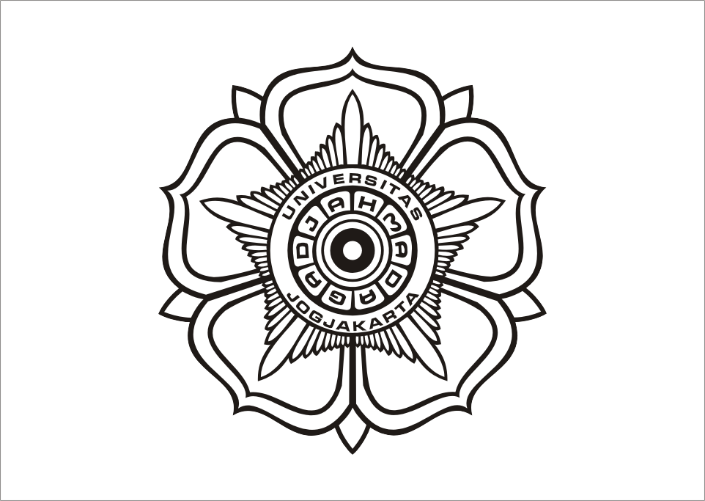
**Makalah Kompetisi Data Mining**

**C-Compiler 2018**

****

**Oleh :**

**KNearestUkhti**

M Fawwaz Mayda (Ilmu Komputer 2017)

Rimba Erlangga (Ilmu Komputer 2016)

1. **Latar Belakang**

Latar belakang pembuatan makalah ini didasari oleh salah satu prasyarat peserta dalam mengikuti perlombaan C-Compiler bidang Data Mining. Terdapat dua penilaian dalam perlombaan ini, yaitu prediksi dataset dan satu lagi adalah pengerjaan makalah ini, yang berisi penjelasan prediksi dataset yang kami buat.

1. **Tujuan dan Manfaat**

Tujuan dan manfaat yang bisa diambil dari kegiatan ini, harapannya adalah sebagai pengukur kemampuan, menambah pengetahuan tentang berbagai algoritma *Machine Learning.* Serta tidak kalah jauh pentingnya menjadi latihan mempraktikkan teori yang telah dipelajari.

1. **Metode Data Mining**

Pada dataset yang diberikan, terdapat 24 *feature/*kolom yang digunakan sebagai faktor untuk prediksi, dan satu kolom untuk target prediksi. Namun, kami menyadari bahwa tidak semua kolom berpengaruh pada hasil prediksi. Akhirnya, sebagai langkah awal, kami memutuskan untuk melakukan *feature selection,* yaitu memilah kolom atau faktor mana saja yang berpengaruh kuat terhadap hasil prediksi. Bagaimana cara kami dalam memilah faktor ini akan dijelaskan di bawah.

Setelah melakukan *feature selection,* kami mengumpulkan semua model/algoritma yang mungkin digunakan dalam proses prediksi. Masalah prediksi kali ini adalah masalah klasifikasi. Model yang kami kumpulkan antara lain:

1. K-Nearest Neighbor
2. Support Vector Machine
3. Decision Tree
4. Random Forest

Namun, pada akhirnya kami memutuskan menggunakan model Decision Tree dikarenakan cukup baik dan cepat dalam waktu training.

1. **Analisis**

Dataset yang diberikan pada kompetisi ini adalah dataset tentang rincian data nasabah kartu kredit yang disajikan dalam kolom A sampai X, serta sebuah kolom Y berupa target prediksi apakah kartu kredit nasabah yang bersangkutan akan dinonaktifkan atau terus dilanjutkan. Namun, karena penamaan kolom dari A sampai Y yang tidak representatif, kami mengubah nama kolom pada data tersebut terlebih dahulu sebagai berikut:

A : 'ID',

kategorik, yaitu ID unik untuk setiap nasabah.

Kolom ini hanya berguna untuk identifikasi dan tidak akan kami gunakan dalam klasifikasi nantinya.

B : 'LIMIT\_BAL',

numerik, yaitu jumlah kredit yang tersisa.

C : 'SEX',

kategorik, jenis kelamin nasabah, yaitu 1 untuk laki-laki dan 2 untuk perempuan .

D : 'EDUCATION',

kategorik, tingkat pendidikan terakhir nasabah,

1 = Master, 2 = Sarjana, 3 = SMA, 4 = Lainnya

E : 'MARRIAGE',

kategorik, status pernikahan nasabah

1 = Sudah Menikah, 2 = Belum Menikah, 3 = Lainnya

F : 'AGE',

numerik, umur nasabah

G-L : 'PAY\_1', 'PAY\_2', 'PAY\_3', 'PAY\_4', 'PAY\_5', 'PAY\_6',

numerik, yaitu ketepatan si nasabah untuk membayar tagihan. ‘PAY\_1’ adalah seberapa tepat nasabah membayar tagihan di bulan pertama dan seterusnya untuk ‘PAY\_6’.

M-R : 'BILL\_AMT1', 'BILL\_AMT2', 'BILL\_AMT3', 'BILL\_AMT4',

'BILL\_AMT5', 'BILL\_AMT6',

Merupakan data numerik, yang merupakan tagihan dibulan ke 1 (‘BILL\_AMT1’)

hingga bulan ke-6 (‘BILL\_AMT6’)

S-X : 'PAY\_AMT1', 'PAY\_AMT2', 'PAY\_AMT3', 'PAY\_AMT4',

'PAY\_AMT5', 'PAY\_AMT6',

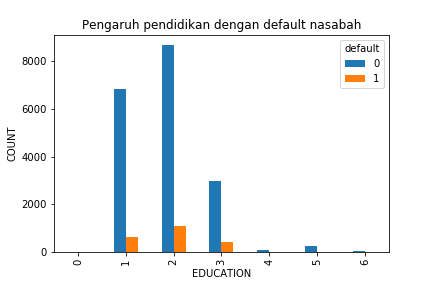
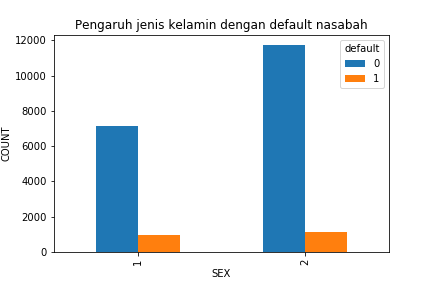
Merupakan data numerik, yang menyatakan jumlah yang dibayar si nasabah pada

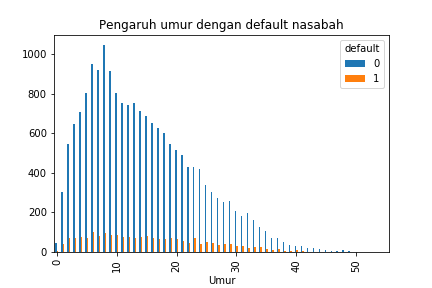
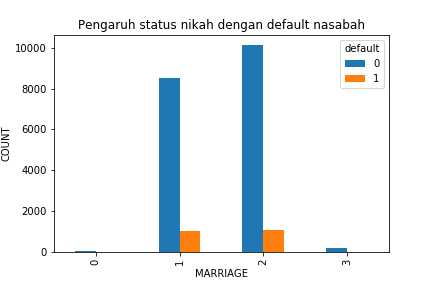
bulan ke 1(‘PAY\_AMT1’) hingga bulan ke-6 (‘PAY\_AMT6’).

Y : 'default'

status yang diberlakukan pada kartu kredit nasabah

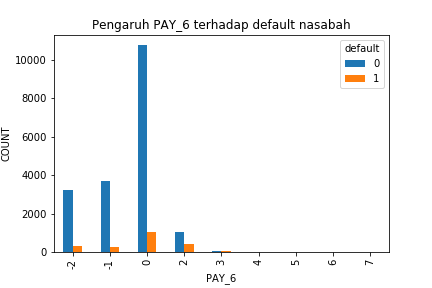
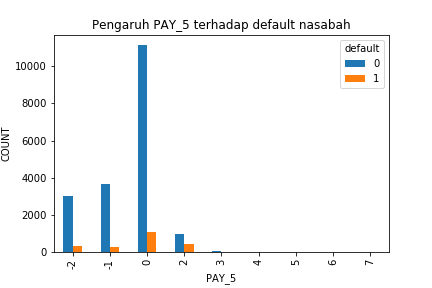
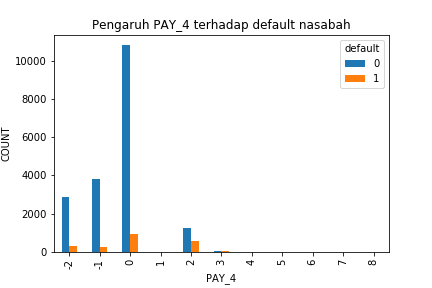
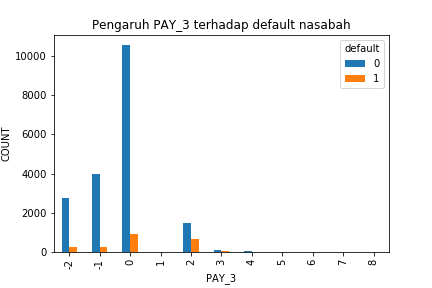
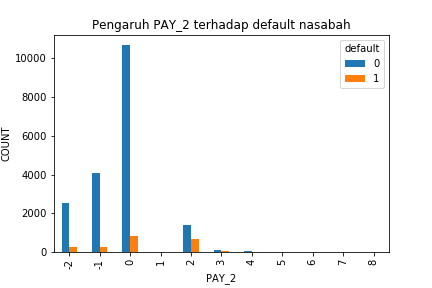
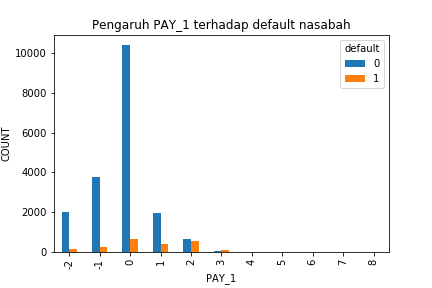
Kemudian kami membuat grafik distribusi dan scatter-plot untuk kolom Y dengan setiap kolom lain. Hasilnya dapat dilihat pada gambar-gambar berikut.





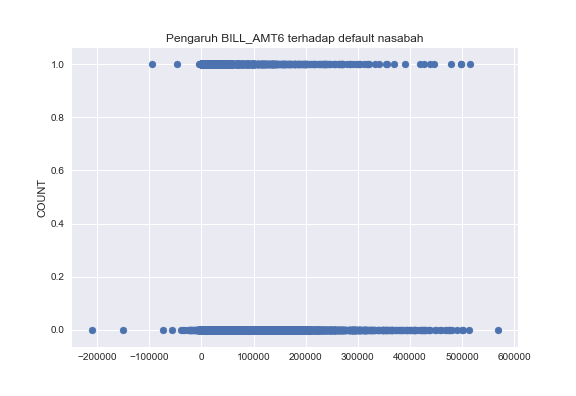
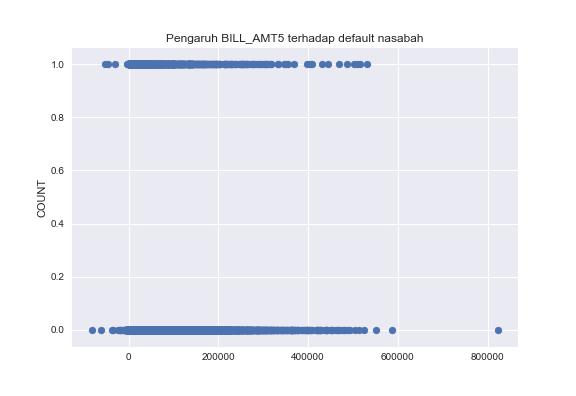
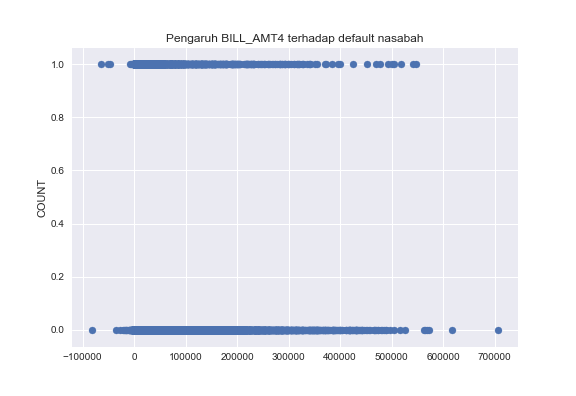
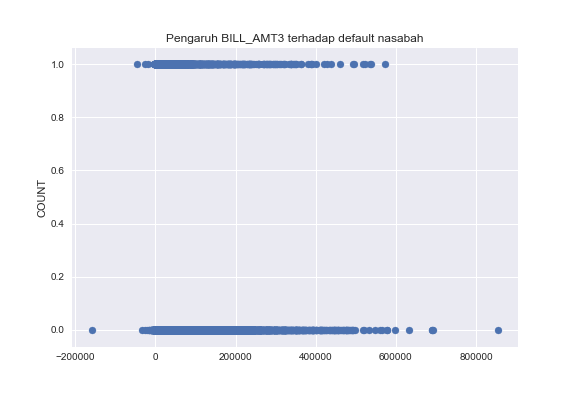
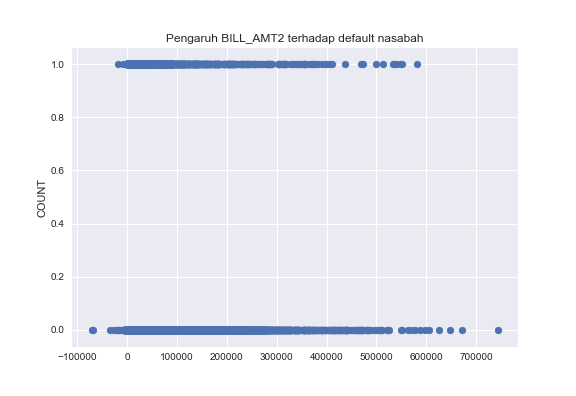
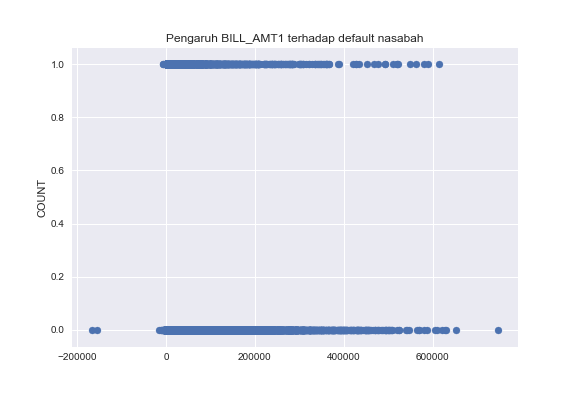
Dilihat dari sini kami mendapat bahwa hasil dari grafik pengaruh kolom-kolom diatas terhadap *default* dari nasabah harus ikut dipertimbangkan dikarenakan ada beberapa *value* dapat memberikan hasil prediksi yang berbeda.

Untuk kolom ‘PAY\_\*’ kami menggunakan *Cross Tabulation* sama seperti kolom sebelumnya untuk mendapat grafik dari pengaruh kolom tersebut terhadap *default* nasabah.



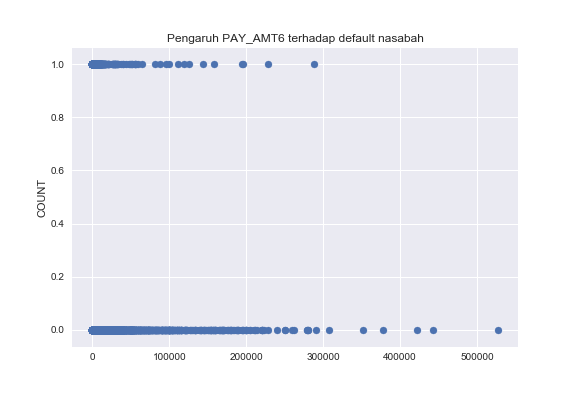
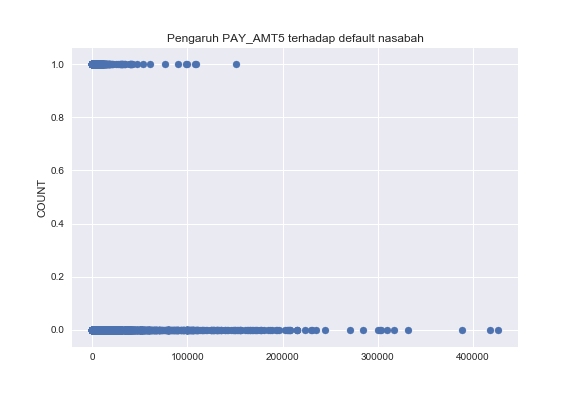
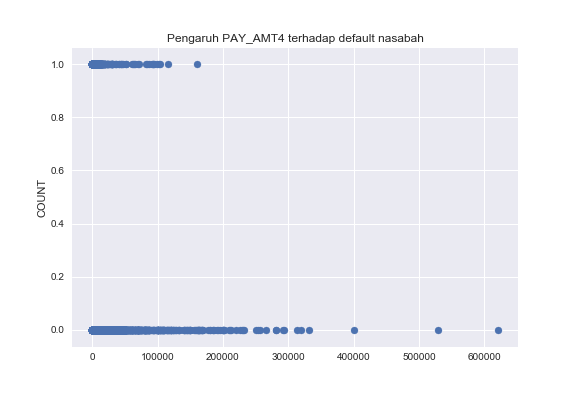
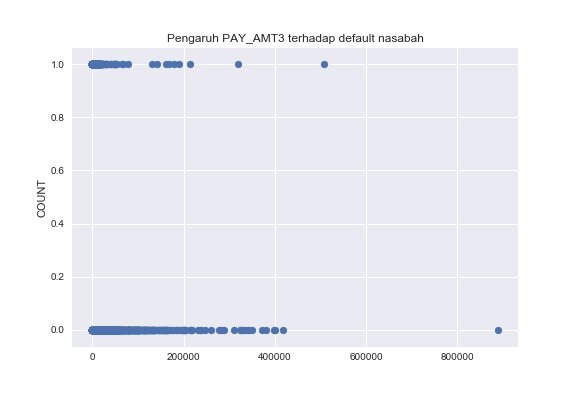
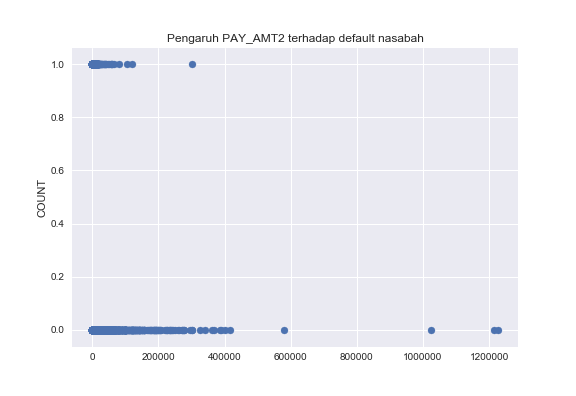
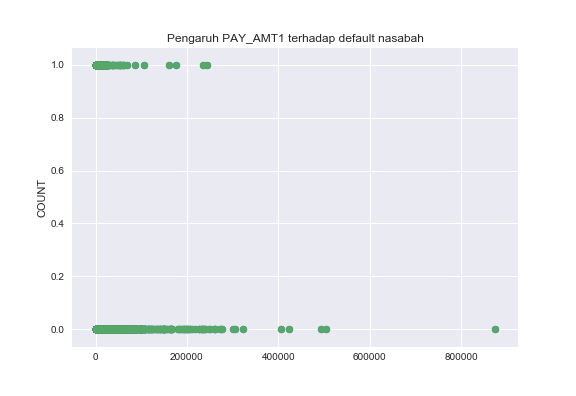
Dari kolom diatas kami menyimpulkan bahwa tidak semua kolom memberikan gambaran yang bagus untuk prediksi kali ini. Disini kami mengamati bahwa kolom ‘PAY\_1’ memberikan dampak yang cukup besar dalam hasil prediksi diikuti oleh kolom ‘PAY\_2’ yang sebenarnya agak kurang lebih dengan kolom ‘PAY\_3’ dan ‘PAY\_4’. Lalu kami juga memilih kolom ‘PAY\_6’ untuk digunakan dalam prediksi dibandingkan dengan kolom ‘PAY\_5’ dikarenakan adanya perbedaan hasil dimana kami melihat kolom ‘PAY\_6’ menghasilkan data yang lebih akurat.

Untuk kolom ‘BILL\_AMT\*’ kami menggunakan *Scatter Plot* untuk mendapatkan grafik pengaruh terhadap default nasabah.



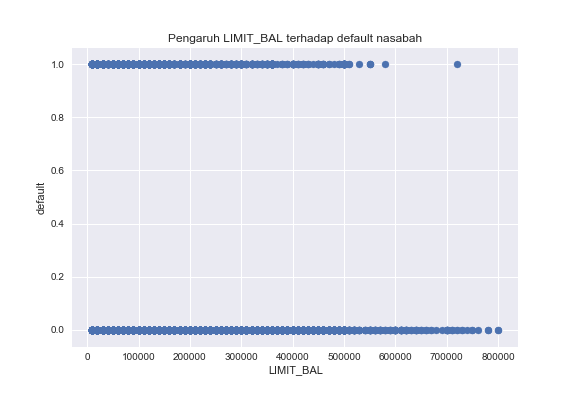
Diantara kolom yang ada kami memilih kolom ‘BILL\_AMT1’, dikarenakan cukup untuk menggambarkan distrubusi data diantara semua kolom ‘BILL\_AMT’ kami hanya memilih kolom ini dikarenakan kolom lainnya terlihat hampir sama.dalam hal kontribusi terhadapa nilai *default.*

Lalu kami juga menggunakan *Scatter Plot* untuk kolom ‘PAY\_AMT\*’ berikut grafik nya:



Kami memutuskan untuk tidak menggunakan kolom ini dikarenakan banyaknya *outliers* yang kami rasa dapat menganggu kinerja algoritma.

Dan kolom terakhir yang akan kami gunakan adalah kolom ‘LIMIT\_BAL’ dikarenakan secara logika jumlah kredit yang dimiliki nasabah akan berkontribusi apakah status kredit nasabah akan masih aktif atau tidak.



Dapat dilihat bahwa, faktor-faktor yang paling berpengaruh yaitu:

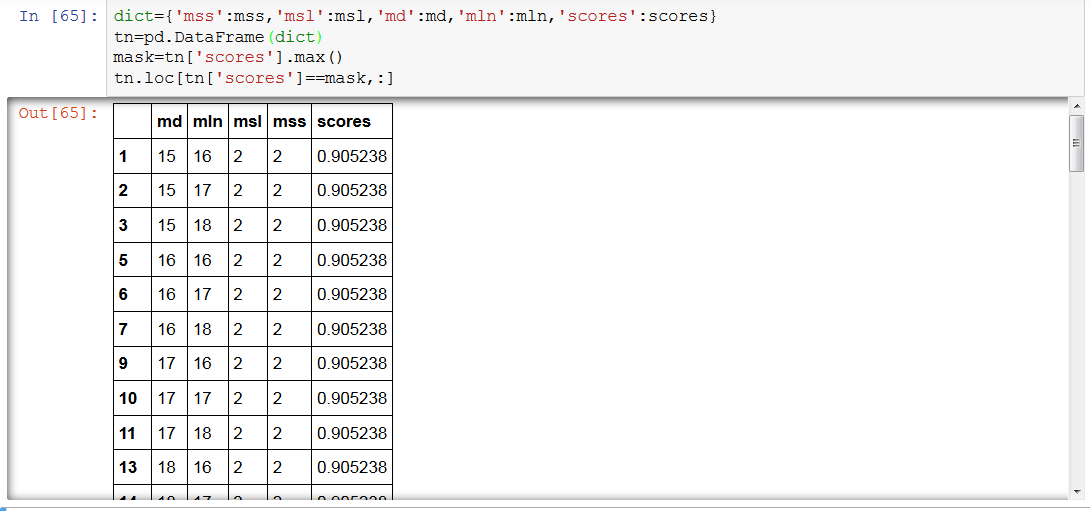
* Limit bal
* Sex
* Marriage
* Age
* Education
* Pay 1, 2, 6
* Bill amt 1

karena alasan yang telah dijelaskan diatas.

Kemudian, pada model yang kami pilih, yaitu Decision Tree, kami lakukan proses training dengan dataset yang sudah diproses dengan *feature selection* tadi*.* Setiap model juga kami lakukan *parameter tuning* untuk meningkatkan performa prediksi. Adapun, parameter yang kami *tuning* pada setiap model beserta *range* nilainya adalah: min\_sample\_split, min\_sample\_leaf, max\_depth, max\_leaf\_nodes



Setelah kami proses, didapat bahwa model terbaik beserta parameternya ada berbagai macam disini hasil tadi kami ubah dalam bentuk table untuk mempermudah pengamatan.



Disini kami coba ambil baris ke-3 untuk sebagai hasil akhir parameter dalam proses prediksi.

1. **Kesimpulan**

Kesimpulan disini adalah bahwa dalam bidang Data Mining tidak hanya diperlukan algoritma tetapi lebih penting adalah analisis data yang bisa memberikan pengaruh terhadap akurasi yang lebih baik. Ini berkaca dari pengalaman kami pada saat menggunakan algoritma Decision Tree tanpa feature selection kami hanya mendapat akurasi sekitar 57% lalu setelah menggunakan feature selection kami dapati akurasi nya meningkat menjadi 60%.