis data, GUI *desktop*, komputasi ilmiah, dan pengembangan perangkat lunak dan *game* [32].

Python memiliki sintaks yang unik, tetapi tidak terlalu rumit. Oleh karena itu, *Python* adalah salah satu bahasa pemrograman dengan tingkat tinggi yang mudah digunakan. Ada beberapa aturan yang harus diikuti saat menulis kode dalam bahasa pemrograman *Python*. Hal ini dilakukan untuk memprediksi kesalahan atau masalah pada program yang dihasilkan. Aturan pertama tata bahasa *Python* adalah menulis *statement* atau perintah. *Statement* adalah instruksi atau perintah yang harus dijalankan oleh komputer [33].

Fitur Bahasa Pemrograman *Python*

Beberapa fitur dan kelebihan yang dimiliki *Python* [34] adalah:

1. Memiliki koleksi kepustakaan yang banyak. Artinya, telah tersedia modul-modul 'siap pakai' untuk berbagai keperluan, seperti pembuatan *game* hingga artificial intelligence (misal: *Tensor Flow*).
2. Memiliki struktur bahasa yang jelas, sederhana, dan mudah dipelajari.
3. Berorientasi prosedural dan objek sekaligus (*multi paradigma*).
4. Memiliki sistem pengelolaan memori otomatis (*garbage collection*) seperti halnya *Java*.
5. Bersifat modular sehingga mudah dikembangkan dengan menciptakan modul-modul baru, baik dibangun dengan bahasa *Python* maupun C/C++

2.1.8. Deposito

Deposito adalah simpanan berjangka yang hanya dapat ditarik hanya dapat dilakukan pada saat jatuh tempo [1]. Dalam Undang-Undang Nomor 10 Tahun 1998, terdapat pengertian deposito itu sendiri, yaitu simpanan yang hanya dapat dibayarkan pada waktu tertentu melalui kesepakatan antara penyimpan dan bank [2].

Deposito memiliki tingkat bunga tahunan lebih tinggi dari cek atau tingkat bunga tabungan biasa. Namun jumlah nasabah yang menabung deposito masih tergolong minim karena meskipun tingkat bunga deposito lebih tinggi dari tabungan biasa, deposito hanya dapat dibayarkan pada waktu tertentu, hal ini membuat deposito justru dapat digunakan sebagai alternatif bentuk investasi [3].

Jenis-jenis deposito [35]:

1. Deposito Berjangka

   Deposito berjangka adalah deposito yang diterbitkan dalam jangka waktu tertentu. Deposito berjangka biasanya antara 1, 3, 6 dan 12 bulan. Deposito berjangka diterbitkan atas nama orang perseorangan dan badan hukum. Pemegang deposito tetap. Bunga deposito tetap dapat ditarik setiap bulan atau pada akhir jangka waktu atau tergantung pada periode.

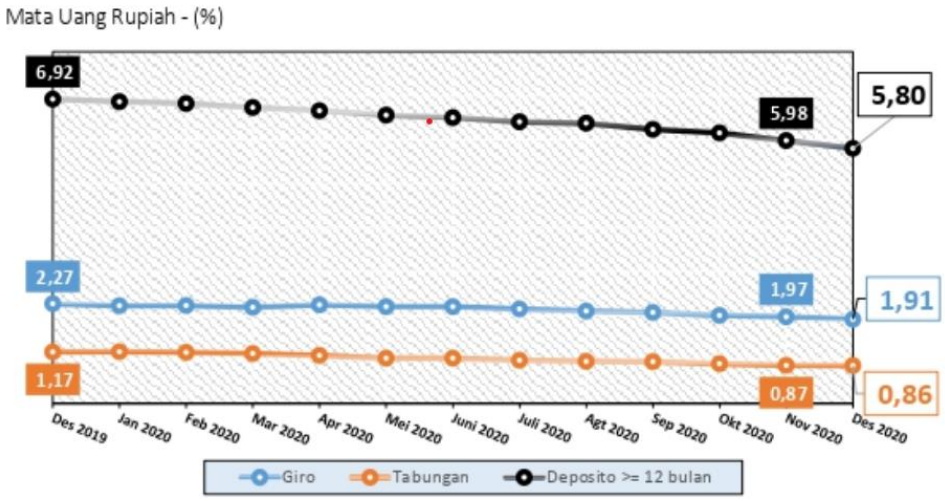
2. Sertifikat Deposito

   Seperti halnya deposito berjangka, sertifikat deposito adalah deposito dengan jangka waktu 1, 3, 6, dan 12 bulan. Satu-satunya perbedaan adalah bahwa sertifikat deposito diterbitkan sebagai sertifikat dan dapat dijual atau dialihkan kepada pihak ketiga. Perbedaan lainnya adalah pembayaran bunga. Tanda terima setoran dapat diterbitkan di muka, tunai atau tunai, dan dapat diterbitkan secara bulanan atau tertunda.

3. *Deposito On Call*

   Deposito on call (DOC) adalah simpanan yang digunakan oleh deposan yang memiliki dana dalam jumlah besar tetapi tidak digunakan dalam jangka waktu tertentu. Jangka waktu penerbitan giro minimal 7 hari dan paling lama 1 bulan. deposit diterbitkan atas nama anda.

Suku Bunga pada deposito merupakan suku bunga yang tinggi dibandingkan dengan tabungan dan giro.



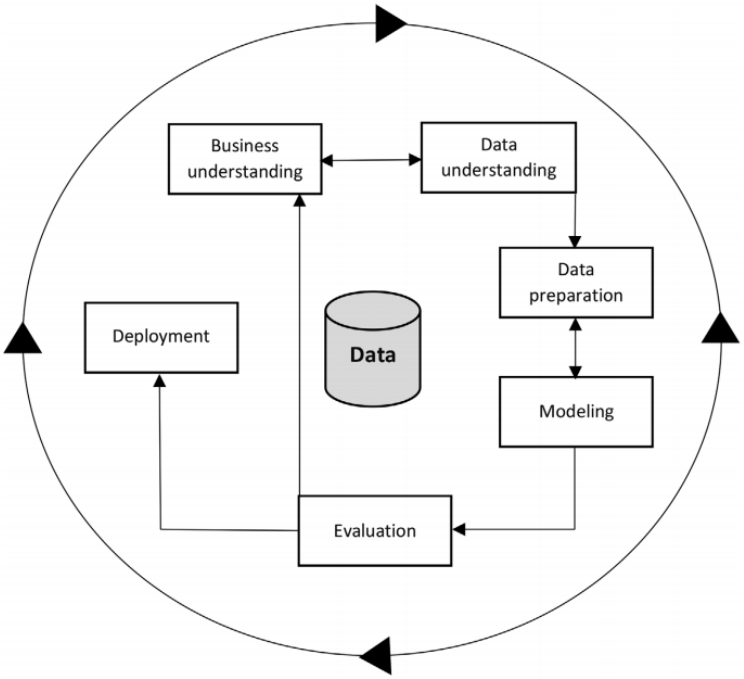
Sumber: Lembaga Penjamin Simpanan (www.lps.go.id)

Gambar 2.1 Perkembangan suku bunga rata-rata

2.1.9. CRISP-DM

Cross-industry Standard Process for Data Mining (CRISP-DM) merupakan *platform* yang mengubah masalah bisnis menjadi tugas *data mining* dan melaksanakan proyek penambangan data secara independen dari area aplikasi dan teknologi yang digunakan. secara luas dianggap sebagai prinsip panduan yang paling relevan dan komprehensif untuk melaksanakan proyek analitik [36].

Berikut enam tahapan yang menunjukkan proses CRISP-DM model dan interaksinya.



Sumber: Sanjiv Jaggia, etc (2020) [36]

Gambar 2.2 Tahapan CRISP-DM

Tahapan CRISP-DM [37].

1. *Business Understanding*
    a. Tujuan dan kebutuhan proyek yang terperinci secara keseluruhan lingkup bisnis atau unit penelitian.
    b. Menerjemahkan tujuan dan batasan ke dalam formula dan definisi masalah data mining.
    c. Siapkan strategi awal untuk mencapai tujuan

2. *Data Understanding*
    a. Mengumpulkan data.
    b. Menggunakan investigasi analisis data untuk mengidentifikasi lebih lanjut data dan mencari pengetahuan awal di dalamnya.
    c. Mengevaluasi kualitas data.
    d. Jika diinginkan, pilih sekelompok kecil data yang mungkin mengandung pola masalah
3. *Data Preparation*
    a. Siapkan data yang tersedia dari awal, karena ini adalah kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini adalah pekerjaan berat yang perlu dilakukan dilakukan secara intensif.
    b. Pilih kasus dan variabel yang Anda inginkan menganalisis dan yang sesuai dengan analisis yang ingin Anda lakukan.
    c. Lakukan perubahan pada beberapa variabel jika diperlukan.
    d. Siapkan data awal supaya siap saat fase pemodelan
4. *Modeling*
    a. Persiapkan data yang tersedia dari awal, karena ini adalah kumpulan data yang akan digunakan pada langkah selanjutnya. Tahap ini merupakan kerja keras yang perlu dilakukan dengan sungguh-sungguh.
    b. Pilih kasus dan variabel yang akan dianalisis dan yang sesuai dengan analisis yang ingin dilakukan.
    c. Lakukan perubahan pada beberapa variabel jika diperlukan.
    d. Siapkan data awal agar siap untuk perangkat pemodelan
5. *Evaluation*
    a. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum diimplementasikan untuk digunakan.
    b. Tentukan apakah sebuah model yang memenuhi tujuan dari fase awal.
    c. Tentukan apakah ada masalah penting dari important bisnis atau penelitian yang tidak ditangani dengan baik.
    d. Membuat keputusan yang berkaitan dengan penggunaan hasil dari enambangan data.

6. *Deployment*
    a. Menggunakan model yang dihasilkan
    b. Pembentukan model tidak menunjukkan selesainya proyek.
    c. Contoh sederhana penerapan adalah laporan generasi.
    d. Contoh penerapan yang kompleks adalah aplikasi proses penambangan data secara paralel dengan departemen yang lain.

Namun dalam penelitian ini yang dilakukan hanya lima tahap saja, yaitu sampai tahap *evaluation*.

2.2. Tinjauan Studi

Tinjauan studi berdasarkan literatur mengenai pembahasan *data mining* telah dilakukan pada penelitian sebelumnya. Berikut adalah penelitian terkait yang sebelumnya dilakukan untuk mendukung teoritis dari penelitian ini.

Pertama, penelitian yang berjudul Perbandingan Metode *SMOTE Random forest* Dan *SMOTE XGBoost* Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada *Imbalance Class Data*, yang dilakukan oleh Muhamad Syukron, Rukun Santoso, Tatik Widiharih (2020). Pada penelitian yang dilakukan Algoritma *SMOTE* digunakan untuk meningkatkan akurasi prediksi dari data yang tidak seimbang (*imbalance*). data yang digunakan dalam penelitian memiliki 24 variabel bebas berupa pasien data diri, gejala hepatitis C, dan hasil pemeriksaan laboratorium. Hasil menunjukkan bahwa *Random forest* memiliki akurasi 74, 79% dan *XGBoost* memiliki akurasi 74, 68% tetapi nilai *recall* kurang dari 2%. Setelah menerapkan *SMOTE* pada kedua algoritma mendapat peningkatan akurasi yaitu *Random forest SMOTE* memiliki akurasi 80, 97% dan *XGBoost SMOTE* memiliki akurasi 78, 63%. *SMOTE Random forest* cenderung memiliki akurasi yang lebih tinggi dibandingkan dengan *SMOTE XGBoost* akan tetapi *SMOTE XGBoost* memiliki nilai *recall* yang lebih tinggi dibandingkan *SMOTE Random forest.* Berdasarkan hasil penelitian disimpulkan bahwa algoritma *SMOTE* dapat memperbaiki model sehingga dapat memprediksi dengan akurat pada semua kelas respon. *SMOTE Random forest* memiliki akurasi keseluruhan yang lebih tinggi dibanding *SMOTE XGBoost.* Variabel yang berpengaruh besar pada pemodelan *SMOTE Random*

forest dan *SMOTE XGBoost* adalah variabel hasil uji laboratorium seperti jumlah sel darah, jumlah enzim ALT, serta jumlah HCV RNA dalam tubuh pasien [38].

Kedua, Penelitian yang berjudul *A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning* yang dilakukan oleh Shutao Wang, Shiyu Liu, Jingkun Zhang, Xiange Che, Yuanyuan Yuan, Zhifang Wang, Deming Kong (2020). Dataset yang digunakan dalam penelitian adalah *diesel data set* yang diperoleh dari http: //www.eigenvector.com/data/swri/index.html. Kumpulan data NIRS diperoleh dan dipublikasikan di Southwest Research Institute (SWRI). terdiri dari 784 sampel diesel yang berbeda NIRS-nya menunjukkan hasil akurasi pada *XGBoost* 75.63%, *Tree-XGBoost* 80.67%, *SMOTE-XGBoost* 86.55% Tree-*SMOTE-XGBoost* 94.96% *Tree-SMOTE-DT* 85.71%. Metode *Tree-SMOTEXGBoost* tidak hanya memiliki akurasi keseluruhan yang tinggi, tetapi juga memiliki akurasi yang relatif seimbang dari setiap kategori. Hal ini menunjukan adanya peningkatan akurasi pada model [39].

Ketiga, Penelitian yang berjudul *iDTi-CSSMOTEB: Identification of Drug–Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost With Over-sampling Technique SMOTE*, yang dilakukan oleh S.M. Hasan Mahmud Wenyu Chen, Hosney Jahan, Yongsheng Liu, Nasir Islam Sujan, dan Saeed Ahmed (2019). Dalam penelitianya mereka menggunakan empat dataset yang dikomparasi dengan mengusulkan model komputasi high-throughput baru, yang disebut *iDTi-CSSMOTEB*, untuk identifikasi DTI berdasarkan struktur obat kimia dan urutan protein. kemudian, urutan protein diekstraksi melalui posisi-spesifik *scoring matrix* (PSSM)-Bigram, komposisi asam amino *pseudo amphiphilic* (AM-PseAAC) dan *dipeptide Deskriptor* PseAAC yang mewakili informasi evolusi dan urutan. Struktur obat kimia direpresentasikan sebagai sidik jari substruktur molekul (MSF) yang menggambarkan keberadaan pecahan atau kelompok. Dan menggunakan teknik *SMOTE Oversampling* untuk mengatasi masalah ketidakseimbangan. dari kumpulan data dan menerapkan algoritma *XGBoost* sebagai pengklasifikasi untuk memprediksi DTI. Hasil prediksi menunjukkan bahwa strategi perubahan kinerja dari *random sampling* ke *Oversampling SMOTE* dari semua metrik evaluasi untuk *dataset reseptor* nuklir

perlu diperhatikan, di mana nilai AUC melonjak menjadi 0, 9350 dari 0, 8611. Di antara empat data set target obat, enzim memperoleh AUC nilai tertinggi untuk kedua teknik penyeimbangan data/ *balancing*, yaitu 0, 9607 untuk *random sampling* dan 0, 9830 untuk *oversampling SMOTE* [40].

Keempat, penelitian yang berjudul *Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique,* yang dilakukan oleh Mengying Lin, Xuefen Zhu, Teng Hua 1, Xinhua Tang, Gangyi Tu dan Xiyuan Chen (2021). Pada penelitianya menggunakan strategi yang menggabungkan algoritma *eXtreme Gradient Boosting* (*XGBoost*) yang ditingkatkan dengan menggunakan minoritas sintetis, teknik *oversampling* dan *nearest neighbor* yang diedit (*SMOTE*-ENN) teknik resampling untuk mendeteksi data yang tidak seimbang sedang dan, akurasi pengujian *XGBoost* yang ditingkatkan adalah sekitar 4% hingga 5% lebih tinggi dari *Decision Tree* dan *random forest*. Sementara itu, hasil pengujian untuk metode perbaikan menunjukkan peningkatan yang signifikan dalam indikator evaluasi, sedangkan nilai *recall* relatif stabil, di atas 90%, dan skor F1 yang sesuai adalah *corresponding* lebih dari 92%. Saat menguji pada kumpulan data dengan tingkat ketidakseimbangan yang berbeda, ada peningkatan yang berbeda sekitar 10% hingga 20% pada nilai *recall*, dan 6% hingga 11% pada skor F1, dengan akurasi pengujian berkisar antara 90, 42% hingga 96, 04% [17].

Kelima, penelitian dengan judul *An Interpretable Aid Decision-Making Model for Flag State Control Ship Detention Based on SMOTE and XGBoost*. Yang dilakukan oleh Jian He, Yong Hao dan Xiaoqiong Wang (2019). Dalam penelitian yang dilakukan mengusulkan model pengambilan keputusan baru berbasis *machine learning,* yang disebut *SMOTE-XGBoost* -Ship detention model (SMO-XGB-SD)), menggunakan algoritma *extreme gradient boosting* (*XGBoost)* dan algoritma *Synthetic Minority Over-sampling Technique* (*SMOTE*) untuk mengidentifikasi apakah sebuah kapal harus ditahan. Hasil verifikasi menunjukkan bahwa algoritma SMO-XGB-SD mengungguli *random forest* (RF), *support vector machine* (SVM), dan algoritma *logistic regression* (LR) [29].

Keenam, penelitian yang berjudul *An Alternative SMOTE Oversampling Strategy for High-dimensional Datasets*. Yang dilakukan oleh Sebasti´an Maldonadoa, Julio L´opezb, dan Carla Vairettia (2019). Dalam penelitianya

melakukan pendekatan *Synthetic Minority Over-sampling Technique* (*SMOTE*) diadaptasi untuk pengaturan biner dimensi tinggi. Metrik jarak baru diusulkan untuk perhitungan lingkungan untuk setiap sampel minoritas, yang memperhitungkan hanya subset dari atribut yang tersedia yang relevan untuk tugas tersebut. Tiga varian untuk metrik jarak dieksplorasi: jarak *Euclidean, Manhattan,* dan *Chebyshev*, dan empat strategi peringkat yang berbeda: *Fisher Score,* Mutual *Information*, *Eigenvector Centrality*, dan *Correlation* Skor. Pada penelitianya dibandingkan dengan berbagai teknik *oversampling* pada dataset dimensi rendah dan tinggi dengan adanya *imbalance data, Oversampling* yang diusulkan menunjukkan hasil yang unggul rata-rata saat dibandingkan.[28].

Ketujuh, penelitian dengan judul *An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study*. Yang dilakukan oleh Saleh R. Mousa, Peter R. Bakhit, and Sherif Ishak. Penelitiannya mengembangkan *extreme gradient boosting classifier* (*XGBoost* ing), menggunakan dataset SHRP2-NDS untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap kejadian kecelakaan dan hampir kecelakaan CNC. Performa dari *XGBoost* classifier dievaluasi terhadap tiga algoritma pembelajaran mesin lainnya. Hasil menunjukkan bahwa model *XGBoosting* mengungguli model lain dengan akurasi deteksi 85% dan mengidentifikasi "perilaku pengemudi" dan "pengaruh persimpangan" sebagai faktor yang paling berkontribusi terhadap deteksi kecelakaan dan hampir kecelakaan (CNC) [41].

Kedelapan, penelitian yang berjudul *Visualization and Analysis in Bank Direct Marketing Prediction*, yang dilakukan oleh Alaa Abu-Srhan, Sanaa Al zghoul, Bara’a Alhammad, dan Rizik Al-Sayyed (2019), dalam penelitiannya melakukan mekanisme visualisasi untuk tugas klasifikasi sederhana. Percobaan dilakukan pada kumpulan data yang tidak seimbang atau *imbalance* untuk data *bank marketing* dari lembaga bank Portugis. Tujuannya adalah untuk memprediksi apakah pelanggan akan berlangganan deposito berjangka atau tidak. Eksperimen dilakukan dengan menggunakan teknik *oversampling* yang berbeda, dengan lima pengklasifikasi umum yang dipilih yaitu *SMOTE*, ADASYN, ROS, ADOMS, SPIDER, AHC. hasil menunjukkan bahwa AHC dengan *SMOTE* mengungguli teknik *oversampling* lainnya. *Naïve Bayes* dan SVM memberikan akurasi prediksi

terbaik dan Nilai Gmean. *Naive Bayes SMOTE* 80.80%, *Naive Bayes* AHC 80.96% [42].

Kesembilan, penelitian yang berjudul *Classification Of A Bank Data Set On Various Data mining Platform Bir Banka Müşteri Verilerinin Farklı Veri Madenciliği Platformlarında Sınıflandırılması*, yang dilakukan oleh Muhammet Sinan Başarslan dan İrem Düzdar Argun (2018). Dalam penelitianya, kumpulan data *bank marketing* pada UCI digunakan dengan membuat model dengan algoritma klasifikasi yang sama di program penambangan data yang berbeda. Kriteria akurasi, presisi dan ukuran digunakan untuk menguji kinerja model klasifikasi. Saat membuat model klasifikasi, set data pengujian dan pelatihan dibagi secara acak untuk mengevaluasi kinerja data set. Program data mining yang digunakan untuk proses tersebut adalah R, Knime, RapidMiner dan WEKA. Dan, algoritma klasifikasi yang digunakan di penelitian ini adalah k-*nearest neighbor (k-nn), Naive Bayes,* dan *C4.5 decision tree.* Dengan hasil perolehan algoritma C4.5 yang mendapatkan nilai tertinggi yaitu *accuracy* dengan R 90% *Precision* dengan Weka 93% dan F-*Measurement* 94% dengan R. [43].

Sepuluh, penelitian yang berjudul *Predicting The Success of Bank marketing using Classification Techniques.* Yang dilakukan oleh S.Vaishnavi, Gattu Divya Lakshmi, Dandu Vinisha Reddy dan Majeti Satya Naga Sulochana (2020), dalam penelitian yang dilakukan menggunakan dataset publik dari UCI *repository* yaitu membandingkan prediksi menggunakan teknik *Data mining* yang berbeda seperti *Decision Tree* (DT), *Naïve Bayes*, *Neural Networks* (NN), KNN dan *support vector machine* (SVM). Kemudian menemukan teknik yang memberikan nilai akurasi prediksi tertinggi. Dari hasil penelitian yang telah dilakukan dengan komparasi beberapa algoritma berikut hasil yang didapat, *Decision Tree* 88.64% *Naive Bayes* 87.54%, SVM 90.35%, ANN 88.97% dan KNN 89.01%. dan berdasarkan hasil yang tadi dapat disimpulkan bahwa algoritma SVM merupakan algoritma terbaik dalam memprediksi nasabah yang melakukan deposit pada penelitian tersebut [44].

Sebelas, penelitian yang berjudul *Comparing Supervised Classification Methods for Financial Domain Problems* yang dilakukan oleh Victor Ulisses Pugliese, Celso Massaki Hirata and Renato Duarte Costa (2020). Pada penelitian melakukan studi komparatif untuk mengidentifikasi metode mana yang berkinerja

terbaik pada masalah analisis risiko, terjadinya penipuan, dan pemberian kredit. Metode yang digunakan *Naive Bayes, Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, XGBoost,* dan *Multilayer Perceptron* untuk mengidentifikasi apakah ada metode yang secara sistematis mengungguli yang lain untuk masalah yang disebutkan di atas dan penerapan Optuna, yang merupakan kerangka kerja optimasi Hyperparameter generasi berikutnya pada metode untuk mencapai hasil yang lebih baik pada lima dataset di Domain Keuangan salah satunya *bank marketing*. Hasil menunjukan *XGBoost* adalah metode terbaik dengan mengadopsi metrik kinerja Skor F1 88.76 % dan AUROC 88.23 %. hasil yang lebih baik dalam menerapkan Optuna mengalami peningkatan yaitu Skor F1 89.56 % dan AUROC 89.11%. Dengan hasil yang didapat disimpulkan bahwa XGBoost adalah metode klasifikasi pembelajaran mesin yang direkomendasikan ketika mengusulkan metode baru untuk masalah analisis. [45]

Dua belas, penelitian yang berjudul *Analysis of applying Genetic Algorithm to Simple Neural Network Based on Bank Direct Marketing Dataset* yang dilakukan oleh Yitong Chen (2018), penelitian dilakukan mengatasi ketidak seimbangan dan mengadopsi GA untuk memilih atribut informatif untuk membentuk MLP's masukan. Membandingkan hasil untuk dataset asli yang tidak seimbang dan dataset yang diperbesar, ditemukan bahwa dataset yang diperbesar sedikit lebih baik bagi MLP. kumpulan data yang diperluas sampel dipilih untuk penggunaan lebih lanjut. Akurasi pengujian rata-rata adalah 77% dan rata-rata skor f1 adalah 44% untuk kumpulan data yang dipilih ini. Meskipun dataset ini tampaknya telah melakukan yang terbaik, skor dan akurasi f1 tidak cukup ideal untuk menghasilkan informasi yang sangat berarti. [46]

Tiga belas, penelitian yang berjudul *Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing* yang dilakukan oleh Khairul Nizam Abd Halim, Abdul Syukor Mohamad Jaya, Ahmad Firdaus Ahmad Fadzi (2020). hasilnya menunjukkan bahwa biner model klasifikasi dilengkapi dengan teknik pembersihan data seperti *Missing common* (MC) dan *Tomek Links* (TL) menunjukkan hasil yang lebih baik dibandingkan dengan *Ignore Missing* (IM). Dari segi data normalisasi, teknik seperti *MaxAbsScaler* (MAS) dan

MinMaxScaler (MMS) secara konsisten menunjukkan kinerja yang lebih baik dari teknik normalisasi lainnya. Model klasifikasi digunakan dalam penelitian ini menggunakan algoritma pra-pemrosesan data kombinasi MC-TL-MMS. Algoritma menggunakan pendekatan ini mampu merekam AUC sebesar 0, 9129 dan 0, 9464 dengan menggunakan 16 fitur dan 20 fitur masing-masing [47].

Empat belas, penelitian yang berjudul "*Handling Imbalanced Data: SMOTE vs. Random Undersampling"* yang dilakukan oleh (Mishra, 2017). Pada penelitian yang dilakukan bertujuan untuk membandingkan hasil dari metode pengambilan sampel yang berbeda yaitu *Randomized Under Sampling, SMOTE* dengan validasi dan tanpa validasi pada kumpulan data yang tidak seimbang atau *imbalance data*, dengan algoritma klasifikasi *Random Forest* dan *XGBoost*. Hasil penelitianya menunjukan *XGBoost* berkinerja lebih baik daripada *Random Forest* untuk semua kombinasi dalam hal *run time, tradeoff* antara sensitivitas dan spesifisitas dan ROC. Dengan *cross validation* hasil yang baik dihasilkan oleh *Random Forest* dengan *SMOTE* yaitu 79, 19% [48].

Lima belas, penelitian yang berjudul *Imbalanced customer Classification for bank direct marketing* yang dilakukan oleh Georgios Marinakos, Sophia Daskalaki (2017), Proses eksperimen menunjukkan bahwa perhitungan kekuatan algoritma yang dikombinasikan dengan beberapa praktik *non-random resampling* dapat meningkatkan efektivitas untuk meramalkan kelas yang diminati. Rata-rata, di semua algoritma dan upaya resampling, prediksi sukses meningkat dari 37% yang buruk ke atas hingga 80% sambil mempertahankan prediksi respons negatif dari 73%, Selain itu, teknik ini dikombinasikan dengan *SMOTE,* terbukti secara konsisten berkinerja baik di semua pengklasifikasi, dan merupakan pilihan yang aman terlepas dari algoritma klasifikasi yang diterapkan [49].

Berdasarkan penelitian sebelumnya dapat diketahui perbedaan dari penelitian yang akan dilakukan, yaitu:

1. Sebagian besar penelitian yang dilakukan dengan dataset serupa hasil akurasi tertinggi yang diperoleh adalah 90.35% dengan algoritma SVM dan 90% dengan algoritma C4.5 menggunakan R, sedangkan penelitian dengan mengatasi *imbalance data* dihasilkan akurasi *Naive Bayes SMOTE* 80.80%, *Naive Bayes* AHC 80.96%.

2. Melakukan peningkatan akurasi terhadap model yang dibangun pada algoritma *XGBoost* dengan teknik *SMOTE* dalam mengatasi *imbalance data* untuk memprediksi calon nasabah deposito menggunakan *Python*.
3. Belum ditemukan penelitian serupa yang dilakukan untuk mengetahui atribut yang paling berkontribusi terhadap model yang dibangun berdasarkan algoritma *SMOTE XGBoost* dengan penerapan SHAPE *value*.

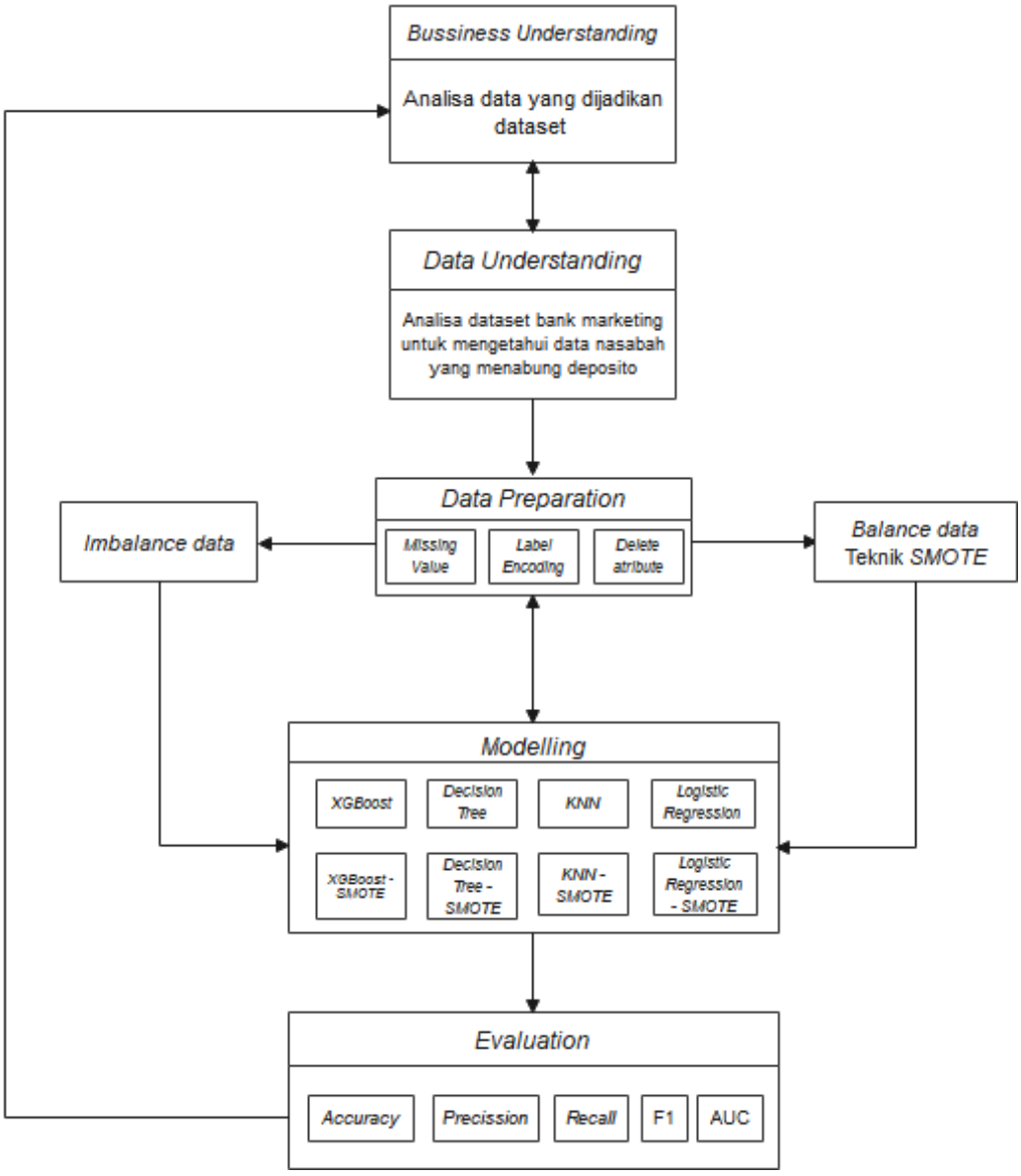
2.3. Tinjauan Organisasi/ Objek Penelitian

Data yang digunakan dalam penelitian ini adalah *Bank marketing* Dataset, merupakan dataset publik yang diperoleh dari UCI *Repository* https: //archive.ics.uci.edu/ml/datasets/bank+marketing, Data tersebut terkait dengan pemasaran langsung dari lembaga perbankan Portugis. pemasaran didasarkan pada panggilan telepon untuk mengetahui nasabah yang akan melakukan deposito dan tidak.

BAB III
METODOLOGI PENELITIAN

3.1. Tahapan Penelitian

Penelitian dilakukan menggunakan metode eksperimen. Eksperimen yang dilakukan dengan menerapkan CRISP-DM untuk mendapatkan informasi yang mendalam tentang nasabah yang akan menabung deposito, dan Model yang digunakan dalam penelitian ini yaitu:



Sumber: Hasil Penelitian (2021)

Gambar 3.1. Tahapan Model Penelitian

Gambar 3.1. menunjukan tahapan model penelitian dengan CRISP-DM, urutan tahapan tidak kaku, tahapan bergerak bolak balik terhadap tahapan yang berbeda. Tanda panah menunjukan alur kepentingan dan ketergantungan antar setiap tahapan. Penjelasan dari kelima tahapan metode dengan CRISP-DM yaitu:

1. Tahap *Business Understanding*

Tahap pertama CRISP-DM berfokus pada pemahaman tujuan penelitian, dan kemudian mengubah pemahaman ini ke dalam rumusan dan definisi masalah data mining. Pada tahap ini, penting untuk memahami bidang masalah dan menemukan solusi yang cocok untuk masalah yang ada, serta memahami faktor-faktor yang dapat mempengaruhi hasil penelitian. Penerapan data mining untuk mengklasifikasikan nasabah yang menabung deposito berdasarkan strategi pemasaran langsung. Diharapkan melalui penelitian ini bank dapat menerapkan strategi pemasaran yang tepat untuk meningkatkan loyalitas nasabah dan memperoleh nasabah baru.

2. Tahap *Data Understanding*

Tahap Data Understanding bertujuan untuk identifikasi dan pahami data yang dipunya. Selain itu, data harus diverifikasi. Data yang digunakan dalam penelitian ini adalah data *Bank marketing* yang diperoleh dari repositori UCI.

3. Tahap *Data Preparation*

Tahap *Data Preparation* mencakup semua aktivitas yang diperlukan untuk membuat kumpulan data akhir atau data yang akan dimasukan kedalam alat pemodelan dari data mentah awal. *Data Preparation* yang dilakukan dalam penelitian ini meliputi cek dan menghapus *missing value*, hapus atribut yang tidak digunakan, mengubah data kategori ke numerik, dan melakukan eksperimen berbeda yaitu data tidak seimbang dan data seimbang.

4. Tahap *Modelling*

Pemodelan yang diterapkan pada penelitian ini dengan membagi data *training* dan *testing* 80:20 dimana 80% dijadikan data *training* dan 20% data *testing*. Ada

dua eksperimen yang dilakukan yaitu menggunakan *imbalance* data dan *balance* data dimana dilakukan *oversampling* menggunakan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada proses latih/ *training*.

Eksperimen menggunakan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression*. Setiap algoritma yang diujikan akan dilakukan *tuning hyperparameter* dengan GridsearchCV untuk menentukan parameter yang akan digunakan berdasarkan dari *best score* dan parameter terbaik yang dihasilkan dari proses uji. GridSearchCV adalah salah satu proses yang digunakan untuk memilih *hyperparameter* terbaik untuk model tertentu. *Hyperparameter* adalah parameter yang ditentukan tanpa proses pengujian, yaitu parameter yang tidak ditentukan oleh mesin. *Hyperparameter* yang diperoleh merupakan parameter terbaik yang akan dimasukkan pada model. [50]. GridSearchCV menjalankan kombinasi *hyperparameter* standar dan menghitung validasi silang rata-rata untuk setiap kombinasi. Dalam penelitian ini akan menggunakan GridSearch dengan *cross validation* 10. *Cross validation* merupakan metode validasi yang membagi data menjadi K bagian, kemudian setiap bagian melewati proses klasifikasi [51]. Hasil dari pemodelan (*modeling*) dapat mengetahui parameter terbaik yang digunakan dalam melakukan evaluasi.

5. Tahap *Evaluation*

Model yang dihasilkan diuji dengan menggunakan *confusion matrix*, yang menentukan tingkat akurasi. *confusion matrix* menggambarkan hasil akurasi, mulai dari prediksi positif benar, prediksi positif palsu, prediksi negatif benar, dan prediksi negatif palsu. Akurasi dihitung berdasarkan dari semua prediksi yang benar (positif dan negatif) dan dari semua data uji. Semakin tinggi nilai akurasi, semakin baik kinerja model. Pengujian juga diukur dengan kurva ROC. Kurva ROC menggambarkan derajat positif sebagai kurva. Uji dilakukan dengan menghitung luas daerah di bawah kurva (AUC). Semakin tinggi nilai ROC dan AUC maka semakin baik pembentukan model klasifikasi.

a. *Confusion Matrix*

Confusion Matrix merupakan visualisasi untuk mengevaluasi dari kinerja model klasifikasi. Untuk melakukan klasifikasi evaluasi komparatif, maka dalam penelitian ini menggunakan *Confusion Matrix*. *Confusion Matrix* ini meliputi informasi tentang kelas yang sebenarnya dan kelas prediksi. Hal ini akan ditemukan pada kolom matriks yang mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kejadian pada kelas tersebut. *Confusion Matrix* adalah salah satu alat ukur berbentuk matrik 2x2 yang digunakan untuk mendapatkan jumlah ketepatan algoritma yang dipakai.

Tabel 3.1. *Confusion Matrix*

True *Condition*	***Predicted Condition***	
	Positive	***Negative***
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

Sumber: Meng, Yang, Qian dan Zhang (2020) [52]

Untuk menghitung performa dari model klasifikasi dapat dihitung dengan menggunakan beberapa cara yaitu:

1) *Accuracy*

Dengan menggunakan rumus di bawah ini akan didapatkan akurasi dari matriks yang mengukur tentang rasio kebenaran dari prediksi dari seluruh data yang dievaluasi.

$$\frac{TP+TN}{TP+FP+TN+FN} \quad \text{..... (3.1)}$$

2) *Precision*

Dengan menggunakan rumus di bawah akan mengukur pola positif yang diprediksi dengan benar dari total pola yang diprediksi di kelas positif.

$$\frac{TP}{TP+FP} \quad \text{.....(3.2)}$$

3) *Recall*

Dengan menggunakan rumus di bawah akan mengukur fraksi pola positif yang diklasifikasikan dengan benar.

$$\frac{TP}{TP+TN} \quad \text{.....(3.3)}$$

4) *F-measure*

Dengan menggunakan rumus di bawah ini merupakan nilai *comprefensif* dari *recall* dan *precision.*

$$\frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} \quad \text{.....(3.4)}$$

Keterangan:

TP : Jumlah kasus positif yang tergolong positif

FP : Jumlah kasus negatif yang tergolong positif

TN : Jumlah kasus negatif yang tergolong negatif

FN : Jumlah kasus positif yang tergolong negatif

b. ROC Curve

Kurva ROC adalah metode untuk melihat, mengelola, dan memilih klasifikasi berdasarkan model yang diperlukan. ROC juga dapat digunakan untuk menjelaskan tingkat sensitivitas dan spesifisitas.

Performance akurasi *Area Under Curve* dapat diklasifikasikan menjadi lima kelompok yang terlihat pada tabel 3.2 sebagai berikut:

Tabel 3.2. *Performance* Keakurasian AUC

Performance	Klasifikasi
0.90 – 1.00	*Excellent Classification*
0.80 – 0.90	*Good Classification*
0.70 – 0.80	*Fair Classification*
0.60 – 0.70	*Poor Classification*
0.50 – 0.60	*Failure*

Sumber: Luque, Carrasco, Martín, Heras (2019) [53]

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

Berdasarkan metodologi penelitian pada bab 3, maka pada bab ini akan dijelaskan implementasi metodologi penelitian yang dilakukan adalah sebagai berikut:

4.1. *Business Understanding*

Pada tahap ini dilakukan pemahaman terhadap objek penelitian. Pemahaman objek dilakukan dengan menggali data dan informasi melalui dataset pada *Bank marketing*. Tujuan dari penelitian ini adalah melakukan klasifikasi terhadap nasabah bank yang menabung deposito. Klasifikasi dilakukan dengan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression.* Diharapkan melalui penelitian ini bank dapat menerapkan strategi pemasaran yang tepat untuk meningkatkan loyalitas nasabah dan memperoleh nasabah baru.

4.2. *Data Understanding*

Tahap *data understanding* merupakan proses pemahaman pada dataset *Bank marketing*. Data ini mengacu pada kegiatan pemasaran langsung dari lembaga perbankan Portugis. Kegiatan pemasaran didasarkan melalui telepon. Untuk mengetahui nasabah menabung deposito "Ya" atau "Tidak". Dalam tahap ini peneliti memahami data dan menentukan model yang akan digunakan dalam penelitian. Berikut ini data yang digunakan sebagai penelitian.

A. Menganalisa dataset

Dalam menganalisa dataset perlu mengetahui data apa yang ada pada dataset. Sehingga dapat mengetahui proses selanjutnya yang harus dilakukan untuk mengolah data set tersebut dengan melihat atribut dan jenis datanya.

Pada Dataset *bank marketing* memiliki 41.188 data dengan 21 atribut. dengan satu target yaitu atribut ‘y’ dengan dua kelas yes dan no, berikut rincian dari dataset *bank marketing* pada Tabel 4.1.

Tabel 4.1 Dataset *Bank marketing*

No	Atribut	Keterangan	Type
1	*Age*	Usia Nasabah	Numerik
2	*Job*	jenis pekerjaan nasabah	Kategori
3	*Marital*	Status perkawinan	Kategori
4	*Education*	Menunjukkan tingkat pendidikan setiap pelanggan	Kategori
5	*Default*	Apakah pelanggan memiliki kredit secara *default*	Kategori
6	*Housing*	Apakah pelanggan memiliki pinjaman perumahan?	Kategori
7	*Loan*	Apakah pelanggan memiliki pinjaman pribadi	Kategori
8	*Contact*	Jenis komunikasi kontak seluler atau telepon	Kategori
9	*Month*	Kontak terakhir bulan	Kategori
10	*Day_of_week*	hari kontak terakhir dalam seminggu	Kategori
11	*Duration*	Durasi kontak terakhir dalam hitungan detik.	Numerik
12	*Campaign*	Jumlah kontak yang dilakukan untuk klien selama penawaran	Numerik
13	*Pdays*	jumlah hari yang berlalu setelah klien terakhir dihubungi dari kampanye	Numerik
14	*Previous*	jumlah kontak yang dilakukan sebelum kampanye ini dan untuk klien ini	Numerik
15	*Poutcome*	hasil dari penawaran sebelumnya	Kategori
16	*Emp.var.rate*	Tingkat variasi pekerjaan	Numerik
17	*Cons.price.idx*	hasil dari penawaran sebelumnya	Numerik
18	*Cons.conf.id*	Indeks kepercayaan konsumen	Numerik
19	*Euribor3m*	Tarif suku bunga euribor 3 bulan	Numerik
20	*Nr.employed*	Jumlah karyawan	Numerik
21	*y*	Apakah nasabah telah menabung deposito berjangka	*Binary*

B. Memvisualisasikan dataset

Agar lebih mudah dalam melihat dataset yang ada pada *bank marketing*, maka divisualisasikan dalam bentuk diagram untuk melihat isi dataset dalam setiap *feature*.

Sumber: Hasil Penelitian (2021)

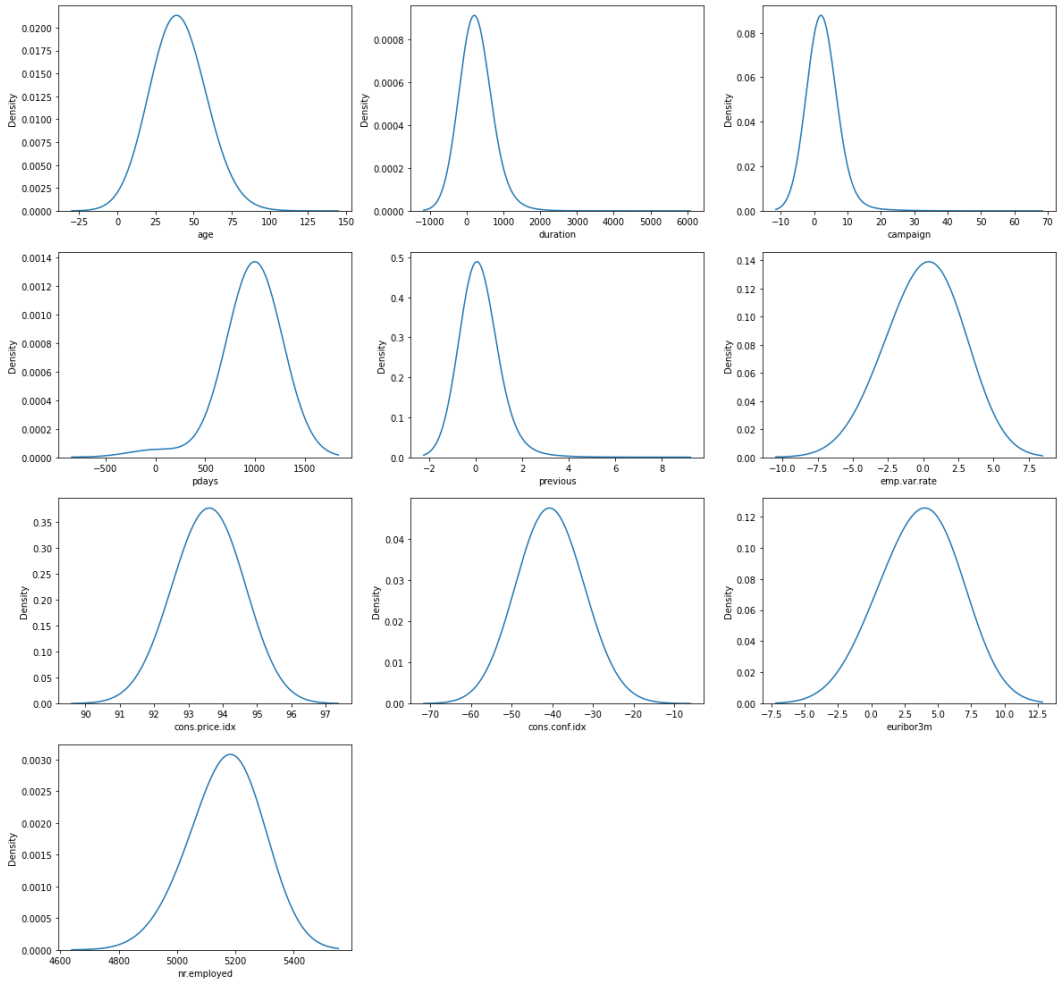
Gambar 4.1 Atribut tipe Kategori

Pada gambar 4.1 dapat dilihat isi data pada *feature* dengan tipe kategori yang ada pada *dataset Bank marketing*.

Sumber: Hasil Penelitian (2021)

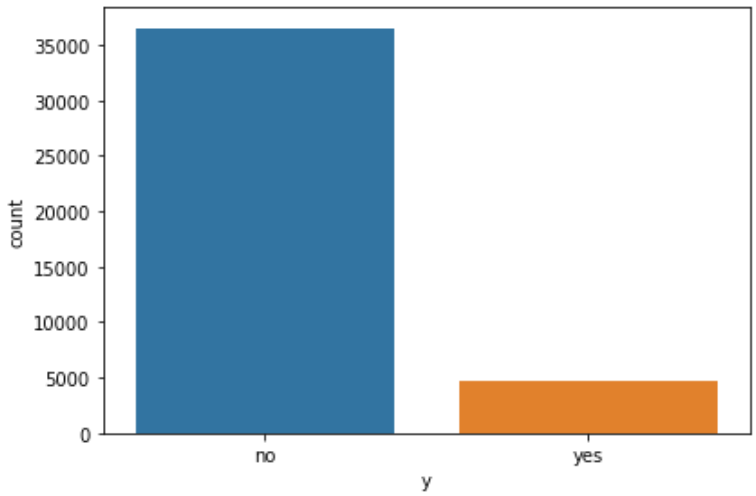
Gambar 4.2 Hubungan variabel kategori dengan target

Pada gambar 4.2 dapat dilihat isi data pada *feature* dengan tipe kategori yang ada pada dataset *Bank marketing*. Nasabah yang bekerja sebagai admin, teknisi, dan buruh lebih cenderung menabung deposito. Nasabah yang sudah menikah memiliki tingkat menabung deposito yang tinggi. Nasabah yang berpendidikan universitas cenderung menabung deposito. Nasabah yang tidak memiliki kredit default lebih cenderung menabung deposito. Selama bulan Mei hingga Agustus nasabah menunjukkan minat yang tinggi untuk menabung deposito. Nasabah yang memiliki pinjaman pribadi tampaknya kurang tertarik dengan deposito. Nasabah yang dihubungi melalui 'seluler' cenderung menabung deposito.



Sumber: Hasil Penelitian (2021)

Gambar 4.3 Atribut tipe numerik



Sumber: Hasil Penelitian (2021)

Gambar 4.4 Data target

Pada gambar 4.4 dapat dilihat data yang menjadi target yaitu atribut "y" memiliki jumlah data yang tidak sama antara "*yes*" dan "*no*". dengan jumlah "*yes*" ada 36548 sedangkan "*no*" ada 4640, artinya dataset *bank marketing* memiliki data yang tidak seimbang atau *imbalance* data.

4.3. *Data Preparation*

Pada tahap *data preparation* ini akan melakukan *data preprocessing* terlebih dahulu terhadap *dataset* sebelum digunakan dan masuk ke tahap *modelling*, bertujuan untuk mendapatkan data yang bersih dan siap untuk digunakan dalam penelitian agar nantinya data yang digunakan dapat lebih maksimal saat diterapkan dengan algoritma data mining.

4.3.1. *Data Preprocessing*

Pada tahap *preprocessing* digunakan untuk membersihkan data. Dan berikut data *preprocessing* yang dilakukan pada penelitian ini.

1. Cek dan mengatasi *Missing Value*

   Tahap ini digunakan untuk melihat apakah terdapat missing value pada dataset, missing value biasanya merupakan data yang hilang atau bernilai "0", jika ada maka perlu untuk diatasi.

Tabel 4.2 Cek *Missing Value*

No	**Atribut**	Missing Value/null
1	*Age*	0
2	*Job*	0
3	*Marital*	0
4	*Education*	0
5	*Default*	0
6	*Housing*	0
7	*Loan*	0
8	*Contact*	0
9	*Month*	0
10	*Day_of_week*	0
11	*Duration*	0
12	*Campaign*	0

13	*Pdays*	0
14	*Previous*	0
15	*Poutcome*	0
16	*Emp.var.rate*	0
17	*Cons.price.idx*	0
18	*Cons.conf.id*	0
19	*Euribor3m*	0
20	*Nr.employed*	0
21	*y*	0

Sumber: Hasil Penelitian (2021)

Berdasarkan hasil cek *missing value* yang dilakukan tidak terdapat *missing value* atau data yang bernilai 0 pada dataset. artinya dataset sudah bersih.

2. *Label Encoding*

Tahap *Label Encoding* digunakan untuk merubah tipe data kategori menjadi numerik. Untuk mengonversi label kata menjadi angka, kita perlu menggunakan generator *label encoding*. *Label Encoding* mengacu pada proses mengubah label kata untuk dijadikan bentuk numerik. Jadi jika data sudah menjadi numerik, maka kita dapat menggunakannya secara langsung untuk memulai pelatihan agar data dapat dihitung untuk melakukan prediksi.

Tabel 4.3 *Label Encoding*

No	**Atribut**	**Type**		**Atribut**	**Type**
1	*Age*	Int64		*Age*	Int64
2	*Job*	Object		*Job*	Int64
3	*Marital*	Object		*Marital*	Int64
4	*Education*	Object		*Education*	Int64
5	*Default*	Object		*Default*	Int64
6	*Housing*	Object	⇨	*Housing*	Int64
7	*Loan*	Object		*Loan*	Int64
8	*Contact*	Object		*Contact*	Int64
9	*Month*	Int64		*Month*	Int64
10	*Day_of_week*	Object		*Day_of_week*	Int64
11	*Duration*	Int64		*Duration*	Int64

12	*Campaign*	Int64		*Campaign*	Int64
13	*Pdays*	Int64		*Pdays*	Int64
14	*Previous*	Int64		*Previous*	Int64
15	*Poutcome*	Object		*Poutcome*	Int64
16	*Emp.var.rate*	Float64	⇨	*Emp.var.rate*	Float64
17	*Cons.price.idx*	Float64		*Cons.price.idx*	Float64
18	*Cons.conf.id*	Float64		*Cons.conf.id*	Float64
19	*Euribor3m*	Float64		*Euribor3m*	Float64
20	*Nr.employed*	Float64		*Nr.employed*	Float64
21	*y*	Object		*y*	Int64

Sumber: Hasil Penelitian (2021)

Dari Tabel 4.3 *Label Encoding* dapat dilihat semua tipe data dari dataset *bank marketing* sudah menjadi numerik.

3. Menghapus atribut *Pdays*

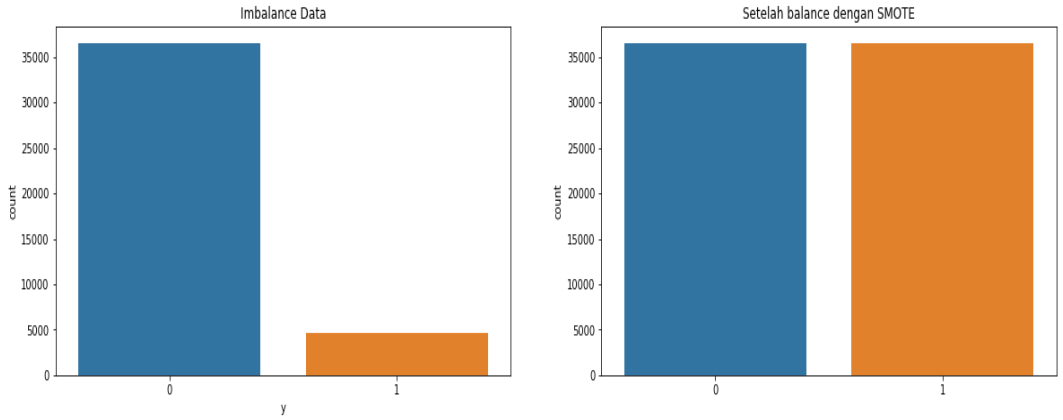
Atribut Pdays tidak digunakan dan dihapus karena memiliki nilai 999 yang artinya nasabah tidak dihubungi sekitar 90%.

4. Mengatasi *Imbalance* data

Berdasarkan Analisa yang dilakukan diketahui bahwa dataset yang digunakan merupakan data tidak seimbang atau *imbalance* data. Untuk mengatasi hal tersebut maka data yang tidak seimbang dilakukan *oversampling* dengan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada tahap *training*.

4.4. *Modelling*

Tahap M*odelling* adalah pemilihan teknik data mining dengan menentukan algoritma yang digunakan. Dalam tahap *Modelling* ini dilakukan teknik pengklasifikasian data dengan algoritma *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti *Decision Tree, K-Nearest Neighbor* dan *Logistic Regression* dengan penerapan *split training testing* 80:20 dimana dari 41188 dibagi dua yaitu 80% (32950) data *training* dan 20% (8238) data *testing.* Pada data *training* dilakukan *oversampling* dengan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE).*



Sumber: Hasil Penelitian (2021)

Gambar 4.5 Mengatasi *Imbalance* data

Pada Gambar 4.5 menunjukkan hasil data asli atau yang tidak seimbang (*imbalance data*) dengan data yang seimbang (*balance*) yang sudah di *oversampling* dengan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*). Dimana terdapat 32950 data *training* yaitu:

Sebelum *oversampling*: jumlah label ‘1’ = 3721

Sebelum *oversampling*: jumlah label ‘0’ = 29229

Setelah *oversampling*: jumlah label ‘1’ = 29229

Setelah *oversampling*: jumlah label ‘0’ = 29229

Jadi setelah di *oversampling* dengan teknik *SMOTE* pada data *training*, jumlah data pada atribut “y” sudah seimbang. Kemudian dioptimalkan dengan *hyperparameter tuning* menggunakan GridsearchCV untuk menentukan parameter yang digunakan.

4.4.1. *Extreme Gradient Boosting (XGBoost)*

Pada model yang dibangun dengan algoritma klasifikasi *XGBoost* menerapkan *split training testing* 80:20 dimana 80 % data *training* dan 20% data *testing*, kemudian model yang dibangun dioptimalkan menggunakan GridsearchCV dengan cv=10, artinya data dibagi menjadi 10 bagian yaitu 9 data *training* dan 1 data *testing* dan dilakukan 10 kali dengan parameter objective='binary: logistic', gamma=0, learning_rate=0.1, max_depth=6, n_estimators=200, min_child_weight=5 dan memvariasikannya seperti min_child_weight: [1, 3, 5, 10]

dan max_depth: [3, 5, 7, 9] yang diujikan untuk mendapatkan *best score* dengan parameter terbaik.

4.4.2. *Decision Tree*

Pada model yang dibangun dengan algoritma klasifikasi *Decision Tree* menerapkan *split training testing* 80:20 dimana 80 % data *training* dan 20% data *testing*, kemudian model yang dibangun dioptimalkan menggunakan GridsearchCV dengan cv=10, artinya data dibagi menjadi 10 bagian yaitu 9 data *training* dan 1 data *testing* dan dilakukan 10 kali dengan parameter criterion='gini', splitter='best', max_depth=22, min_samples_split=3, min_samples_leaf=1, random_state=0 dan memvariasikannya seperti criterion: ['gini', 'entropy'], splitter: ['best', 'random'], max_depth: [20, 22, 28, 32, 37, 38, 42, 45, 50, 70], min_samples_leaf: [1, 2, 3, 4, 5] yang diujikan untuk mendapatkan *best score* dengan parameter terbaik.

4.4.3. *K-Nearest Neighbor*

Pada model yang dibangun dengan algoritma klasifikasi *K-Nearest Neighbor* menerapkan *split training testing* 80:20 dimana 80 % data *training* dan 20% data *testing*, kemudian model yang dibangun dioptimalkan menggunakan GridsearchCV dengan cv=10, artinya data dibagi menjadi 10 bagian yaitu 9 data *training* dan 1 data *testing* dan dilakukan 10 kali dengan parameter n_neighbors=6, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=-1 dan memvariasikannya seperti n_neighbors: [1, 2, 3, 4, 5, 6, 7, 8], weights: ["uniform", "distance"], leaf_size: [10, 20, 30, 40] yang diujikan untuk mendapatkan *best score* dengan parameter terbaik.

4.4.4. *Logistic Regression*

Pada model yang dibangun dengan algoritma klasifikasi *logistic Regression* menerapkan *split training testing* 80:20 dimana 80 % data *training* dan 20% data *testing*, kemudian model yang dibangun dioptimalkan menggunakan GridsearchCV dengan cv=10, artinya data dibagi menjadi 10 bagian yaitu 9 data *training* dan 1 data *testing* dan dilakukan 10 kali dengan parameter penalty='l2', C=1.0, random_state=None, solver='lbfgs', max_iter=150, multi_class='auto', verbose=0

dan memvariasikannya seperti C: [1.0, 2.0, 3.0, 4.0, 5.0, 6.0], max_iter: [10, 100, 1000], penalty: ['l1', 'l2'] multi_class: ['auto', 'ovr', 'multinomial'] yang diujikan untuk mendapatkan *best score* dengan parameter terbaik.

4.5. *Evaluation*

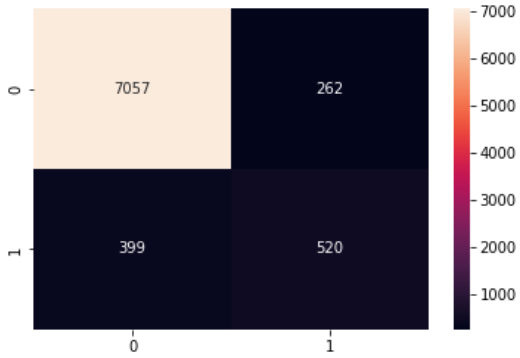
Dari hasil *modeling* yang telah dilakukan sebelumnya didapat *best score* dan parameter terbaik yang kemudian diterapkan untuk menghasilkan *Confusion Matrix* dan Kurva ROC dari algoritma klasifikasi *XGBoost, Decision Tree, KNN* dan *Logistic Regression* untuk dibandingkan hasil akurasi yang telah didapat.

4.5.1. *Extreme Gradient Boosting (XGBoost)*

Pada algoritma *Extreme Gradient Boosting (XGBoost)* didapatkan *best score* 0.920 dengan parameter max_depth= 3 dan min_child_weight= 5. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *XGBoost* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *Extreme Gradient Boosting (XGBoost). Confusion Matrix* menunjukkan nilai *accuracy, precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.6 *Confusion matrix XGBoost*

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma klasifikasi *Extreme Gradient Boosting (XGBoost).*

True Positif: bernilai 7057

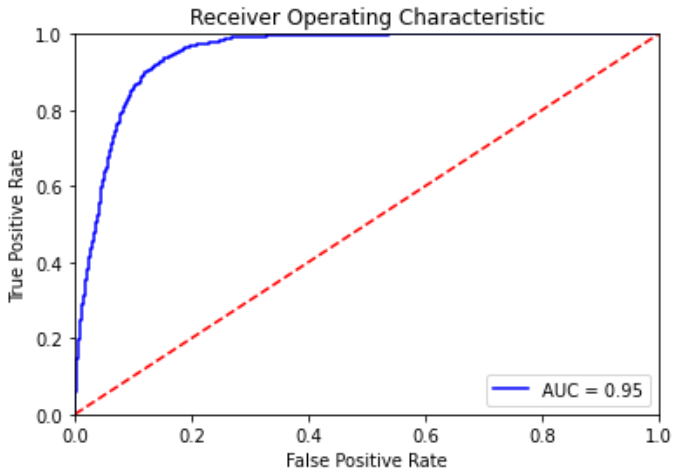
True Negatif: bernilai 520

False Positif: bernilai 262

False Negatif: bernilai 399

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *XGBoost*



Sumber: Hasil Penelitian (2021)

Gambar 4.7 ROC *XGBoost*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0.95 dimana diagnosa hasilnya adalah ***Excellent Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.4. Hasil Evaluasi *XGBoost*

Evaluasi	Hasil
Accuracy	92%
Precision	96%
Recall	94%
F1 Score	95%

AUC	0.95

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7057+262}{7057+262+520+399} = \mathbf{0.919}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7057}{7057+262} = \mathbf{0.964}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7057}{7057+399} = \mathbf{0.946}$$

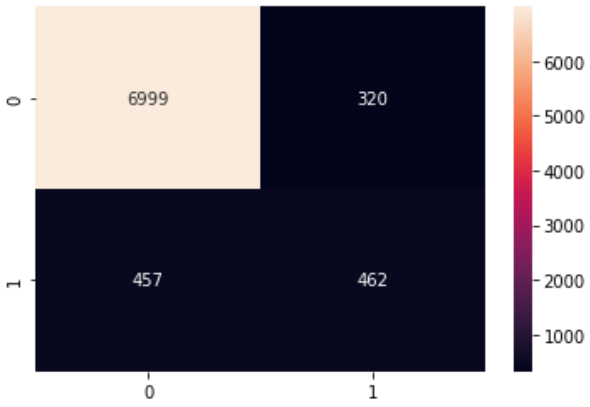
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}964\ x\ 0{,}946}{0{,}964\ x\ 0{,}946} = \mathbf{0.955}$$

4.5.2. *Decision Tree*

Pada algoritma *Decision Tree* didapatkan *best score* 0.904 dengan parameter criterion= gini, max_depth= 22, min_samples_leaf= 4, splitter= random. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *Decision Tree* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *Decision Tree.*



Sumber: Hasil Penelitian (2021)

Gambar 4.8 *Confusion matrix Decision Tree*

Gambar diatas menunjukkan hasil *confusion Matrix* dari algoritma klasifikasi *Decision Tree*.

True Positif: bernilai 6999

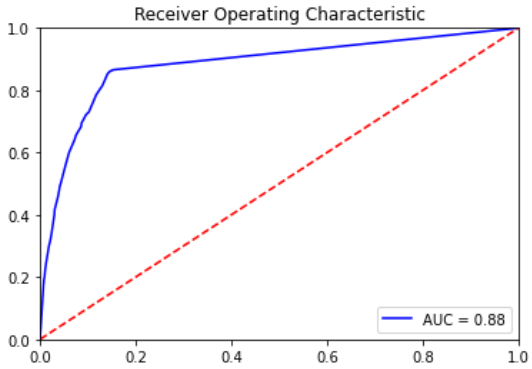
True Negatif: bernilai 462

False Positif: bernilai 320

False Negatif: bernilai 457

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *Decision Tree*



Sumber: Hasil Penelitian (2021)

Gambar 4.9 ROC *Decision Tree*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0,88 dimana diagnosa hasilnya adalah ***Good Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.5. Hasil Evaluasi *Decision Tree*

Evaluasi	Hasil
Accuracy	91%
Precision	96%
Recall	94%
F1 Score	95%
AUC	0.88

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{6999+462}{6999+462+320+457} = \mathbf{0.905}$$

$$Precision = \frac{TP}{TP+FP} = \frac{6999}{6999+320} = \mathbf{0.956}$$

$$Recall = \frac{TP}{TP+TN} = \frac{6999}{6999+462} = \mathbf{0.938}$$

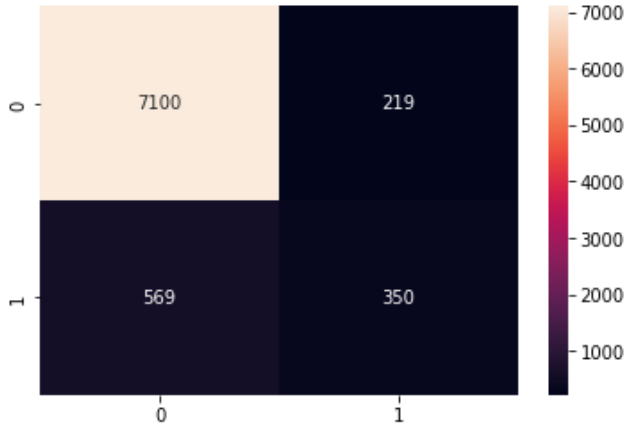
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}956\ x\ 0{,}938}{0{,}956\ x\ 0{,}938} = \mathbf{0.947}$$

4.5.3. *K-Nearest Neighbor (KNN)*

Pada algoritma KNN didapatkan *best score* 0.912 dengan parameter leaf_size= 10, n_neighbors= 8, weights= uniform. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *KNN* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *KNN*. Pada *Confusion Matrix* menunjukkan nilai *accuracy, precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.10 *Confusion matrix* KNN

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma klasifikasi KNN.

True Positif: bernilai 7100

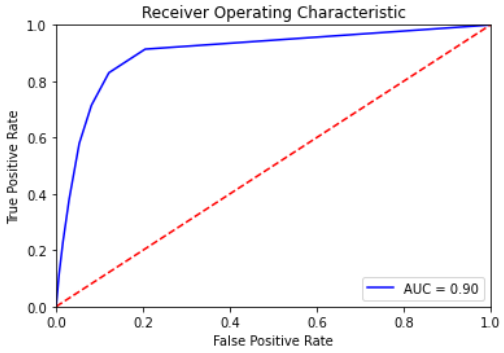
True Negatif: bernilai 350

False Positif: bernilai 219

False Negatif: bernilai 569

2. Kurva ROC

   Berikut ini merupakan Kurva ROC dari algoritma KNN



Sumber: Hasil Penelitian (2021)

Gambar 4.11 ROC KNN

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0,90 dimana diagnosa hasilnya adalah ***Excellent Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.6. Hasil Evaluasi KNN

Evaluasi	**Hasil**
Accuracy	90%
Precision	97%
Recall	92%
F1 Score	95%
AUC	0.90

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7100+350}{7100+350+219+569} = \mathbf{0.903}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7100}{7100+219} = \mathbf{0.978}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7100}{7100+569} = \mathbf{0.918}$$

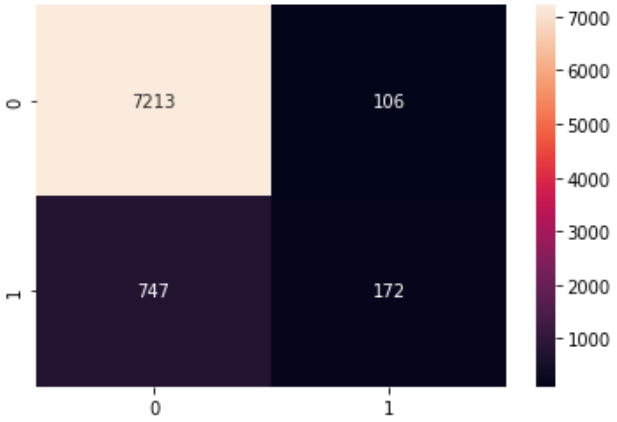
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}978\ x\ 0{,}918}{0{,}978+0{,}918} = \mathbf{0.947}$$

4.5.4. Logistic Regression

Pada algoritma *Logistic Regression* didapatkan *best score* 0.921 dengan parameter C= 3.0, max_iter =10, multi_class= auto, penalty =l2. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma *Logistic Regression* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah menampilkan *Confusion Matrix* menggunakan algoritma *Logistic Regression*. Pada *Confusion Matrix* menunjukkan nilai *accuracy*, *precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.12 *Confusion matrix Logistic Regression*

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma *Logistic Regression*.

True Positif: bernilai 7213

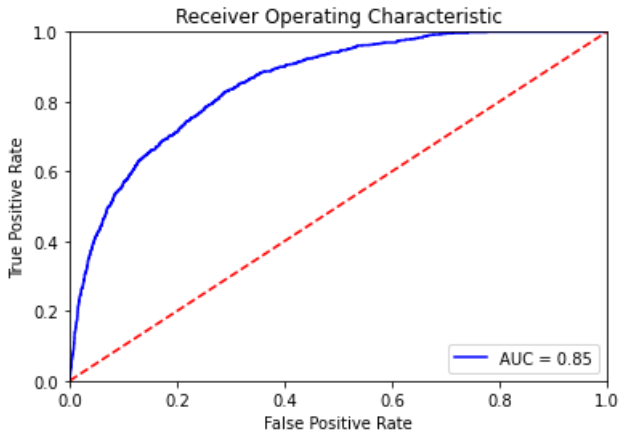
True Negatif: bernilai 172

False Positif: bernilai 106

False Negatif: bernilai 747

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *Logistic Regression*



Sumber: Hasil Penelitian (2021)

Gambar 4.13. ROC *Logistic Regression*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0,85 dimana diagnosa hasilnya adalah ***Good Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.7. Hasil Evaluasi *Logistic Regression*

Evaluasi	**Hasil**
Accuracy	91%
Precision	95%
Recall	90%
F1 Score	94%
AUC	0.93

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7213+172}{7213+172+106+147} = \mathbf{0.896}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7213}{7213+106} = \mathbf{0.985}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7213}{7213+747} = \mathbf{0.906}$$

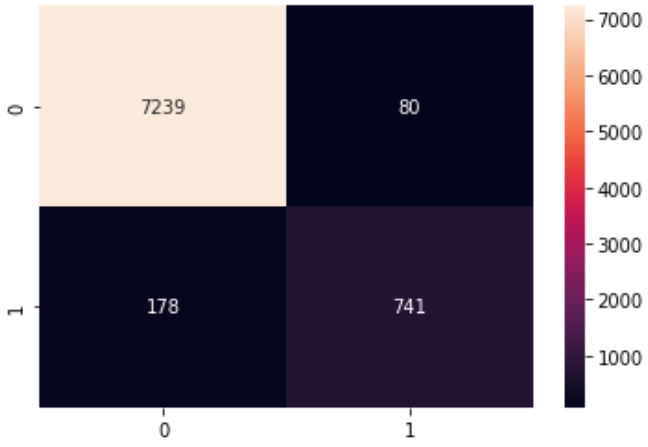
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}985x\ 0{,}906}{0{,}985x\ 0{,}906} = \mathbf{0.951}$$

4.5.5. *XGBoost – SMOTE*

Pada algoritma *XGBoost–SMOTE* didapatkan *best score* 0.953 dengan parameter max_depth= 7, min_child_weight= 1. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *Extreme Gradient Boosting (XGBoost) – SMOTE* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini pada Tabel 4.4 adalah menampilkan *Confusion Matrix* menggunakan algoritma *Extreme Gradient Boosting (XGBoost) – SMOTE* dengan. Pada *Confusion Matrix* menunjukkan nilai *accuracy*, *precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.14. *Confusion matrix XGBoost- SMOTE*

Gambar diatas menunjukkan hasil *confusion matrix* dari algoritma klasifikasi *Extreme Gradient Boosting – SMOTE*.

True Positif: bernilai 7239

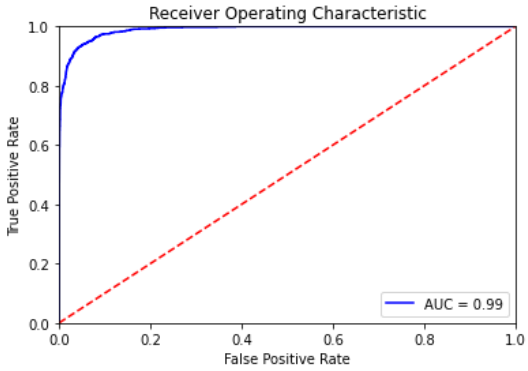
True Negatif: bernilai 741

False Positif: bernilai 80

False Negatif: bernilai 178

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *XGBoost- SMOTE*



Sumber: Hasil Penelitian (2021)

Gambar 4.15. ROC *XGBoost-SMOTE*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0, 99 dimana diagnosa hasilnya adalah ***Excellent Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.8 Hasil Evaluasi *XGBoost- SMOTE*

Evaluasi	**Hasil**
Accuracy	97%
Precision	98%
Recall	97%
F1 Score	95%

AUC	0.99

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7239+741}{7239+6995+326+354} = \mathbf{0.968}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7239}{7239+80} = \mathbf{0.989}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7239}{7239+178} = \mathbf{0.976}$$

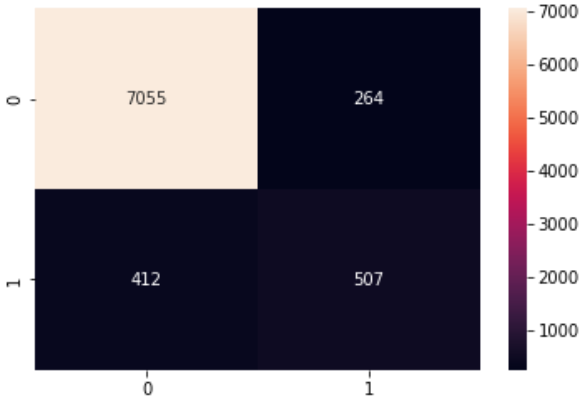
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}955\ x\ 0{,}951}{0{,}955\ +0{,}951} = \mathbf{0.982}$$

4.5.6. *Decision Tree – SMOTE*

Pada algoritma *Decision Tree-SMOTE* didapatkan *best score* 0.937 dengan parameter criterion= entropy, max_depth= 22, min_samples_leaf= 1, splitter= best. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *Decision Tree-SMOTE* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *Decision Tree-SMOTE*.



Sumber: Hasil Penelitian (2021)

Gambar 4.16. *Confusion matrix Decision Tree-SMOTE*

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma klasifikasi *Decision Tree-SMOTE.*

True Positif: bernilai 7055

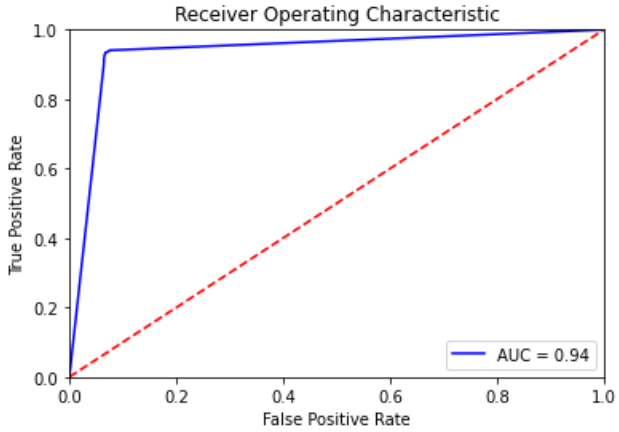
True Negatif: bernilai 7055

False Positif: bernilai 246

False Negatif: bernilai 412

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *Decision Tree-SMOTE*



Sumber: Hasil Penelitian (2021)

Gambar 4.17. ROC *Decision Tree-SMOTE*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0, 93 dimana diagnosa hasilnya adalah ***Excellent Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.9. Hasil Evaluasi *Decision Tree-SMOTE*

Evaluasi	**Hasil**
Accuracy	92%
Precision	96%
Recall	94%

F1 Score	95%
AUC	0.94

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7055+507}{7055+507+264+412} = \mathbf{0.917}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7055}{7055+264} = \mathbf{0.963}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7055}{7055+412} = \mathbf{0.944}$$

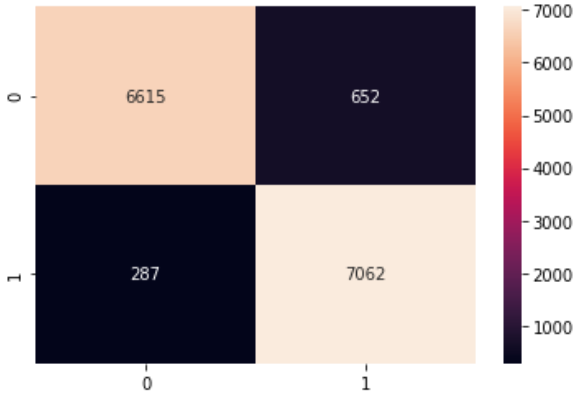
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}932\ x\ 0{,}932}{0{,}932\ +0{,}932} = \mathbf{0.954}$$

4.5.7. *KNN - SMOTE*

Pada algoritma *KNN-SMOTE* didapatkan *best score* 0.946 dengan parameter leaf_size= 10, n_neighbors= 2, weights= uniform. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma klasifikasi *KNN-SMOTE* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *KNN-SMOTE* dengan. Pada *Confusion Matrix* menunjukkan nilai *accuracy*, *precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.18 *Confusion matrix KNN-SMOTE*

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma klasifikasi *KNN-SMOTE*.

True Positif: bernilai 6890

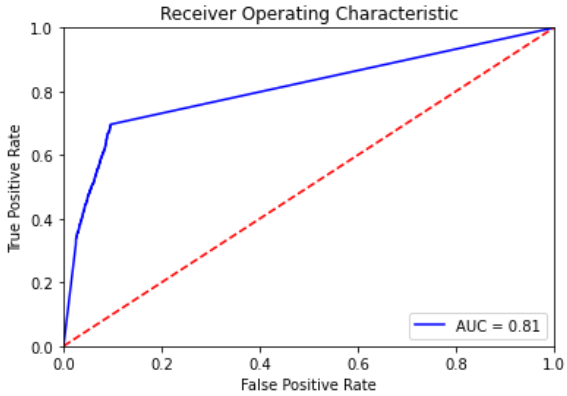
True Negatif: bernilai 466

False Positif: bernilai 429

False Negatif: bernilai 453

2. Kurva ROC

Berikut ini merupakan Kurva ROC dari algoritma *KNN-SMOTE*



Sumber: Hasil Penelitian (2021)

Gambar 4.19 ROC *KNN-SMOTE*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0, 81 dimana diagnosa hasilnya adalah ***Good Classification.***

3. Hasil

Dari *confusion matrix* dan ROC kita bisa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.10. Hasil Evaluasi *XGBoost*

Evaluasi	**Hasil**
Accuracy	89%
Precision	94%
Recall	94%
F1 Score	94%
AUC	0.81

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{6890+466}{6890+466+429+453} = \mathbf{0.892}$$

$$Precision = \frac{TP}{TP+FP} = \frac{6890}{6890+429} = \mathbf{0.943}$$

$$Recall = \frac{TP}{TP+TN} = \frac{6890}{6890+453} = \mathbf{0.938}$$

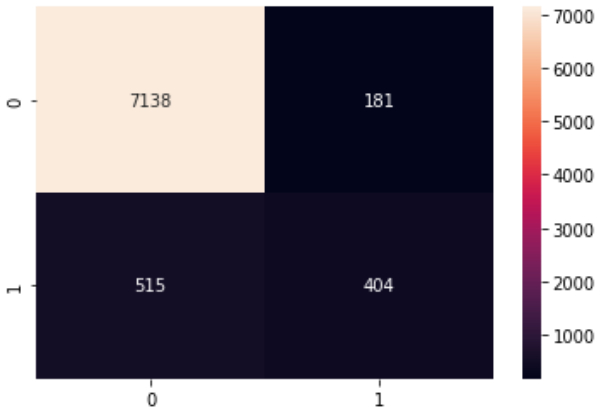
$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}943\ x\ 0{,}938}{0{,}943\ x\ 0{,}938} = \mathbf{0.939}$$

4.5.8. *Logistic Regression – SMOTE*

Pada algoritma *Logistic Regression-SMOTE* didapatkan *best score* 0.860 dengan parameter C= 3.0, max_iter= 1000, multi_class= auto, penalty= l2. Berikut adalah *Confusion Matrix* dan Kurva ROC algoritma *Logistic Regression-SMOTE* yang dihasilkan dari penelitian yang dilakukan:

1. *Confusion Matrix*

   Confusion Matrix memberikan sebuah penilaian *performance* klasifikasi berdasarkan objek *true* dan *false*. *Confusion Matrix* berisi terkait informasi aktual dan prediksi pada sistem klasifikasi. Berikut dibawah ini menampilkan *Confusion Matrix* menggunakan algoritma *Logistic Regression-SMOTE.* Pada *Confusion Matrix* menunjukkan nilai *accuracy*, *precision* dan *recall*.



Sumber: Hasil Penelitian (2021)

Gambar 4.20. *Confusion matrix Logistic Regression-SMOTE*

Gambar diatas menunjukkan hasil *Confusion Matrix* dari algoritma klasifikasi *Logistic Regression-SMOTE.*

True Positif: bernilai 7138

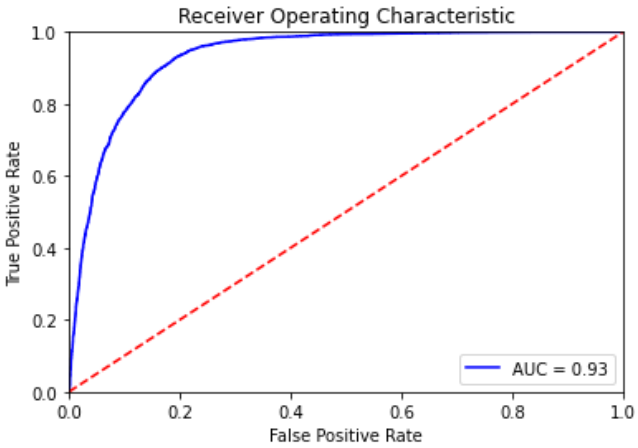
True Negatif: bernilai 404

False Positif: bernilai 181

False Negatif: bernilai 515

2. Kurva ROC

   Berikut ini merupakan Kurva ROC dari algoritma *Logistic Regression-SMOTE*



Sumber: Hasil Penelitian (2021)

Gambar 4.21 ROC *Logistic Regression-SMOTE*

Kurva ROC dengan nilai AUC (*Area Under Curve*) yang dihasilkan dari gambar diatas sebesar 0, 93 dimana diagnosa hasilnya adalah ***Excellent Classification.***

3. Hasil

Dari *confusion matrix* dan ROC 55itab isa melihat hasil keseluruhan dari tabel berikut:

Tabel 4.11. Hasil Evaluasi *Logistic Regression-SMOTE*

Evaluasi	Hasil
Accuracy	92%
Precision	98%
Recall	93%
F1 Score	95%

AUC	0.93

Sumber: Hasil Penelitian (2021)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7138+404}{7138+404+181+515} = \mathbf{0.915}$$

$$Precision = \frac{TP}{TP+FP} = \frac{7138}{7138+181} = \mathbf{0.975}$$

$$Recall = \frac{TP}{TP+TN} = \frac{7138}{7138+515} = \mathbf{0.932}$$

$$F1\ Score = \frac{2\ x\ precision\ x\ Recall}{precision+\ Recall} = \frac{2\ x\ 0{,}975\ x\ 0{,}932}{0{,}975\ x\ 0{,}932} = \mathbf{0.953}$$

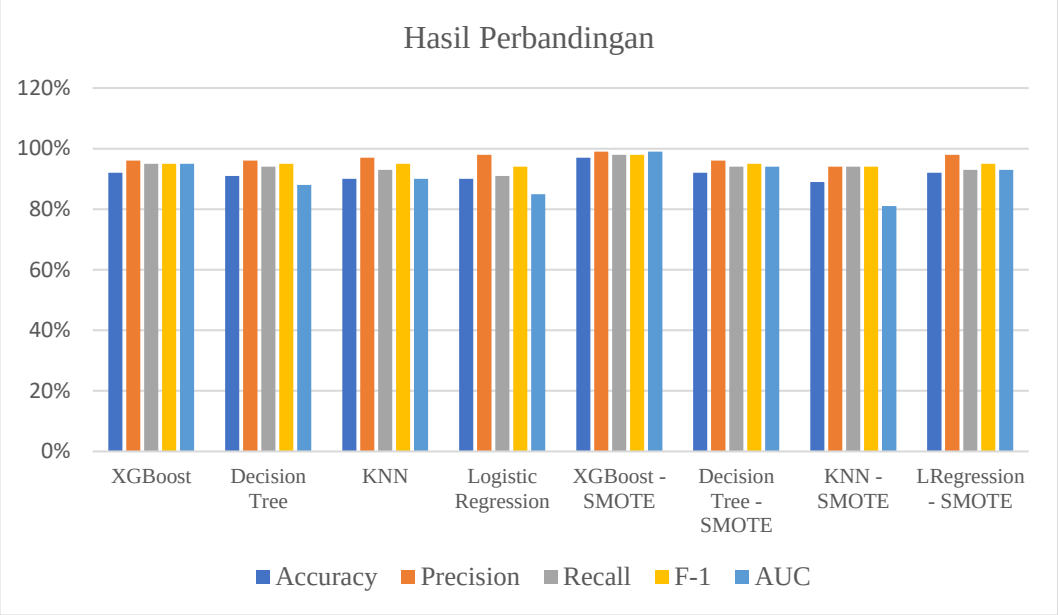
4.5.9. Rangkuman Hasil

Adapun perbandingan rangkuman hasil *accuracy*, *precision*, *recall* dan AUC Algoritma klasifikasi *Extreme Gradient Boosting (XGBoost)* yang dikomparasi dengan algoritma lain seperti, *Decision Tree*, *K-Nearest Neighbor* dan *Logistic Regression* yang dibedakan menjadi dua data tidak seimbang dan data seimbang sebagai berikut:

Tabel 4.12 Perbandingan hasil penelitian

Algoritma	***Accuracy***	***Precision***	***Recall***	***F-1***	***AUC***
XGBoost	92%	96%	95%	95%	0.95
Decision Tree	91%	96%	94%	95%	0.88
KNN	90%	97%	93%	95%	0.90
Logistic Regression	90%	98%	91%	94%	0.85
XGBoost - SMOTE	**97%**	99%	98%	98%	**0.99**
Decision Tree - SMOTE	92%	96%	94%	95%	0.94
KNN - *SMOTE*	89%	94%	94%	94%	0.81
L Regression - *SMOTE*	92%	98%	93%	95%	0.93

Sumber: Hasil penelitian (2021)



Sumber: Hasil penelitian (2021)

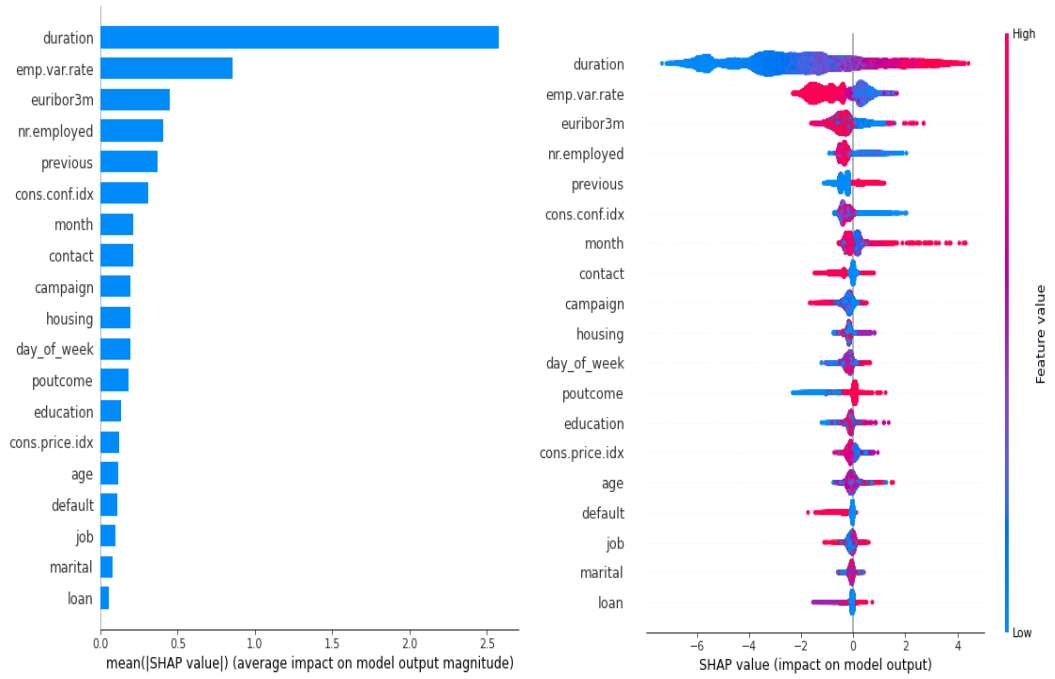
Gambar 4.22 Gambar Hasil perbandingan

Berdasarkan Tabel 4.22 Hasil perbandingan perhitungan dengan algoritma *XGBoost* mendapatkan nilai akurasi 92% dengan AUC 0.95, perhitungan dengan algoritma *Decision Tree* mendapatkan nilai akurasi 91% dengan AUC 0.88, perhitungan dengan algoritma KNN mendapatkan nilai akurasi 90% dengan AUC 0.90, perhitungan dengan algoritma *Logistic Regression* mendapatkan nilai akurasi 90% dengan AUC 0.85, Sedangkan perhitungan dengan algoritma *XGBoost – SMOTE* mendapatkan nilai akurasi 97% dengan AUC 0.99, perhitungan dengan algoritma *Decision Tree – SMOTE* mendapatkan nilai akurasi 92% dengan AUC 0.92, perhitungan dengan algoritma KNN – *SMOTE* mendapatkan nilai akurasi 89% dengan AUC 0.81, perhitungan dengan algoritma *Logistic Regression – SMOTE* mendapatkan nilai akurasi 92% dengan AUC 0.93.

Dapat disimpulkan dari hasil komparasi diatas algoritma *XGBoost-SMOTE* mendapat akurasi tertinggi dibanding algoritma lainnya berdasarkan nilai akurasi 97% dan AUC 0.99 dan *best score* 0.953 dengan parameter max_depth= 7, min_child_weight= 1. Kemudian berdasarkan hasil tersebut alhoritma *XGBoost-SMOTE* akan diterapkan pada *feature important* dengan *SHAPE value* untuk melihat atribut apa yang paling berkontribusi terhadap model yang dibangun.

4.5.10. *Feature Important* dengan *SHAP Value*

Untuk menginterpretasikan model yang diimplementasikan pada *XGBoost* dengan baik, penelitian ini menerapkan SHAP *value*, hal ini dilakukan untuk mengetahui *feature* yang paling mempengaruhi akurasi terhadap model yang telah dibuat berdasarkan tingkatan atau ranking, terutama pada algoritma *XGBoost* yang memiliki hasil akurasi tertinggi dari algoritma lainya. Berikut adalah *feature important* berdasarkan hasil penelitian yang sudah dilakukan:



Sumber: Hasil penelitian (2021)

Gambar 4.23 *SHAPE value*

Berdasarkan gambar SHAP *Value* yang dihasilkan dengan algoritma *XGBoost – SMOTE* yang merupakan algoritma dengan akurasi terbaik. *feature* diurutkan berdasarkan besaran efek terhadap model yang dibangun. Pertama dapat disimpulkan bahwa kontribusi fitur bervariasi dan berbeda, dengan beberapa fitur spesifik yang berkontribusi jauh lebih banyak daripada fitur lainnya. Seperti *feature* duration yang lebih mendominasi dari *feature* lainya. Kemudian diketahui bahwa dalam beberapa *feature* berkontribusi sangat sedikit pada keluaran model, seperti loan, merital, job.

Berdasarkan hal tersebut diatas disimpulkan bahwa ada tingkat tertentu hubungan linier antara kontribusi beberapa fitur dan nilainya. Dapat mengambil contoh *feature* yang paling berkontribusi pada model yaitu *duration,* hal ini dapat diartikan bahwa semakin lama durasi nasabah berbicara dengan pegawai *bank marketing* untuk menawarkan deposito, artinya ada ketertarikan nasabah untuk mencari tahu info lebih detail terkait produk yang ditawarkan, sebaliknya semakin sedikit durasi komunikasi antara nasabah dengan *customer service,* maka potensi untuk menabung deposito rendah karena tidak tertarik dengan penawaran yang disampaikan.

Maka strategi dalam mengatasi hal tersebut pihak bank harus melatih *customer service* terutama yang berada dalam divisi *marketing* untuk menguasai *public speaking* dan *product knowledge* atau penguasaan pengetahuan produk yang baik khususnya deposito. Sehingga nasabah yang ditawarkan deposito akan tertarik dan akan menabung deposito karena penyampaian informasi yang lengkap, jelas dan mudah dipahami serta memiliki solusi dari pertanyaan atau masalah yang disampaikan oleh calon nasabah deposito.

BAB V
PENUTUP

5.1. Kesimpulan

Berdasarkan dari penelitian yang telah dilakukan maka dapat diambil beberapa kesimpulan yaitu penerapan data mining dapat digunakan untuk memprediksi calon nasabah yang akan menabung deposito. Dataset yang digunakan merupakan dataset tidak seimbang (*imbalance data*), kemudian data yang tidak seimbang diolah dengan *oversampling* menggunakan teknik *SMOTE* pada data *training*, dan dari data yang tidak seimbang dan seimbang dimodelkan dengan algoritma *Extreme Gradient Boosting (XGBoost) Decision Tree, K-Nearest Neighbor* dan *Logistic Regression*. Menunjukan bahwa data yang melalui proses *oversampling* dengan *SMOTE* mengalami peningkatan serta memiliki akurasi dan AUC lebih tinggi.

Pada semua model yang diuji algoritma *XGBoost-SMOTE* memiliki hasil paling tinggi yaitu akurasi 97% dan AUC 99% termasuk kategori *Excellent Classification*. Hal ini menunjukan bahwa *XGBoost-SMOTE* mampu memprediksi nasabah yang akan menabung deposito di perbankan lebih baik dibanding algoritma lain yang diujikan.

Berdasarkan *Feature Important* dari model yang dibangun dengan algoritma *XGBoost-SMOTE* menggunakan SHAP *Value* didapatkan informasi bahwa *feature "duration"* memiliki kontribusi lebih terhadap model yang dibangun dibandingkan dengan *feature* yang lain. Dapat disimpulkan bahwa ketertarikan nasabah terhadap produk yang ditawarkan terutama deposito melalui panggilan telepon dapat dipengaruhi oleh nasabah yang menggali informasi lebih saat panggilan berlangsung. Artinya strategi yang harus dilakukan oleh perbankan dalam melakukan penawaran produk secara langsung harus melatih *customer service* terutama yang berada dalam divisi *marketing* untuk menguasai *public speaking* yang bagus dan *product knowledge* atau penguasaan produk yang baik. Sehingga nasabah yang ditawarkan deposito akan tertarik dan akan menabung deposito karena penyampaian yang jelas dan mudah dipahami serta informasi yang lengkap.

5.2. Saran

Berdasarkan hasil penelitian dapat diberikan beberapa saran yang dapat dilakukan untuk penelitian selanjutnya sebagai berikut:

1. Penelitian selanjutnya dapat menerapkan metode *over sampling* yang berbeda dalam mengatasi *imbalance* data agar mendapatkan hasil yang lebih maksimal.
2. Penelitian selanjutnya dapat menggunakan algoritma yang lain serta dapat juga menggunakan algoritma *Deep Learning*.
3. Penelitian selanjutnya dapat memvariasikan lebih banyak *hyperparameter* yang diuji untuk melihat performa model yang lebih baik.
4. Diharapkan untuk pihak yang berkaitan dengan *bank marketing* ataupun pemasaran lain yang dilakukan secara langsung, untuk bisa meningkatkan kualitas layanan dan penyampaian informasi yang jelas untuk bisa menarik pelanggan.

DAFTAR REFERENSI

[1] T. Online *et al.*, “ANALISIS FATWA DSN MUI TENTANG DEPOSITO DITINJAU DARI ASPEK USHUL FIQH Oleh : Kartini Kata Kunci : analisis fatwa dsn MUI , deposito , Aspek ushul fiqh Pendahuluan Sebagaimana syariah dewasa diketehui ini bank menggunakan sistim dan operasi nya berdasarka,” vol. 1, no. 1, 2021.

[2] N. 10 T. 1998 UU RI, *Undang-Undang RI No. 10 Tahun 1998 tentang Perbankan*. 1998, p. 182.

[3] R. Ula, “PENGARUH CAPITAL ADEQUACY RATIO (CAR), INFLASI , DAN SUKU BUNGA SERTIFIKAT BANK INDONESIA (SBI) TERHADAP TINGKAT SUKU BUNGA DEPOSITO BERJANGKA (Studi Pada Perusahaan Bank Pembangunan Daerah di Indonesia Periode 2010-2015),” vol. 56, no. 1, 2018.

[4] S. M. Kostic, M. Duricic, M. I. Simic, and M. V. Kostic, “Data Mining and Modeling use Case in Banking Industry,” *2018 26th Telecommun. Forum, TELFOR 2018 - Proc.*, pp. 1–4, 2018, doi: 10.1109/TELFOR.2018.8611897.

[5] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Penerbit Andi, 2020.

[6] J. Suntoro, *Data Mining: Algoritma dan Implementasi dengan Pemrograman PHP*. 2019.

[7] P. R. S. Moro,P. Cortez, “Bank Marketing Data Set,” *2012*, 2012. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[8] J. SUNTORO, “DATA MINING : Algoritma dan Implementasi dengan Pemrograman php.” p. 179, 2019.

[9] Sudirman, A. P. Windarto, and A. Wanto, “Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.

[10] P. Wongchinsri and W. Kuratach, “A survey - Data mining frameworks in credit card processing,” *2016 13th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2016*, 2016, doi:

10.1109/ECTICon.2016.7561287.

[11] A. Hemeida, R. Mansour, and M. E. Hussein, “Multilevel Thresholding for Image Segmentation Using an Improved Electromagnetism Optimization Algorithm,” *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 4, p. 102, 2019, doi: 10.9781/ijimai.2018.09.001.

[12] L. Muflikhah, D. E. Ratnawati, and R. R. M. Putri, *Buku Ajar Data Mining*, Cetakan pe. Malang: Universitas Brawijaya Press, 2018.

[13] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, San Fr. CA, USA*, pp. 785–794, 2016.

[14] H. Shi, H. Wang, Y. Huang, L. Zhao, C. Qin, and C. Liu, “A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification,” *Comput. Methods Programs Biomed.*, vol. 171, pp. 1–10, 2019, doi: 10.1016/j.cmpb.2019.02.005.

[15] D. Yu *et al.*, “Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier,” *Thorac. Cancer*, vol. 11, no. 1, pp. 95–102, 2020, doi: 10.1111/1759-7714.13204.

[16] C. Hu, J. Yan, and C. Wang, “Advanced Cyber-Physical Attack Classification with Extreme Gradient Boosting for Smart Transmission Grids,” *IEEE Power Energy Soc. Gen. Meet.*, vol. 2019-Augus, 2019, doi: 10.1109/PESGM40551.2019.8973679.

[17] M. Lin, X. Zhu, T. Hua, X. Tang, G. Tu, and X. Chen, “Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique,” *Remote Sens*, pp. 1–22, 2021, doi: https://doi.org/10.3390/rs13132577.

[18] M. Alqahtani, A. Gumaei, H. Mathkour, and M. M. Ben Ismail, “A genetic-based extreme gradient boosting model for detecting intrusions in wireless sensor networks,” *Sensors (Switzerland)*, vol. 19, no. 20, 2019, doi: 10.3390/s19204383.

[19] S. B. Yang and T. L. Chen, “Uncertain decision tree for bank marketing classification,” *J. Comput. Appl. Math.*, vol. 371, p. 112710, 2020, doi: 10.1016/j.cam.2020.112710.

[20] H. H. Patel and P. Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[21] S. Rizkia, E. Budi Setiawan, and D. Puspandari, “Analisis Sentimen Kepuasan Pelanggan Terhadap Internet Provider Indihome di Twitter Menggunakan Metode Decision Tree dan Pembobotan TF-IDF,” *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9683–9693, 2019.

[22] I. A. Nikmatun and I. Waspada, “Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor,” *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.

[23] W. Hou, D. Li, C. Xu, H. Zhang, and T. Li, “An Advanced k Nearest Neighbor Classification Algorithm Based on KD-tree,” *Proc. 2018 IEEE Int. Conf. Saf. Prod. Informatiz. IICSPI 2018*, pp. 902–905, 2019, doi: 10.1109/IICSPI.2018.8690508.

[24] D. A. Adeniyi, Z. Wei, and Y. Yongquan, “Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method,” *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001.

[25] I. Aghaei and A. Sokhanvar, “Factors influencing SME owners’ continuance intention in Bangladesh: a logistic regression model,” *Eurasian Bus. Rev.*, vol. 10, no. 3, pp. 391–415, 2020, doi: 10.1007/s40821-019-00137-6.

[26] S. Solutions, “What is Logistic Regression?,” *Statistics Solutions*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/ (accessed Jul. 02, 2021).

[27] S. S. Nasim, C. M. Mizan, T. Chakroborty, S. Ghosh, and S. Karmakar, *A Survey of Load Balanced Job Scheduling Schemes in Cloud Computing*, vol. 1187. 2021.

[28] S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” *Appl. Soft Comput. J.*, vol. 76, pp. 380–389, 2019, doi: 10.1016/j.asoc.2018.12.024.

[29] J. He, Y. Hao, and X. Wang, “An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGboost,” *J. Mar.*

Sci. Eng., vol. 9, no. 2, pp. 1–19, 2021, doi: 10.3390/jmse9020156.

[30] B. Raharjo, *Mudah Belajar Python Untuk Aplikasi Dekstop Dan Web Edisi Revisi*. Bandung, 2019.

[31] J. Enterprise, *Python untuk Programmer Pemula*. Jakarta: PT Elex Media Komputindo, 2019.

[32] J. Hao and T. K. Ho, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language," *J. Educ. Behav. Stat.*, vol. 44, no. 3, pp. 348–361, 2019, doi: 10.3102/1076998619832248.

[33] K. Gede and R. Aditya, "Program Menghitung Banyak Bata pada Ruangan Menggunakan Bahasa Python," vol. 2, no. 1, 2021.

[34] Jubilee Enterprise, *Python untuk Programmer Pemula*. PT Elex Media Kamputindo, 2019.

[35] C. Diana and M. S. Tanjung, "SISTEM TRANSAKSI DAN PERHITUNGAN BUNGA DEPOSITO PADA BANK NAGARI CABANG PARIAMAN," *CC-By Attrib. 4.0 Int.*, pp. 1–8, 2019, doi: 10.31219/osf.io/exct7.

[36] S. Jaggia, A. Kelly, K. Lertwachara, and L. Chen, "Applying the CRISP-DM Framework for Teaching Business Analytics," *Decis. Sci. J. Innov. Educ.*, vol. 18, no. 4, pp. 612–634, 2020, doi: 10.1111/dsji.12222.

[37] T. Darmawan, A. S. Birawa, E. Eryanto, and T. Mauritsius, "Credit classification using crisp-dm method on bank abc customers," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2375–2380, 2020, doi: 10.30534/ijeter/2020/28862020.

[38] M. Syukron, R. Santoso, and T. Widiharih, "PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IMBALANCE CLASS DATA," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, doi: 10.14710/j.gauss.v9i3.28915.

[39] S. Wang *et al.*, "A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning," *Fuel*, vol. 282, no. July, p. 118848, 2020, doi: 10.1016/j.fuel.2020.118848.

[40] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan, and S. Ahmed,

"IDTi-CSsmoteB: Identification of Drug-Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost with Over-Sampling Technique SMOTE," *IEEE Access*, vol. 7, pp. 48699–48714, 2019, doi: 10.1109/ACCESS.2019.2910277.

[41] S. R. Mousa, P. R. Bakhit, and S. Ishak, "An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: A naturalistic driving study," *Can. J. Civ. Eng.*, vol. 46, no. 8, pp. 712–721, 2019, doi: 10.1139/cjce-2018-0117.

[42] A. Abu-Srhan, S. Al zghoul, B. Alhammad, and R. Al-Sayyed, "Visualization and analysis in bank direct marketing prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 651–657, 2019, doi: 10.14569/ijacsa.2019.0100785.

[43] M. S. Basarslan and I. D. Argun, "Classification of a bank data set on various data mining platforms | Bir Banka Müşteri Verilerinin Farkli Veri Madenciliği Platformlarinda Siniflandirilmasi," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018*, 2018, pp. 1–4.

[44] L. Geetha, K. Anusha, D. Ganeshgoud, S. M. Gouse, and V. N. Mandhala, "Predicting the success of bank marketing system using classification techniques," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 2, pp. 1961–1968, 2020, doi: 10.5373/JARDCS/V12I2/S20201241.

[45] V. U. Pugliese, C. M. Hirata, and R. D. Costa, "Comparing supervised classification methods for financial domain problems," *ICEIS 2020 - Proc. 22nd Int. Conf. Enterp. Inf. Syst.*, vol. 1, no. Iceis, pp. 440–451, 2020, doi: 10.5220/0009426204400451.

[46] Y. Chen, "Analysis of applying Genetic Algorithm to Simple Neural Network Based on Bank Direct Marketing Dataset," pp. 1–6, 2018.

[47] K. Nizam, A. Halim, A. Syukor, M. Jaya, A. Firdaus, and A. Fadzil, "Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 3, pp. 272–277, 2020, doi: 10.35940/ijitee.c8472.019320.

[48] S. Mishra, "Handling Imbalanced Data : SMOTE vs . Random

Undersampling," *Int. Res. J. Eng. Technol.*, vol. 4, no. 8, pp. 317–320, 2017, [Online]. Available: https://irjet.net/archives/V4/i8/IRJET-V4I857.pdf.

[49] G. Marinakos and S. Daskalaki, "Imbalanced customer classification for bank direct marketing," *J. Mark. Anal.*, vol. 5, no. 1, pp. 14–30, 2017, doi: 10.1057/s41270-017-0013-7.

[50] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, p. 1502, 2016, doi: 10.12928/telkomnika.v14i4.3956.

[51] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017, doi: 10.33096/ilkom.v9i2.139.188-195.

[52] Y. Meng, N. Yang, Z. Qian, and G. Zhang, "What makes an online review more helpful: An interpretation framework using xgboost and shap values," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 3, pp. 466–490, 2021, doi: 10.3390/jtaer16030029.

[53] J. Muschelli, "ROC and AUC with a Binary Predictor: a Potentially Misleading Metric," *J. Classif.*, vol. 37, no. 3, pp. 696–708, 2020, doi: 10.1007/s00357-019-09345-1.

DAFTAR RIWAYAT HIDUP

A. Biodata Mahasiswa

NIM	: 14002371
Nama	: Riki Supriyadi
Tempat, Tanggal Lahir	: Tegal, 26 Juli 1996
Jenis Kelamin	: Laki-laki
Alamat	: Ds. Karangjati RT 02/01 Kecamatan Tarub Kabupaten Tegal

B. Riwayat Pendidikan Formal dan Non Formal

1. SDN Karangjati 01 — Lulus Tahun 2009
2. SMP N 1 Tarub — Lulus Tahun 2012
3. SMA N 1 Kramat — Lulus Tahun 2015
4. UBSI Kota Tegal (D3) — Lulus Tahun 2018
5. Universitas BSI Bandung (S1) — Lulus Tahun 2019

C. Riwayat Pengalaman Organisasi / Pekerjaan

1. Senat Mahasiswa UBSI Tegal (2016-2017)
2. Komite Mahasiswa Anti Narkotika dan Obat Terlarang (2016-2018)
3. Paguyuban Duta Wisata Kabupaten Tegal (2017-Sekarang)
4. UBSI (2019-Sekarang)

Jakarta, 26 Juli 2021

(Riki Supriyadi)

LAMPIRAN

Lampiran 1. Sampel Dataset

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
5	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
6	59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
7	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
8	24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
9	25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
10	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
11	25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
12	29	blue-collar	single	high.school	no	no	yes	telephone	may	mon	137	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
13	57	housemaid	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
14	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
15	54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
16	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
17	46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
18	50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
19	39	management	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
20	30	unemployed	married	high.school	no	no	no	telephone	may	mon	38	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
21	55	blue-collar	married	basic.4y	unknown	yes	no	telephone	may	mon	262	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
22	55	retired	single	high.school	no	yes	no	telephone	may	mon	342	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
23	41	technician	single	high.school	no	yes	no	telephone	may	mon	181	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
24	37	admin.	married	high.school	no	yes	no	telephone	may	mon	172	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
25	35	technician	married	university.degree	no	no	yes	telephone	may	mon	99	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
26	59	technician	married	unknown	no	yes	no	telephone	may	mon	93	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
27	39	self-employed	married	basic.9y	unknown	no	no	telephone	may	mon	233	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
28	54	technician	single	university.degree	unknown	no	no	telephone	may	mon	255	2	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
29	55	unknown	married	university.degree	unknown	unknown	unknown	telephone	may	mon	362	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
30	46	admin.	married	unknown	no	no	no	telephone	may	mon	348	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
31	59	technician	married	unknown	no	yes	no	telephone	may	mon	386	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
32	49	blue-collar	married	unknown	no	no	no	telephone	may	mon	73	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
33	54	management	married	basic.4y	unknown	yes	no	telephone	may	mon	230	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
34	54	blue-collar	divorced	basic.4y	no	no	no	telephone	may	mon	208	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
35	55	unknown	married	basic.4y	unknown	yes	no	telephone	may	mon	336	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
36	34	services	married	high.school	no	no	no	telephone	may	mon	365	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
37	52	technician	married	basic.9y	no	yes	no	telephone	may	mon	1666	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
38	41	admin.	married	university.degree	no	yes	no	telephone	may	mon	577	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
39	56	technician	married	basic.4y	no	yes	no	telephone	may	mon	137	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
40	58	management	unknown	university.degree	no	yes	no	telephone	may	mon	366	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
41	32	entrepreneur	married	high.school	no	yes	no	telephone	may	mon	314	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
42	38	admin.	single	professional.course	no	no	no	telephone	may	mon	160	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
43	57	admin.	married	university.degree	no	no	yes	telephone	may	mon	212	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
44	44	admin.	married	university.degree	unknown	yes	no	telephone	may	mon	188	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
45	42	technician	single	professional.course	unknown	no	no	telephone	may	mon	22	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
46	57	admin.	married	university.degree	no	yes	yes	telephone	may	mon	616	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
47	40	blue-collar	married	basic.9y	no	no	yes	telephone	may	mon	178	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
48	35	admin.	married	university.degree	no	yes	no	telephone	may	mon	355	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
49	45	blue-collar	married	basic.9y	no	yes	no	telephone	may	mon	225	2	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
50	54	admin.	married	high.school	no	no	no	telephone	may	mon	160	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
51	39	housemaid	married	basic.4y	no	no	yes	telephone	may	mon	266	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
52	60	admin.	married	high.school	no	no	no	telephone	may	mon	253	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
53	53	admin.	single	professional.course	no	no	no	telephone	may	mon	179	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
54	55	blue-collar	married	basic.4y	unknown	no	no	telephone	may	mon	269	2	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
55	55	technician	married	professional.course	unknown	yes	no	telephone	may	mon	135	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
56	50	management	married	university.degree	unknown	no	yes	telephone	may	mon	161	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
57	45	services	married	high.school	unknown	yes	no	telephone	may	mon	787	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
58	55	unemployed	married	professional.course	unknown	yes	yes	telephone	may	mon	145	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
59	25	technician	single	university.degree	no	yes	no	telephone	may	mon	174	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no

Lampiran 2. Hasil penelitian

Data *Imbalance*:

1. ***XGBoost***

	precision	recall	f1-score	support
0	0.95	0.96	0.96	7319
1	0.66	0.57	0.61	919
accuracy			0.92	8238
macro avg	0.81	0.77	0.78	8238
weighted avg	0.92	0.92	0.92	8238

2. ***Decision Tree***

	precision	recall	f1-score	support
0	0.94	0.96	0.95	7319
1	0.59	0.50	0.54	919
accuracy			0.91	8238
macro avg	0.76	0.73	0.75	8238
weighted avg	0.90	0.91	0.90	8238

3. ***KNeighbors***

	precision	recall	f1-score	support
0	0.93	0.97	0.95	7319
1	0.62	0.38	0.47	919
accuracy			0.90	8238
macro avg	0.77	0.68	0.71	8238
weighted avg	0.89	0.90	0.89	8238

4. ***Logistic Regression***

	precision	recall	f1-score	support
0	0.91	0.99	0.94	7319
1	0.62	0.19	0.29	919
accuracy			0.90	8238
macro avg	0.76	0.59	0.62	8238
weighted avg	0.87	0.90	0.87	8238

Lanjutan (Hasil penelitian)

Data *Balance - SMOTE*:

1. ***XGBoost-SMOTE***

	precision	recall	f1-score	support
0	0.98	0.99	0.98	7319
1	0.90	0.81	0.85	919
accuracy			0.97	8238
macro avg	0.94	0.90	0.92	8238
weighted avg	0.97	0.97	0.97	8238

2. ***Decision Tree-SMOTE***

	precision	recall	f1-score	support
0	0.94	0.96	0.95	7319
1	0.66	0.55	0.60	919
accuracy			0.92	8238
macro avg	0.80	0.76	0.78	8238
weighted avg	0.91	0.92	0.91	8238

3. ***KNeighbors-SMOTE***

	precision	recall	f1-score	support
0	0.94	0.94	0.94	7319
1	0.52	0.51	0.51	919
accuracy			0.89	8238
macro avg	0.73	0.72	0.73	8238
weighted avg	0.89	0.89	0.89	8238

4. ***Logistic Regression-SMOTE***

	precision	recall	f1-score	support
0	0.93	0.98	0.95	7319
1	0.69	0.44	0.54	919
accuracy			0.92	8238
macro avg	0.81	0.71	0.75	8238
weighted avg	0.91	0.92	0.91	8238