Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 3. Februari 2024

p-ISSN: 2301-5373
e-ISSN: 2654-5101

# Default Risk Prediction Using Decision Tree Study Case of Home Credit

Dewa Nyoman Agung Adipurwa Mahandiri[a1], Agus Muliantara[a2]

[a]Informatics Department, Faculty of Math and Natural Sciences, Udayana University
Bali, Indonesia
[1]adiagung707@email.com
[2]muliantara@unud.ac.id

***Abstract***

*Consuming loans with any service has become a trend in modern society. However, that trend gives some risk for the loan company such as Home Credit. Home Credit needs to create an automation analytic for predicting customers that might be default in future. So, we build a machine learning model using the Decision Tree algorithm to resolve that risk. The Decision Tree model can give mean score 85% accuracy, 91% precision ,and 92% recall score for Home Credit study case.*

***Keywords:*** *Credit, Default Risk Prediction, Machine Learning, Decision Tree, CRIPS DM*

## 1. Introduction

Seiring berjalannya perkembangan zaman, kualitas hidup masyarakat bergerak ke tingkat yang lebih tinggi disertai dengan perubahan gaya hidup yang semakin modern dan terbuka. Dengan hal tersebut, konsumsi pinjaman dengan berbagai metode seperti kredit ataupun *pay later* secara bertahap diterima oleh masyarakat. Berdasarkan data dari Statistik Fintech Lending Indonesia oleh Otoritas Jasa Keuangan (OJK), peningkatan entitas peminjam terus terjadi selama dua tahun terkahir. Peningkatan sebesar 19,8 juta peminjam tercatat dari Agustus 2021 hingga Agustus 2022. Kemudian peningkatan lebih besar terjadi pada rentang Agustus 2020 hingga Agustus 2021 sebesar 41 juta peminjam. Sejalan dengan peningkatan tersebut, pinjaman pribadi dengan layanan *peer-to-peer lending* cenderung lebih diminati oleh masyarakat terutama anak muda karena minimnya persyaratan dan prosedur sehingga waktu dibutuhkan lebih pendek, tetapi tetap aman sebab telah diawasi oleh OJK. Kemudahan tersebut menyebabkan banyak masyarakat yang mencoba untuk mendapatkan pinjaman. Namun, setiap lembaga atau perusahaan yang memberikan layanan peer-to-peer lending harus tetap membatasi dan memilah peminjam berdasarkan analisis kemampuan peminjam membayar kembali pinjaman. Pembatasan ini bertujuan untuk mengurangi kemungkinan risiko kegagalan kredit. Risiko kegagalan kredit yang paling sering terjadi pada layanan peer-to-peer lending adalah risiko kegagalan awal atau *default risk* [1].

*Default risk* adalah risiko yang diambil oleh pemberi pinjaman bila peminjam tidak dapat melakukan pembayaran yang diperlukan dari kewajiban utang pinjamannya [2]. Analisis terhadap default risk merupakan faktor penting dalam lembaga keuangan karena memungkinkan untuk memberikan pinjaman hanya kepada konsumen yang memiliki kredit baik. Analisis default risk secara konvensional dilakukan oleh lembangan dengan kartu penilain kredit pada setiap konsumen. Kartu penilaian kredit menganalisis secara statistik kelayakan kredit konsumen dan membantu lembaga keuangan untuk menolak atau memperpanjang kredit [3]. Dengan nilai kredit yang diperoleh, konsumen akan diklasifikasikan sebagai “kredit baik” hingga “kredit buruk”. Namun sebagaimana cara konvensional lainnya, hal tersebut akan sulit berlaku pada lembaga keuangan yang telah menerima ribuan bahkan ratusan juta kredit sehingga diperlukan teknologi untuk membantu otomatisasi prediksi default risk.

Machine learning adalah pendekatan yang tepat untuk data analisis dan prediksi yang terotomatisasi analisis konvensional. Sebagai subdomain dari a*rtificial intelligence, machine learning* dapat belajar dari data yang diberikan, mengidentifikasi pola, dan mengambil keputusan berdasarkan informasi yang diekstrak dengan meminimalisir intervensi manusia[4]. Dengan menggunakan *machine learning*,

proses analisis konsumen yang tepat untuk diberikan kredit dapat lebih cepat dilakukan bahkan dapat dilakukan prediksi berdasarkan catatan kredit yang pernah dilakukan sebelumnya ataupun hanya berdasarkan informasi dari latar belakang konsumen sebagai dasar data latihan untuk mesin dengan menggunakan algoritma tertentu. Berbagai algoritma dapat diuji untuk membangun model *machine learning* dengan akurasi, presisi, dan *recall* suatu model salah satunya decision tree. Banyak penelitian telah menggunakan algoritma *decision tree* untuk membangun model prediksi yang akurat. Oleh sebab itu, penelitian ini dilakukan untuk membentuk model prediksi *default risk* dari lembaga kredit Home Credit dengan menggunakan model *decision tree* dengan metode *Cross-Industry Standard Process for Data Mining* (CRISP DM).

## 2. Reseach Methods

*Cross-Industry Standard Process for Data Mining* adalah model proses data mining (data mining framework) yang diprakarsai oleh 5 perusahan yaitu Integral Solution Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA. Framework ini kemudian dikembangkan kembali oleh ratusan perusahaan dan organisasi sebagai metodologi terbuka (open source) bagi data mining. Model proses CRISP DM memberikan gambaran tentang siklus hidup pengembangan proyek data mining melalui 6 tahapan yaitu business understanding, data understanding, data preparation, modeling, evaluation, dan deployment [5].

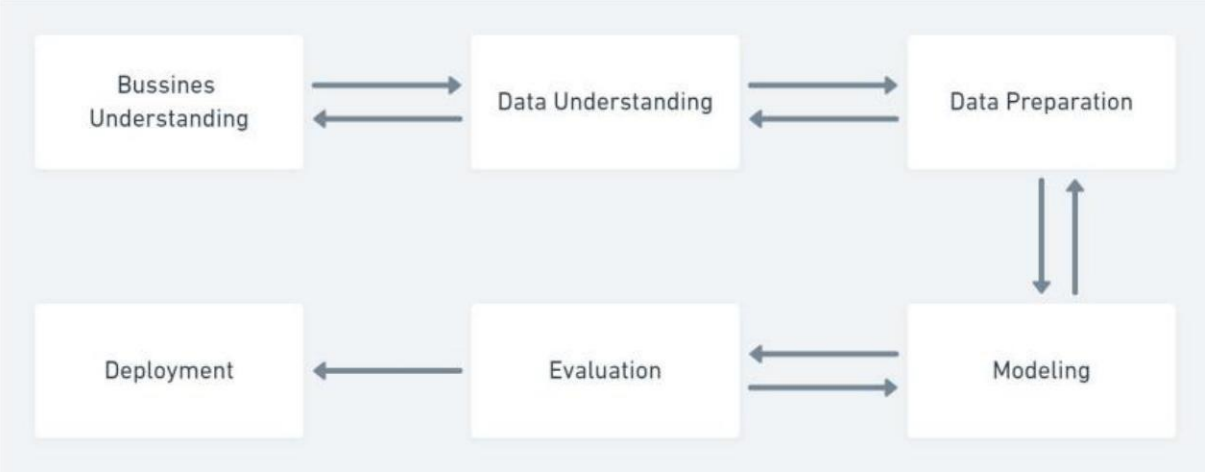



Figure 1. CRISP DM Diagram

### 2.1. Bussiness Understanding

Kebutuhan konsumsi terkadang datang secara insidental seperti mengganti alat rumah tangga utama, memperbaiki rumah, ataupun membayar pengeluaran yang tidak direncanakan. Dalam kondisi ini, konsumen mungkin kekurangan uang tunai untuk segera menangani kewajiban pembayaran. Oleh karena itu, layanan pinjaman hadir untuk memberi konsumen bantuan keuangan jangka pendek atau pencapaian tujuan jangka pendek.

Home Credit adalah penyedia layanan pinjaman *multi-channel* internasional untuk pembiayaan konsumen. Home Credit berupaya untuk memperluas ekspansi pelayanannya bagi masyarakat yang tidak memiliki rekening bank. Dengan penargetan lebih banyak konsumen maka kemungkinan Home Credit untuk meningkatkan dan menempati ini teratas sebagai penyedia pinjaman bukan bank. Namun, semakin luas cakupan penawaran perusahaan, semakin banyak risiko yang harus ditangani oleh Home Credit salah satunya *default risk*. Home Credit harus mampu menentukan kemungkinan konsumen yang akan mengalami *default*. Jika tingkat *default* tinggi, perusahaan akan mengalami kerugian dalam ekspansinya. Tujuan Home Credit adalah memberikan menawarkan pinjaman kepada individu yang berkemungkinan tinggi dapat membayar kembali pinjaman dan menolak permintaan pinjaman untuk yang mungkin mengalami *default* dimasa depan. Tantangan yang harus dihadapi adalah data dari latar belakang peminjam yang sangat beragam. Dalam penelitian ini, saya menggunakan *machine learning* dengan algoritma *decision tree* untuk memprediksi apakah konsumen akan mengalami default.

### 2.2. Data Understanding

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Kebutuhan Data terdiri atas 8 set dengan 2 data utama yaitu data latih “application_train.csv” dan data tes “application_test.csv”. Data latih terdiri atas 307.551 tinjauan dari 122 variabel. Pada data latih ini terdapat variabel “TARGET” yang yang menunjukan apakan konsumen mengalami kesulitan dalam membayar kredit. Selanjutnya, data tes terdiri atas 48744 tinjaun dan 121 variabel. Data test tidak mencakup variabel “TARGET” karena data tes akan digunakan untuk menguji output dari model apakah mampu menghasilkan prediksi target yang akurat. Secara tidak langsung, variabel target pada data latih telah menyatakan hampir semua konsumen dengan nilai variabel “TARGET” sama dengan 1 (benar) akan mengalami *default*. Data set lainnya adalah pecahan dari cakupan data latih dan data tes untuk mempermudah memahami setiap bagian variabel yang serumpun seperti dataset “berau.csv”, “credit_card_balance” , dan lainnya. Oleh sebab itu, untuk mempersingkat waktu penelitian maka data set hanya fokus pada data latih dan data tes sebagai data utama. Sebanyak
122 variabel pada data latih dapat dipecah menjadi 5 rumpun yaitu informasi pribadi konsumen, informasi pinjaman, informasi demografi konsumen, dokumen yang diberikan konsumen, dan pertanyaan yang diajukan ke biro kredit.

### 2.3. **Data Preparation**

Sebelum melakukan analisis, hal pertama yang perlu dilakukan adalah membuat data set mentah dapat digunakan. Langkah pertama dengan membersihkan data yaitu mencari cara untuk menangani data yang hilang atau bernilai Null. Penanganan dapat mempertimbangkan beberapa acuan yaitu makna keberadaan data dan besarnya data, sehingga dapat diputuskan apakah akan mengecuailkan variabel keseluruhan, menghapus satu baris selaras dengan data yang mengandung data hilang, atau melakukan imputasi pada baris data yang mengandung data hilang. Dalam analisis ini, data hilang diimputasi dengan menggunakan median dan modus. Median untuk jenis data numerik dan modus untuk jenis data kategorik atau objek, proses tersebut dilakukan untuk kedua data utama. Jumlah data hilang tertinggi pada kedua jenis data lebih dari 21.000.

Langkah kedua, mengubah data kategorik menjadi data numerik agar dapat diperhitungkan secara komputasi dengan cara melabelkan setiap variasi. Variabel “GENDER” akan diubah menjadi 0 untuk perempuan (F) dan 1 untuk laki-laki (M), serta data kategorikal lainnya.

Langkah ketiga, menemukan dan menangani outliers dengan metode IQR.

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1 \quad (1)$$

$$\text{Lower Bound} = (\mathrm{Q}_1 - (1.5 \times \mathrm{IQR}) \quad (2)$$

$$\text{Upper Bound} = (\mathrm{Q}_3 + (1.5 \times \mathrm{IQR}) \quad (3)$$

$$\text{Lower Bound} < \text{Data Outliers} < \text{Upper Bound} \quad (4)$$

Data outliers dapat mengganggu model dan menyebabkan bias pada model sehingga perlu ditangani. Untuk menangani data outliers dapat digunakan teknik yang sama dengan penanganan data hilang. Namun, penangangan data outliers perlu mempertimbangkan apakah data tersebut time series atau tidak memiliki hubungan yang cukup berpengaruh antara data lainnya. Pada penelitian ini, penanganan dirasa tidak perlu dilakukan karena cukup banyak data yang bersifat times series dan tidak ada variabel dengan data outliers yang sangat tinggi hingga mendekati setengah dari data keseluruhan.

Langkah terakhir adalah visualisasi data yang telah siap untuk digunakan, proses visualisasi ini akan dapat digunakan lagi untuk melakukan data understanding atau dapat melanjutkan langkah berikutnya pada modeling.

## 2.4. Modeling

Berdasarkan tujuan pada pemahaman bisnis, penyelesaian masalah yang diperlukan adalah menentukan apakah konsumen dikategorikan akan *default* atau tidak. Dengan demikian, penyelesaian masalah ini berupa pembentukan model kalsifikasi.

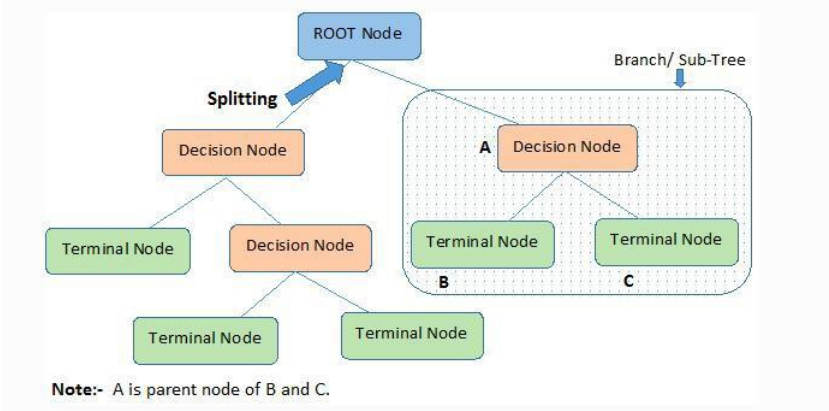



Figure 2. Decision Tree Plot

Pada penelitian ini, decision tree akan digunakan sebagai pilihan model klasifikasi untuk menentukan apakah konsumen tersebut bernilai "Ya" atau "Tidak" sebagai konsumen yang berpotensi akan default. Proses modeling terdiri atas, pembuangan variabel yang tidak dibutuhkan dalam pembentukan model pada data latih. Selanjutnya, pembagian data menjadi data untuk latih dan validasi berdasarkan data set latih dengan rasio 25%-50% untuk data validasi [6]. Berdasarkan partisi data tersebut, diperoleh 206.032 data latih dan 101.479 data validasi. Dengan menggunakan modul sklearn, model decision tree dibentuk sebagai berikut

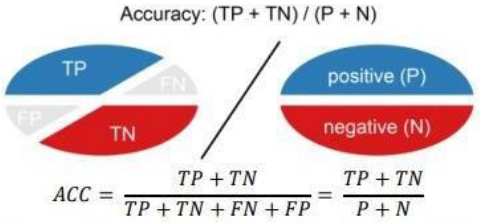
```
#Buat data x dan y
x=data_train.drop(['TARGET','SK_ID_CURR'],axis=1)

y=data_train['TARGET']

#Partisi data dengan 67:33
x_train, x_valid, y_train, y_valid = train_test_split(x,y,test_size=.33,random_state=1)

#Panggil modul untuk model Decision Tree
model = DecisionTreeClassifier()
model.fit(x_train, y_train)

#Lakukan prediksi model
y_pred_dt = model.predict(x_valid)
```

Table 1. Model Decision Tree

Setelah model berhasil dibangun, maka perlu evaluasi untuk menilai model apakah sudah baik atau belum

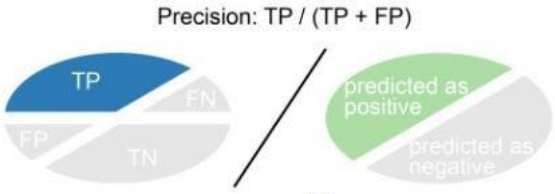
## 2.5. Evaluation

Evaluasi digunakan untuk menilai kinerja dari model machine learning. Ada banyak metode atau cara untuk menilai kinerja sebuah model. Secara umum, motode yang sering digunakan untuk menilai kinerja model klasifikasi yaitu accuracy, precision, recall.



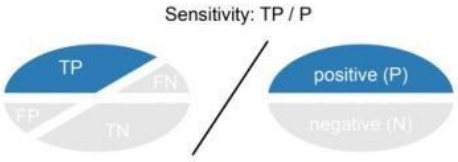
p-ISSN: 2301-5373
e-ISSN: 2654-5101

Accuracy adalah perhitungan jumlah dari dua prediksi akurat (TP + TN) dibagi banyak data set (P+N). Akurasi memiliki rentang dari terbaik sampai terburuk (1.0 - 0.00) [7].




$$PREC = \frac{TP}{TP + FP}$$

Precision adalah perhitungan nilai prediksi benar positif (TP), dibagi dengan total nilai prediksi positif (TP + FP) [7].




$$SN = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Recall adalah perhitungan nilai prediksi benar positif (TP), dibagi dengan nilai positif (P) [7].

### 2.6. Deployment

Tahap terakhir model akan dibentuk menjadi sebuah servis berupa API menggunakan Fast API. Sebelum dapat dihadirkan dalam API, model yang telah siap digunakan harus disimpan dalam disk dengan bentuk file. Salah satu bentuk yang umum digunakan adalah model.sav dengan menggunakan modul pickle.

## 3. Result and Discussion

Adapun beberapa temuan dari penelitian tentang default risk prediction berdasarkan data set Home Credit. Pada penelitian ini model yang diimplementasikan adalah Decision Tree dengan beberapa eksperimen untuk memperoleh kinerja model Decision Tree terbaik.

Eksperimen pertama dengan melakukan partisi 80% data latih dan 20% data validasi

| | DecisionTree 80:20 |
|---|---|
| accuracy_score | 0.851259 |
| recall_score | 0.926132 |
| precision_score | 0.910831 |

Eksperimen pertama dengan melakukan partisi 70% data latih dan 30% data validasi

| | DecisionTree 70:30 |
|---|---|
| accuracy_score | 0.850836 |
| recall_score | 0.925846 |
| precision_score | 0.910635 |

Eksperimen pertama dengan melakukan partisi 67% data latih dan 33% data validasi

| | DecisionTree 67:33 |
|---|---|
| accuracy_score | 0.851575 |
| recall_score | 0.925877 |
| precision_score | 0.911505 |

Eksperimen pertama dengan melakukan partisi 60% data latih dan 40% data validasi

| | DecisionTree 60:40 |
|---|---|
| accuracy_score | 0.850022 |
| recall_score | 0.925419 |
| precision_score | 0.910196 |

Eksperimen pertama dengan melakukan partisi 50% data latih dan 50% data validasi

| | DecisionTree 50:50 |
|---|---|
| accuracy_score | 0.849736 |
| recall_score | 0.925401 |
| precision_score | 0.909899 |

Berdasarkan eksperimen model yang dibangun menggunakan algorimat Decision Tree dengan bergaram rasio partisi menghasilkan rata rata nilai akurasi 85%. Nilai akurasi 85% sudah tergolong nilai yang sangat baik untuk akurasi suatu model machine learning yang disertai dengan rata-rata recall sebesar 92% dan precision sebesar 91%

## 4. Conclusion

Pada penelitian ini, dataset dibagi menjadi berbagai rasio untuk menemukan model Decision Tree terbaik dalam mendeteksi risiko default pada konsumen berdasarkan 5 bagian dara konsumen yaitu informasi pribadi konsumen, informasi pinjaman, informasi demografi konsumen, dokumen yang diberikan konsumen, dan pertanyaan yang diajukan ke biro kredit. Dengan melakukan eksperimen tersebut diperoleh model Decision Tree dengan rata-rata akurasi yaitu 85% diikut nilai recall dan precision masing-masing sebesar 92% dan 91%.. Dengan demikian, model ini sudah cukup baik untuk digunakan memprediksi risiko *default* pada konsumen. Meskipun model ini sudah cukup baik, kami rasa masih perlu penelitian lebih lanjut terkait kemungkinan model lainnya dapat memperoleh nilai akurasi yang lebih tinggi.

## References

[1] A. Krichene, “Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank,” *Journal of Economics, Finance and Administrative Science*, vol. 22, no. 42, pp. 3–24, 2017, doi: 10.1108/JEFAS-02-2017-0039.

[2] Y. R. Chen, J. S. Leu, S. A. Huang, J. T. Wang, and J. I. Takada, “Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3079701.

[3] Z. Khemais, D. Nesrine, and M. Mohamed, "Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression," *Int J Econ Finance*, vol. 8, no. 4, 2016, doi: 10.5539/ijef.v8n4p39.

[4] J. Y. Seo, "Machine Learning in Consumer Credit Risk Analysis: A Review," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, 2020, doi: 10.30534/ijatcse/2020/328942020.

[5] M. F. M. Salleh and T. Suryanto, "Fraud Detection on Banking Industry in South Sumatera: A Study on the Role of Internal Auditors'," *International Journal of Shari'ah and Corporate Governance Research*, vol. 2, no. 2, 2019, doi: 10.46281/ijscgr.v2i2.399.

[6] R. R. Picard and K. N. Berk, "Data splitting," *American Statistician*, vol. 44, no. 2, 1990, doi: 10.1080/00031305.1990.10475704.

[7] Ž. Vujović, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

*This page is intentionally left blank.*