Komparasi Algoritma Klasifikasi untuk Prediksi *Leads* yang akan Menjadi Customer

Abednego Steven Sihite [1], Hendra Bunyamin [1,*]

[1] Program Studi S1 Teknik Informatika; Universitas Kristen Maranatha; Jln. Prof. Drg. Surya Sumantri No. 65, Sukawarna, Bandung, Indonesia, 62222015154; e-mail: 1972009@maranatha.ac.id, hendra.bunyamin@it.maranatha.edu

* Korespondensi: e-mail: hendra.bunyamin@it.maranatha.edu

Diterima: 24 Oktober 2023 ; Review: 21 November 2023; Disetujui: 11 Desember 2023

Cara sitasi: Sihite AS, Bunyamin H. 2023. Komparasi Algoritma Klasifikasi untuk Prediksi *Leads* yang akan Menjadi Customer. Bina Insani ICT Journal. Vol 10(2): 188 - 199.

Abstrak: *Leads* adalah sekelompok orang yang memiliki potensi untuk menjadi pelanggan dari sebuah produk dan jasa dari sebuah bisnis. Oleh karena itu, mengetahui segmentasi *leads* dan memprediksi *leads* yang akan menjadi pelanggan merupakan hal yang sangat penting bagi sebuah perusahaan. Beruntungnya di era digital ini, teknologi dapat digunakan untuk membantu pekerjaan manusia seperti memprediksi *leads* yang akan menjadi pelanggan. Salah satu teknologi yang dapat digunakan adalah penerapan machine learning. Proses penerapan machine learning dalam memprediksi *leads* yang akan menjadi pelanggan meliputi preprocessing data, analisis data eksploratif, serta pembuatan dan interpretasi model machine learning itu sendiri. Algoritma klasifikasi yang diujicobakan pada penelitian ini adalah Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, Naive Bayes, dan K-Nearest Neighbors. Dataset yang digunakan dalam penelitian ini adalah data sekunder yang diambil dari platform Kaggle dengan total 9.240 baris dan 37 kolom. Preprocessing data dapat membersihkan kualitas dari dataset sehingga dataset tersebut dapat digunakan sebagai data training dan data testing dengan kualitas yang baik. Analisis Data Eksplorasi dapat menghasilkan informasi penting dari sebuah dataset. Terakhir, pembangunan dan interpretasi model dapat menjelaskan bagaimana sebuah fitur dalam dataset dapat mempengaruhi prediksi sebuah model. Pada bagian algoritma, berdasarkan hasil uji coba didapatkan bahwa algoritma Random Forest merupakan algoritma dengan nilai terbaik dalam memprediksi prospek yang akan menjadi pelanggan. Terakhir, pada tahap interpretasi model dengan menggunakan metode partial dependence plot menghasilkan kesimpulan bahwa semakin lama calon pelanggan menghabiskan waktu di website perusahaan, maka semakin besar kemungkinan calon pelanggan tersebut menjadi pelanggan, kemudian *leads* yang teridentifikasi melalui Lead Add Form memiliki peluang lebih besar untuk menjadi pelanggan dan yang terakhir calon pelanggan atau lead yang bekerja secara profesional memiliki peluang lebih besar untuk menjadi pelanggan

Kata kunci: algoritma klasifikasi, analisis data eksploratori, interpretasi model, *partial dependence plot*

Abstract: *Leads are a group of people who have the potential to become customers of a product and service from a business. Therefore, knowing the segmentation of leads and predicting leads that will become customers is very important for a company. Luckily in this digital age, technology can be used to help human work such as predicting leads that will become customers. One of the technologies that can be used is the application of machine learning. The process of applying machine learning in predicting leads that will become customers includes data preprocessing, exploratory data analysis, and the creation and interpretation of the machine learning model itself. The classification algorithms tested in this research are Logistic Regression, Support Vector Machine, Decision Trees, Random Forest,*

Naive Bayes, and K-Nearest Neighbors. The dataset used in this research is secondary data taken from the Kaggle platform with a total of 9,240 rows and 37 columns. Data preprocessing can clean the quality of the dataset so that the dataset can be used as training data and test data with good quality. Exploratory Data Analysis can generate important information from a dataset. Finally, model building and interpretation can explain how a feature in the dataset can affect the prediction of a model. In the algorithm section, based on the test results, it is found that the Random Forest algorithm is the algorithm with the best score in predicting leads that will become customers. Finally, at the model interpretation stage using the partial dependence plot method results in the conclusion that the longer a prospective customer spends on the company website, the more likely it is that the prospective customer will become a customer, then leads identified through the Lead Add Form have a greater chance of becoming customers and finally prospective customers or leads who work professionally have a greater chance of becoming customers.

***Keywords**: classification algorithm, exploratory data analysis, model interpretation, partial dependence plot*

1. Pendahuluan

Leads adalah sekumpulan orang yang berpotensi untuk menjadi pelanggan atau customer suatu produk dan jasa dari sebuah usaha. Oleh karena itu, mengetahui segmentasi leads dan memprediksi leads yang akan menjadi customer sangatlah penting bagi sebuah perusahaan. Penelitian ini sendiri mengambil sebuah kasus pada perusahaan edukasi kursus online X, dimana perusahaan X ingin lebih berfokus melakukan komunikasi untuk menjual produknya kepada leads yang benar-benar berpotensi menjadi customer. Hal tersebut dilakukan agar perusahaan X tidak perlu menghubungi setiap leads yang ada, dan usaha persuasif yang dilakukan perusahaan X untuk menarik customer dapat menjadi lebih efisien dan efektif. Perusahaan X sendiri harus melakukan prediksi dan klasifikasi terhadap sekumpulan leads untuk mengetahui leads yang benar-benar akan menjadi customer. Hal tersebut menjadi sebuah permasalahan karena cukup sulit bagi manusia untuk melakukan prediksi satu persatu terhadap sekumpulan data leads yang banyak jumlahnya.

Untungnya pada zaman digital ini, teknologi dapat digunakan untuk membantu pekerjaan manusia, salah satunya dalam hal memprediksi seperti memprediksi leads yang akan menjadi customer. Machine Learning sendiri adalah salah satu teknologi yang dapat digunakan untuk membantu pekerjaan dalam memprediksi sebuah kejadian atau output [1]. Proses penerapan machine learning terdiri dari beberapa proses yang mempunyai fungsinya masing-masing, mulai dari preprocessing data, exploratory data analysis, sampai dengan pembuatan dan interpretasi model machine learning itu sendiri. Model machine learning pada penerapannya juga mempunyai banyak jenis algoritma. Algoritma sendiri adalah salah satu dari sekian banyak unsur penting yang akan mempengaruhi baik atau buruknya kinerja model machine learning dalam melakukan prediksi pada suatu kasus tertentu. Selain algoritma, factor penentu lainnya adalah hyperparameters pada model, kualitas data, dan lain-lain.

Oleh karena itu, dalam penelitian ini beberapa algoritma klasifikasi model pembelajaran mesin akan diujicobakan untuk diukur performanya dalam menyelesaikan permasalahan memprediksi leads yang akan menjadi customer. Hal tersebut dilakukan untuk mengetahui algoritma terbaik dalam menyelesaikan permasalahan prediksi leads yang akan menjadi customer. Seluruh proses penerapan machine learning mulai dari preprocessing data, exploratory data analysis, dan sampai dengan tahapan pembuatan serta interpretasi model machine learning juga akan dilakukan untuk mengetahui fungsi dan nilai manfaat dari setiap proses tersebut.

.

2. Metode Penelitian

2.1 Data

Data yang digunakan pada penelitian ini adalah data leads perusahaan X. Perusahaan X bergerak di bidang edukasi dan menjual kursus-kursus online berstandar industri profesional. Perusahaan ini memasarkan kursusnya di beberapa situs web dan mesin pencari seperti Google. Setelah orang atau pengunjung mengakses situs web, mereka menelusuri kursus atau mengisi formulir untuk kursus atau menonton beberapa video. Ketika orang-orang ini mengisi formulir dengan memberikan alamat email atau nomor telepon mereka, mereka diklasifikasikan

sebagai leads. Data sekunder yang digunakan pada penelitian ini sendiri terdiri dari 9240 baris dan 37 kolom. Data didapatkan dari platform penyedia dataset yaitu Kaggle dengan sumber sebagai berikut (https://www.kaggle.com/datasets/ashydv/leads-dataset). Berikut adalah contoh gambaran dataset dengan total 37 atribut.

	0	1	2	3
Prospect ID	7927b2df-8bba-4d29-b9a2-b6e0beafe620	2a272436-5132-4136-86fa-dcc88c88f482	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	0cc2df48-7cf4-4e39-9de9-19797f9b38cc
Lead Number	660737	660728	660727	660719
Lead Origin	API	API	Landing Page Submission	Landing Page Submission
Lead Source	Olark Chat	Organic Search	Direct Traffic	Direct Traffic
Do Not Email	No	No	No	No
Do Not Call	No	No	No	No
Converted	0	0	1	0
TotalVisits	0.0	5.0	2.0	1.0
Total Time Spent on Website	0	674	1532	305
Page Views Per Visit	0.0	2.5	2.0	1.0
Last Activity	Page Visited on Website	Email Opened	Email Opened	Unreachable
Country	NaN	India	India	India
Specialization	Select	Select	Business Administration	Media and Advertising
How did you hear about X Education	Select	Select	Select	Word Of Mouth
What is your current	Unemployed	Unemployed	Student	Unemployed

Sumber: Hasil Penelitian (2023)

Gambar 1. Dataset bagian pertama

What is your current occupation	Unemployed	Unemployed	Student	Unemployed
What matters most to you in choosing a course	Better Career Prospects	Better Career Prospects	Better Career Prospects	Better Career Prospects
Search	No	No	No	No
Magazine	No	No	No	No
Newspaper Article	No	No	No	No
X Education Forums	No	No	No	No
Newspaper	No	No	No	No
Digital Advertisement	No	No	No	No
Through Recommendations	No	No	No	No
Receive More Updates About Our Courses	No	No	No	No
Tags	Interested in other courses	Ringing	Will revert after reading the email	Ringing
Lead Quality	Low in Relevance	NaN	Might be	Not Sure
Update me on Supply Chain Content	No	No	No	No
Get updates on DM Content	No	No	No	No
Lead Profile	Select	Select	Potential Lead	Select
City	Select	Select	Mumbai	Mumbai
Asymmetrique Activity Index	02.Medium	02.Medium	02.Medium	02.Medium

Sumber: Hasil Penelitian (2023)

Gambar 2. Dataset bagian kedua

2.2 Alir Penelitian

1. *Preprocessing Data*: Proses preprocessing data adalah proses pembersihan data dan pengecekan kualitas data. Proses ini penting untuk dilakukan agar kualitas data yang akan digunakan terjaga dan baik. Contoh preprocessing data dalam penelitian ini adalah mengubah

beberapa nilai NaN pada beberapa kolom di dataset menjadi sebuah nilai atau fitur asli yang dapat digunakan. Selain itu, proses merubah data kategorik menjadi sebuah matriks (one hot encoding) agar data tersebut dapat diterima sebagai masukan terhadap model juga dapat dilakukan di tahap preprocessing data.
2. *Exploratory Data Analysis*: Sebelum melakukan pelatihan model, akan lebih baik jika data dieksplor dan dianalisis terlebih dahulu. Proses exploratory data analysis dapat mengekstraksi informasi-informasi mengenai data yang belum diketahui sebelumnya. Semakin banyak informasi mengenai data yang didapatkan, maka proses pembuatan dan pelatihan model dapat menjadi lebih efektif serta efisien. Penentuan fitur utama pada data yang akan berpengaruh besar pada hasil prediksi juga dapat dilakukan dalam proses exploratory data analysis.
3. Pembuatan dan Pelatihan Model dengan Algoritma: Proses pembuatan dan pelatihan model akan dilakukan bila data sudah diproses sedemikian rupa dan siap untuk digunakan. Penentuan hyperparameter, fungsi aktivasi, dan komponen-komponen utama untuk model lainnya akan dilakukan pada tahapan ini. Terakhir, pelatihan model berdasarkan algoritma yang diteliti akan dilakukan dengan data dan hyperparameter yang sudah disiapkan serta ditentukan sebelumnya.
4. *Model Tuning*: *Model Tuning* diperlukan untuk mengecek apakah performa model masih bisa ditingkatkan lebih lagi. Tuning hyperparameters adalah salah satu cara yang dapat dilakukan untuk melakukan *tuning* terhadap model dan meningkatkan performa model. K-fold Cross Validation sendiri adalah metode yang dipakai untuk *model tuning* pada penelitian ini, sehingga Randomised Search Cross-Validation dapat digunakan untuk *hyperparameters tuning*.
5. Interpretasi Hasil dan Pembahasan: Tahapan selanjutnya pada penelitian ini setelah tahapan evaluasi model adalah tahap interpretasi hasil dan pembahasan. Proses interpretasi akan menggunakan metode model-agnostic. Interpretasi model agnostic mengacu pada metode atau teknik untuk menguraikan atau memahami bagaimana model pembelajaran mesin (machine learning) membuat keputusan atau prediksi tanpa bergantung pada detail internal dari model itu sendiri. Dalam konteks ini, "agnostic" berarti tidak terkait dengan jenis atau arsitektur model tertentu. Partial Dependence Plots (PDP) merupakan salah satu metode interpretasi model agnostic yang bertujuan untuk menunjukkan dampak relatif satu atau dua fitur pada prediksi model dengan mempertahankan nilai fitur lainnya tetap. Proses interpretasi model juga dapat menjelaskan fitur apa saja yang mempengaruhi prediksi model. Lalu untuk proses perbandingan setiap model itu sendiri akan dilakukan dengan cara melihat performa yang dihasilkan setiap model, seperti f1-score dan akurasi.
6. Kesimpulan: Tahapan ini akan meringkas hasil pembahasan sebelumnya untuk mendapatkan kesimpulan dari penelitian yang telah dilakukan. Tahapan ini juga akan menjelaskan hal-hal apa saja yang dapat ditingkatkan untuk penelitian ke depannya.

3. Hasil dan Pembahasan

3.1 Preprocessing Data

Tahap awal adalah memuat dan memahami data. Memuat dan memahami data merupakan tahapan yang tidak boleh dilewatkan agar data dapat dipahami dengan baik. Hal tersebut meliputi memahami format data, struktur data, dan jenis data. Pemahaman yang baik akan data dapat memudahkan pekerjaan dalam menentukan teknik analisis yang tepat dan persiapan data yang perlu dilakukan untuk pelatihan model.

```
# load data
df_leads = pd.read_csv('/kaggle/input/leadscsv/Leads.csv')
```

Sumber: Hasil Penelitian (2023)

Gambar 3. Memuat dataset menjadi dataframe

```
df_leads.head()
```

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0

Sumber: Hasil Penelitian (2023)

Gambar 4. Cek 5 data pertama menggunakan function head

```
# checking data data types
df_leads.dtypes
```

```
Prospect ID                                    object
Lead Number                                     int64
Lead Origin                                    object
Lead Source                                    object
Do Not Email                                   object
Do Not Call                                    object
Converted                                       int64
TotalVisits                                   float64
Total Time Spent on Website                     int64
Page Views Per Visit                          float64
Last Activity                                  object
Country                                        object
Specialization                                 object
How did you hear about X Education             object
```

Sumber: Hasil Penelitian (2023)

Gambar 5. Cek Tipe Data

Data Cleansing merupakan proses cleaning data mentah yang sudah dimuat sebelumnya agar kualitas data latih dapat menjadi lebih baik kualitasnya. Hal pertama yang dapat dilakukan dalam tahapan data cleaning adalah melakukan pengecekan apakah ada duplikasi pada data atau tidak. Proses pembuangan dan penanganan missing values juga akan dilakukan di tahapan data cleansing.

```
# delete columns that have more than 3000 missing values
for i in df_leads.columns:
    if df_leads[i].isna().sum()>3000:
        df_leads.drop(i, axis=1, inplace=True)
```

```
df_leads.isnull().sum()
```

Sumber: Hasil Penelitian (2023)

Gambar 6. Menghapus kolom/atribut dengan jumlah missing values di atas 3000

```
df_leads['Lead Source'].fillna('Google', inplace=True)
```

Sumber: Hasil Penelitian (2023)

Gambar 7. Contoh Mengisi Data Kosong dengan Data Pengganti

```
df_leads['City'].fillna('Select', inplace=True)
```

Sumber: Hasil Penelitian (2023)

Gambar 8. Contoh Kedua Mengisi Data Kosong dengan Data Pengganti

Pemisahan Dataset merupakan pemisahan data yang sudah melalui proses pembersihan menjadi 2 subset bagian data, yaitu data latih dan data uji. Pemisahan data akan dilakukan dengan menggunakan fungsi train_test_split dari pustaka sklearn. Keseluruhan data akan dipisah menjadi 80 persen data latih dan 20 persen data uji. Label y untuk data latih dan data uji juga akan dibuat di tahap ini. Pembuatan label y dilakukan dengan cara menerapkan fungsi values.reshape pada kolom converted pada data latih dan uji yang sudah dipisahkan sebelumnya, dan hasil keluaran fungsi tersebut dimasukkan kedalam variable.

```
train_full, test_X = train_test_split(df_leads, test_size=0.2, random_state=1)
```

```
train_y = train_full['converted'].values.reshape(-1,1)
test_y = test_X['converted'].values.reshape(-1,1)
```

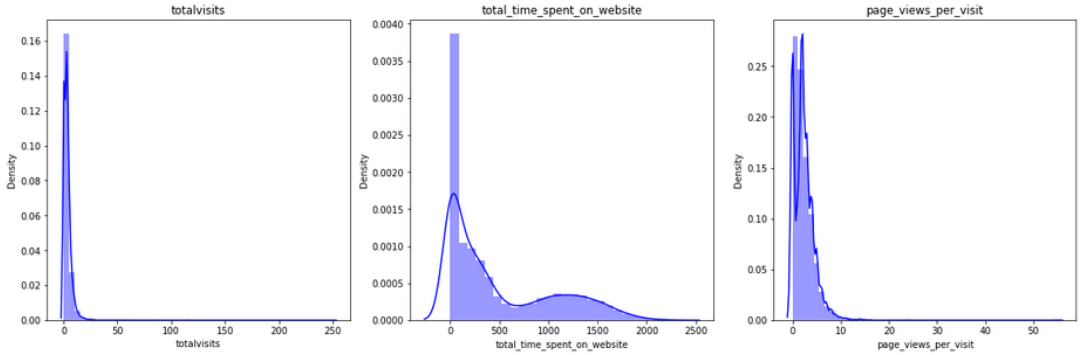
Sumber: Hasil Penelitian (2023)

Gambar 9. Membagi Dataset menjadi Data Latih dan Data Uji

3.2 Exploratory Data Analysis (EDA)

Tahap awal dari EDA adalah menentukan metrics untuk model. Pada dataset sendiri ada kolom converted yang menunjukkan bahwa calon customer sudah menjadi customer atau tidak. Distribusi nilai converted ini juga penting untuk dicek apakah dataset balanced atau tidak. Hal tersebut perlu dilakukan karena dataset balanced atau tidak dapat menentukan metrics pengukur performa model mana yang harus digunakan (*accuracy*, f1-score, dan lain-lain). Pengecekan dapat dilakukan dengan melakukan visualisasi data menggunakan fungsi countplot dari pustaka seaborn. Berdasarkan visualisasi data yang dilakukan, dapat diambil kesimpulan bahwa dataset tidak seimbang. Oleh karena itu, accuracy tidak dapat dijadikan sebagai metrics utama pengukuran performa model [support-sas]. Metrics f1-score dapat digunakan sebagai pengukur performa model selain metrics accuracy.

Tahap berikutnya adalah analisis dan visualisasi data. Analisis yang dilakukan dari visualisasi keseluruhan data dapat memberikan informasi dan insight berguna untuk digunakan ke depannya dalam proses pelatihan serta pembuatan model. Pustaka seaborn dan matplotlib akan digunakan untuk membantu proses visualisasi data. Berdasarkan visualisasi data yang telah dilakukan ada beberapa informasi penting yang didapatkan, sebagai berikut; kebanyakan calon customer tidaklah sedang mempunyai pekerjaan, calon customer yang sudah bekerja di industri profesional lebih banyak berakhir menjadi customer dibandingkan tidak menjadi customer, calon customer yang terakhir kali dihubungi menggunakan SMS lebih banyak berakhir menjadi customer dibandingkan dengan metode komunikasi lainnya pada kolom Last Activity dan Last Notable Activity, calon customer yang menghabiskan waktu lebih banyak di website lebih banyak berakhir menjadi customer



Sumber: Hasil Penelitian (2023)

Gambar 10. Contoh Visualisasi Data terhadap Atribut Numerik

Berikutnya adalah *feature importance*. Feature importance adalah metode untuk menentukan fitur yang paling penting dalam memprediksi dengan menggunakan model machine learning. Dalam feature importance, setiap variabel input diberi skor berdasarkan kontribusinya terhadap prediksi akhir. Ada berbagai cara untuk menghitung feature importance, salah satunya adalah dengan menggunakan mutual information. Dalam konteks machine learning, mutual information dapat digunakan untuk mengukur relevansi variabel- variabel input terhadap variabel target, sehingga variabel-variabel input yang tidak relevan dapat dihilangkan dari model. Pustaka sklearn sudah menyediakan tools untuk menghitung mutual information. Sembilan kolom variabel kategori dengan mutual information terbesar akan diambil untuk menjadi fitur pada pelatihan model, dan sisanya akan dibuang. Sedangkan untuk fitur numerik tidak akan ada yang dibuang, karena jumlahnya terhitung sedikit.

```
def calculate_mi(series):
    return mutual_info_score(series, df_leads['converted'])

data_mi = df_leads[categorical].apply(calculate_mi)
data_mi = data_mi.sort_values(ascending=False).to_frame(name='Mi')
data_mi
```

	Mi
what_is_your_current_occupation	0.095197
last_activity	0.088512
last_notable_activity	0.073860
lead_source	0.061560
what_matters_most_to_you_in_choosing_a_course	0.060058
lead_origin	0.056251
specialization	0.014624
do_not_email	0.010327
a_free_copy_of_mastering_the_interview	0.000797

Sumber: Hasil Penelitian (2023)

Gambar 11. Feature Importance berdasarkan Mutual Information

Tahap selanjutnya adalah one-hot encoding. Model pembelajaran mesin hanya bisa menerima input berupa angka dalam bentuk matriks, sedangkan data yang ada belum semuanya berbentuk angka. Oleh karena itu, diperlukan metode one hot encoding untuk merubah data kategori menjadi sebuah matriks yang dapat diterima oleh model pembelajaran mesin. Dalam one hot encoding, setiap nilai kategori dari suatu variabel dikonversi menjadi vektor biner (0 atau 1), dengan panjang vektor yang sama dengan jumlah kategori unik pada variabel tersebut. Proses one hot encoding sendiri dapat dilakukan dengan bantuan

DictVectorizer dari pustaka sklearn. DictVectorizer bekerja dengan mengambil sebuah dictionary menjadi sebuah input dan melakukan vektorisasi terhadap dictionary tersebut. Hasil vektor tersebut akan ditaruh dan dibentuk menjadi sebuah matriks.

One Hot Encoding using DictVectorizer

```
dv = DictVectorizer(sparse=False)
dv.fit(train_dict)
dv.fit(test_dict)
```

Sumber: Hasil Penelitian (2023)

Gambar 12. One Hot Encoding menggunakan DictVectorizer (bagian 1)

```
X_train = dv.transform(train_dict)
X_test = dv.transform(test_dict)
```

Sumber: Hasil Penelitian (2023)

Gambar 13. One Hot Encoding menggunakan DictVectorizer (bagian 2)

3.3. Pembuatan dan Pelatihan Model

Seluruh algoritma yang dipakai untuk membuat model pembelajaran mesin sudah disediakan oleh pustaka sklearn. Tools untuk menghitung metrics pengukuran juga sudah disediakan oleh pustaka sklearn. Oleh karena itu, seluruh model dibangun menggunakan pustaka sklearn. Pada tahapan ini, setiap model dengan algoritma berbeda akan dibuat dan dilatih dengan data latih yang sudah disiapkan sebelumnya. Proses hyperparameters tuning juga dilakukan dalam tahapan ini untuk mencapai performa terbaik dari setiap model algoritma yang diujicobakan.

3.4. Pembahasan dan Interpretasi

Dikarenakan dataset yang tidak seimbang atau *imbalance* maka F1-Score dapat digunakan sebagai pengukur performa model selain dengan metrics accuracy. pemisahan dataset sendiri dibagi menjadi rasio 8:2. Berdasarkan pelatihan model dengan dataset yang ada dan beberapa algoritma berbeda yaitu Logistic Regression, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, dan LightGBM didapatkan hasil performa prediksi setiap model terhadap data uji dengan algoritma yang berbeda sebagai berikut:

Tabel 1. Skor performa algoritma

Metrics	Logistic Regression	Support Vector Machine	Naïve Bayes	Decision Trees	Random Forest	K-Nearest Neighbors	Light GBM
Accuracy	82%	72%	78%	80%	83%	77%	84%
F1	75%	60%	68%	73%	77%	69.7%	79%
Hyper Parameters Tuned Accuracy	83%	81%	78.5%	83%	84.4%	78%	84%
Hyper Parameters Tuned F1	77%	75%	70%	78%	79.4%	73%	79.2%

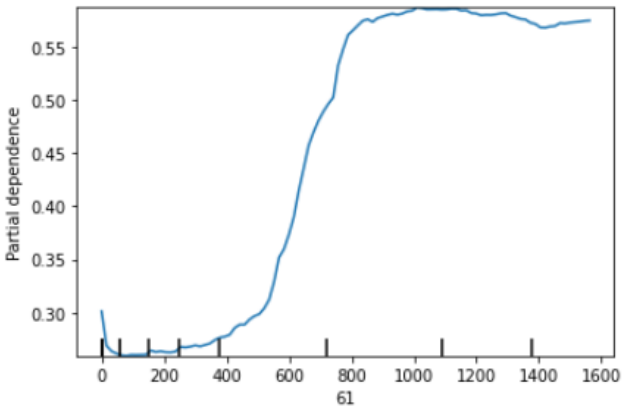
Metrics	Logistic Regression	Support Vector Machine	Naïve Bayes	Decision Trees	Random Forest	K-Nearest Neighbors	Light GBM
Accuracy	82%	72%	78%	80%	83%	77%	84%
F1	75%	60%	68%	73%	77%	69.7%	79%
Hyper Parameters Tuned Accuracy	83%	81%	78.5%	83%	84.4%	78%	84%
Hyper Parameters Tuned F1	77%	75%	70%	78%	79.4%	73%	79.2%

Sumber: Hasil Penelitian (2023)

Ada beberapa informasi yang dapat diambil dari hasil uji coba tersebut, yaitu:

a. Hyperparameters tuning terbukti dapat meningkatkan performa model dalam memprediksi bila dilihat dari akurasi dan f1-score yang dihasilkan.
b. Algoritma Support Vector Machine mengalami kenaikan skor paling signifikan setelah melalui proses hyperparameters tuning dibandingkan dengan algoritma lain yang diujicobakan. Performa model Support Vector Machine yang semakin baik setelah tuning dapat memberikan petunjuk bahwa data tidaklah terlalu independen.
c. Algoritma Naive Bayes menjadi algoritma dengan skor terburuk setelah proses hyperparameters tuning dibandingkan dengan algoritma lain yang diujicobakan.
d. Algoritma LightGBM menjadi algoritma dengan skor terbaik sebelum semua algoritma mengalami proses hyperparameters tuning.
e. Algoritma Random Forest menjadi algoritma dengan skor terbaik setelah semua algoritma mengalami proses hyperparameters tuning.

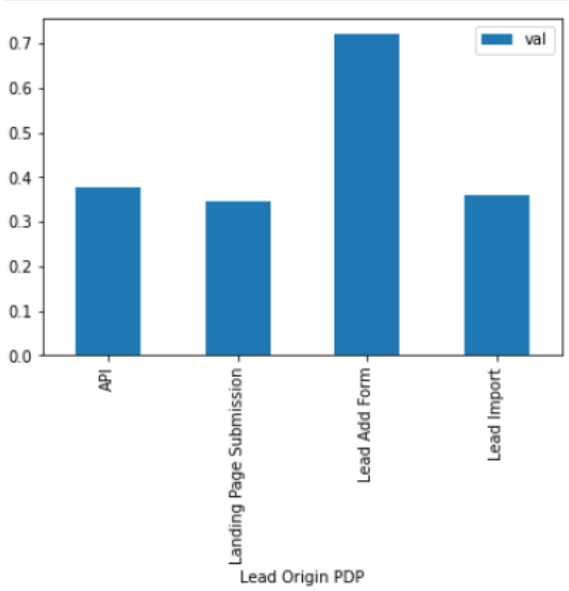
Random Forest menjadi algoritma pembangun model dengan skor terbaik, oleh karena itu interpretasi model akan dilakukan kepada model Random Forest. Interpretasi dilakukan dengan cara melihat hubungan suatu fitur dataset terhadap model. Model Random Forest sendiri mempunyai fungsi feature_importances_ untuk mengecek fitur terpenting yang mempengaruhi model itu sendiri. Dengan mengeksekusi fungsi feature_importances_ dan mengurutkan hasilnya berdasarkan nilai terkecil sampai terbesar, diperoleh hasil bahwa total_time_spent_on_website, last_notable_activity=SMS Sent, dan what_is_your_current_occupation=Working Professional adalah 3 fitur terpenting yang mempengaruhi model. Tiga fitur terpenting yang sudah didapatkan sebelumnya akan dicek kembali untuk memastikan apakah fitur-fitur tersebut memang benar mempunyai pengaruh paling besar terhadap model. Pengecekan selanjutnya dilakukan menggunakan partial dependence. Hasil visualisasi interpretasi model menggunakan partial dependence adalah sebagai berikut:



Sumber: Hasil Penelitian (2023)

Gambar 14. PDP Total Spent on Website

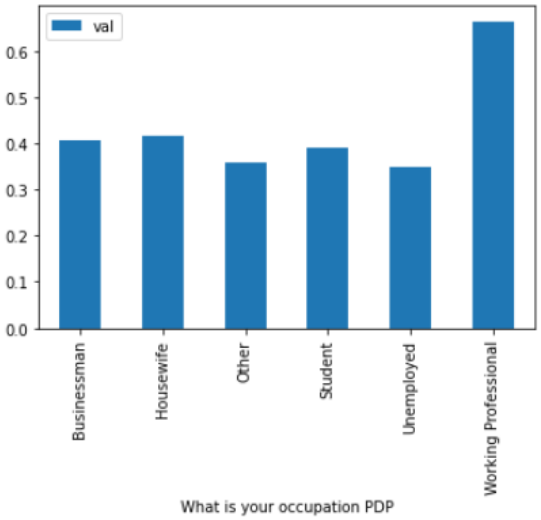
Berdasarkan Gambar 14, dapat diambil kesimpulan bahwa semakin lama seorang calon pelanggan menghabiskan waktu pada website perusahaan, maka semakin besar juga peluang calon pelanggan akan menjadi seorang pelanggan.



Sumber: Hasil Penelitian (2023)

Gambar 15. PDP Lead Origin

Berdasarkan gambar 15, dapat diambil kesimpulan bahwa data calon pelanggan terbanyak berasal dari Lead Add Form. Lalu data calon pelanggan kedua terbanyak berasal dari API. Terakhir, data calon pelanggan yang berasal dari Landing Page Submission dan Lead Import tidaklah jauh berbeda tetapi lebih sedikit persentasenya dibandingkan data yang berasal dari API dan Lead Add Form.

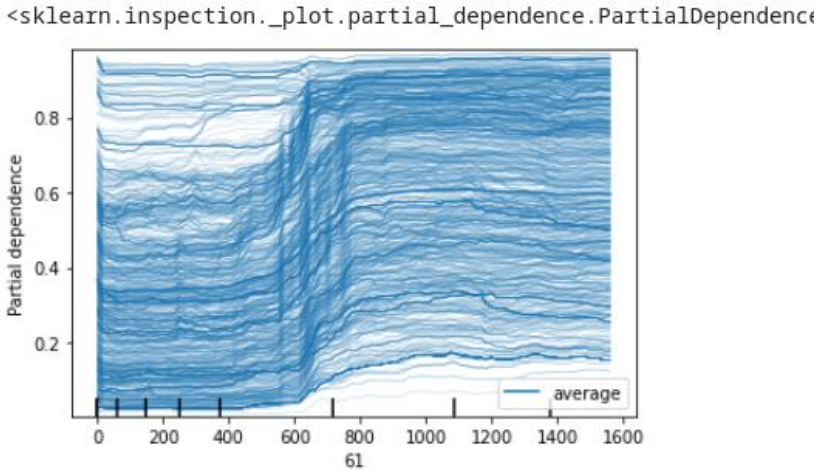


Sumber: Hasil Penelitian (2023)

Gambar 16. PDP What is your occupation

Berdasarkan gambar 16, dapat diambil kesimpulan bahwa sebagian besar calon pelanggan bekerja sebagai pekerja profesional. Menariknya, ibu rumah tangga adalah profesi calon pelanggan dengan jumlah kedua terbesar.

Selain menggunakan PDP, interpretasi juga akan dilakukan menggunakan metode ICE pada fitur “Total Time Spent on Website”. Hasil visualisasi interpretasi model menggunakan ICE adalah sebagai berikut:



Sumber: Hasil Penelitian (2023)

Gambar 17. ICE Total Time Spent on Website

Berdasarkan hasil visualisasi diatas, dapat diambil beberapa kesimpulan yaitu:

a. Fitur total_time_spent_on_website, last_notable_activity=SMS Sent, dan what_is_your_current_occupation=Working Professional memang mempunyai pengaruh yang cukup besar terhadap prediksi model.
b. Semakin lama calon customer menghabiskan waktu di dalam website perusahaan, semakin besar kemungkinan bahwa calon customer tersebut akan menjadi customer.
c. Lead yang diidentifikasi melalui Lead Add Form mempunyai kemungkinan lebih besar dalam menjadi customer.
d. Calon customer atau lead yang bekerja secara profesional mempunyai kemungkinan lebih besar dalam menjadi customer.
e. Terjadi kenaikan probabilitas sebuah lead akan menjadi customer bila calon customer tersebut menghabiskan waktunya pada website kurang lebih selama 700 menit.

4. Kesimpulan

4.1. Simpulan

Berdasarkan penelitian yang dilakukan, kesimpulan yang dapat diambil dapat dirumuskan menjadi 5 (lima) poin.

Pertama, proses preprocessing data dan exploratory data analysis sangat berpengaruh terhadap penerapan machine learning. Dengan melakukan proses preprocessing data, sebuah dataset yang sebelumnya masih berantakan, kotor, dan kurang baik untuk dijadikan data latih serta data uji dapat diproses sedemikian rupa menjadi dataset dengan kualitas baik serta siap untuk dijadikan data latih serta data uji model pembelajaran mesin. Proses exploratory data analysis juga mempunyai pengaruh besar dalam penerapan machine learning dalam penelitian ini. Berdasarkan proses exploratory data analysis pada penelitian ini, banyak informasi penting yang dapat diketahui dari sebuah dataset seperti proporsi keseimbangan data, perilaku calon customer atau leads, dan informasi penting lainnya.

Kedua, algoritma Random Forest menjadi algoritma dengan skor terbaik setelah proses hyperparameters tuning dibandingkan algoritma lainnya yang telah diujicoba dalam penelitian ini. Algoritma Support Vector Machine menjadi algoritma dengan skor terburuk sebelum proses hyperparameters tuning. Walaupun begitu algoritma Support Vector Machine mengalami kenaikan skor yang signifikan setelah proses hyperparameters tuning dan membuat algoritma Naive Bayes menjadi algoritma dengan skor terburuk setelah proses hyperparameters tuning.
Algoritma Random Forest (RF) dan LightGBM mendapatkan skor yang cukup baik membuktikan bahwa model ensembel mempunyai performa lebih baik dibandingkan model non-ensembel terhadap kasus dan dataset di penelitian ini. Walaupun begitu algoritma seperti Logistic Regression dan SVM mempunyai performa yang cukup baik walaupun kedua model algoritma tersebut bukanlah model ensemble. Hal tersebut dapat memberikan petunjuk bahwa dataset tidak terlalu independent.

Ketiga, proses interpretasi model menghasilkan temuan bahwa fitur total_time_spent_on_website, last_notable_activity=SMS Sent, dan what_is_your_current_occupation=Working Professional adalah fitur dengan pengaruh paling besar terhadap prediksi model. Proses interpretasi ini juga menghasilkan beberapa informasi penting yaitu semakin lama calon customer menghabiskan waktu di dalam website perusahaan, semakin besar kemungkinan bahwa calon customer tersebut akan menjadi customer, lalu leads yang diidentifikasi melalui Lead Add Form mempunyai kemungkinan lebih besar dalam menjadi customer dan terakhir calon customer atau lead yang bekerja secara profesional mempunyai kemungkinan lebih besar dalam menjadi customer.

Terakhir, algoritma RF sudah cukup baik performanya untuk menyelesaikan permasalahan dalam penelitian ini dengan dataset yang juga masih sedikit terbatas dan belum terlalu besar jumlahnya. Waktu pelatihan model dengan dataset yang tersedia juga tidak terlalu panjang membuat algoritma RF dapat menjadi algoritma pilihan untuk menyelesaikan permasalahan prediksi leads yang akan menjadi customer.

4.2 Saran

Penelitian ini masih mempunyai banyak kekurangan dan masih dapat dikembangkan lebih jauh. Pada penelitian ini masih belum menjelaskan secara detail mengapa suatu algoritma tertentu dapat menjadi algoritma terbaik dalam memprediksi leads yang akan menjadi customer. Oleh karena itu, baik adanya jika kedepannya dapat dilakukan penelitian yang mampu menjelaskan secara detail mengapa algoritma tertentu dapat menjadi algoritma terbaik dalam kasus memprediksi leads yang akan menjadi customer.

Referensi

[1] Luis G. Serrano, “Grokking Machine Learning: What is Machine Learning? It is common sense”, Manning Publication, 2021.

[2] Tim Editorial Rumah.com, “Lead Adalah: Ini Penjelasannya dan 5 Cara Jitu Mendapatkan Lead”,rumah.com/panduan-properti/lead-adalah-ini-penjelasannya-dan-5-cara-jitu-mendapatkan-lead-65744.

[3] Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E., “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics”, 24(1), 44-65, 2015.

[4] Maalouf, Maher. "Logistic regression in data analysis: an overview." International Journal of Data Analysis Techniques and Strategies 3.3 : 281-299, 2011.

[5] Luis G. Serrano, “Grokking Machine Learning: Support Vector Machine and the Kernel Method”, Manning Publication, 2021.

[6] Awad, Mariette, Rahul Khanna, Mariette Awad, and Rahul Khanna. "Support vector machines for classification." Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 2015.

[7] Binus Bandung, “Algoritma Naive Bayes,” [Online]. Available: https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes, 2022

[8] Jadhav, Sayali D., and H. P. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques." International Journal of Science and Research (IJSR) 5.1, 2016.

[9] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, “Decision tree classifier for network intrusion detection with GA-based feature selection,” in Proceedings of the 43rd annual Southeast regional conference- Volume 2, 2005.

[10] Shaik, Anjaneyulu Babu, and Sujatha Srinivasan. "A brief survey on random forest ensembles in classification model." International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2. Springer Singapore, 2019

[11] Jason Brownlee, “Machine Learning Algorithms From Scratch, Machine Learning Mastery,” 2016.