TESIS

KLASIFIKASI KELAYAKAN KREDIT BANK MENGGUNAKAN METODE NAIVE BAYES DENGAN PENGELOMPOKAN METODE K-MEANS

Disusun oleh:

Nama	**: Taufik Fitriyadi**
NIM	**: 18.51.1162**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2021

TESIS

KLASIFIKASI KELAYAKAN KREDIT BANK MENGGUNAKAN METODE NAIVE BAYES DENGAN PENGELOMPOKAN METODE K-MEANS

CLASSIFICATION OF FEASIBILITY OF BANK CREDIT USING THE NAIVE BAYES METHOD WITH GROUPING K-MEANS METHOD

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Taufik Fitriyadi**
NIM : **18.51.1162**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2021

HALAMAN PENGESAHAN

KLASIFIKASI KELAYAKAN KREDIT BANK MENGGUNAKAN METODE NAIVE BAYES DENGAN PENGELOMPOKAN METODE K-MEANS

CLASSIFICATION OF FEASIBILITY OF BANK CREDIT USING THE NAIVE BAYES METHOD WITH GROUPING K-MEANS METHOD

Dipersiapkan dan Disusun oleh

Taufik Fitriyadi

18.51.1162

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, 04 Februari 2021

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 Februari 2021
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

KLASIFIKASI KELAYAKAN KREDIT BANK MENGGUNAKAN METODE NAIVE BAYES DENGAN PENGELOMPOKAN METODE K-MEANS

CLASSIFICATION OF FEASIBILITY OF BANK CREDIT USING THE NAIVE BAYES METHOD WITH GROUPING K-MEANS METHOD

Dipersiapkan dan Disusun oleh

Taufik Fitriyadi

18.51.1162

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 04 Februari 2021

Pembimbing Utama

Dr. Kusrini, M.Kom
NIK. 190302106

Anggota Tim Penguji

Dr. Wing Wahyu Winarno, MAFIS, Ak.
NIK. 555195

Dr. Andi Sunyoto, M.Kom
NIK. 190302052

Pembimbing Pendamping

M. Rudyanto Arief, S.T, M.T
NIK. 190302098

Dr. Kusrini, M.Kom
NIK. 190302106

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 Februari 2021
Direktur Program Pascasarjana

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Taufik Fitriyadi
NIM : 18.51.1162
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Tuliskan Judul Tesis Bahasa Indonesia

Dosen Pembimbing Utama : Dr. Kusrini, M. Kom
Dosen Pembimbing Pendamping : M. Rudyanto Arief, S.T, M.T

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 04 Februari 2021
Yang Menyatakan,

Taufik Fitriyadi

HALAMAN PERSEMBAHAN

Penelitian tesis ini saya persembahkan kepada Allah Subhanahu wata'ala sebagai bentuk syukur saya terhadap ilmu yang saya dapatkan, saya jabarkan pada laporan ini agar berguna dalam konstribusi ilmu bidang IT. Semoga dapat diterima sebagai suatu amal kebaikan. Selanjutnya karya ini saya persembahkan kepada kedua Orang Tua tersayang Ayah dan Ibu, Keluarga kecil saya teruntuk Istri dan Anak atas segala bentuk dukungan, do'a dan kebaikan yang dilakukan sehingga memberikan saya energi positif dan saya dapat menyelesaikan studi serta penelitian ini dengan baik.

Penelitian ini juga saya persembahkan untuk almamater saya, Universitas AMIKOM Yogyakarta dan juga para pembaca semoga semua yang terdapat dalam naskah laporan penelitian tesis ini dapat memberikan wawasan tambahan dan kontribusi keilmuan yang baik dan bermanfaat.

HALAMAN MOTTO

"*Pengetahuan yang baik adalah yang memberi manfaat, bukan yang hanya diingat.*" - Imam Al-Safi'i

"*Sebaik-Baik Manusia Adalah Yang Paling Bermanfaat Bagi Manusia*" - HR. Ahmad

KATA PENGANTAR

Segala puji penulis penjatkan kepada Allah Subhana Wata'ala atas segala limpahan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Tesis ini dengan judul "Klasifikasi Kelayakan Kredit Bank Menggunakan Metode Naive Bayes Dengan Pengelompokan Metode k-Means" ini.

Penyusunan laporan ini tidak lepas atas bimbingan dan sumbangsih dari berbagai pihak. Oleh karena itu dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas Amikom Yogyakarta yang berkenan memberika kesempatan untuk menimba ilmu di Universitas AMIKOM Yogyakarta ini.
2. Ibu Dr. Kusrini, M.Kom. selaku dosen pembimbing utama dan bapak M. Rudyanto Arief, MT selaku dosen pembimbing pendamping yang telah banyak memberikan ilmu, waktu, dan segenap perhatiannya selama membimbing saya dalam menyelesaikan tesis ini.
3. Bapak Dr. Wing Wahyu Winarno, MAFIS, Ak. dan bapak Dr. Andi Sunyoto, M.Kom selaku dosen penguji yang telah banyak memberikan ilmu, waktu, dan segenap perhatiannya sera saran dalam proses perbaikan tesis ini.
4. Kedua orang tua bapak suwar dan ibu satem, kedua mertua saya bapak kiban dan ibu sutiyem, beserta adik (Fakhri, Adit, Nina) dan tak lupa istri saya tercinta Leni Apriliyani beeserta anak saya Arsennio yang selalu mendukung dan memberikan doa dan motivasinya.

5. Rekan-rekan mahasiswa Universitas AMIKOM Yogyakarta yang telah berjuang bersama menyelesaikan studi S2.
6. Segenap Dosen Universitas AMIKOM Yogyakarta yang telah banyak memberikan ilmu selama menimba ilmu dikampus ini.
7. Terakhir kepada semua pihak yang telah membantu, yang tadak dapat diuraikan satu persatu.

Penulis menyadari bahwa masih ada langit diatas langit, dan begitu juga dengan Tesis ini yang masih sangat perlu disempurnakan dan dikembangkan lagi. Oleh karena itu, penulis membuka diri untuk saran dan kritik yang membangun atas nama ilmu pengetahuan. Penulis juga berhadap bahwa tesis ini dapat dijadikan rujukan ataupun sumbangsi terhadap ilmu pengetahuan untuk dikemudian hari.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Yogyakarta, 04 Februari 2021

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Bank adalah salah satu fasilitas yang menyediakan pinjaman bagi calon nasabah. Nasabah membutuhkan kredit untuk sejumlah kepentingan misalnya kebutuhan pokok, modal usaha ataupun guna keperluan lainnya. Dalam proses kegiatan nya lembaga Bank seringkali mendapat masalah sehingga menimbulkan keputusan yang salah dapat menyebabkan proses pembiayaan angsuran macet serta menghasilkan kerugian. Dalam penelitian ini penulis menggunakan metode Naive Bayes dan K-Means dalam melaksanakan penentuan pengelompok kan kelayakan kredit.

Pengumpulan data yang dilaksanakan menggunakan studi dokumen dengan mengambil German Credit Data UCI machine learning. Data yang diperlukan terdapat 1000 record untuk dilakukan sebagai data latihan serta data pengujian. Algoritma K-Means berfungsi dalam tahap pengelompokan dengan atribut yang bersifat numerik sebanyak 3 atribut meliputi age, credit, duration credit.

Hasil dari pengelompokan K-Means berturut - urut dari variabel age, credit dan duration credit sebesar 3, 4 dan 3 kelompok. Metode Naive bayes berfungsi dalam tahap penilaian probabilitas. Hasil penelitian yang dilakukan menghasilkan akurasi nilai 75,01 % dengan penggunaan metode Naive Bayes ditambah pengelompokan K-means Clustering, Sedangkan dengan perhitungan Naive bayes dengan fungsi Gauss menghasilkan nilai rata-rata 74,91%.

Kata Kunci: Naive Bayes, Kredit, Bank

ABSTRACT

A bank is one of the facilities that provides loans for prospective customers. Customers need credit for a number of purposes, such as basic needs, business capital or other purposes. In the process of its activities, the Bank institution often gets into problems, causing wrong decisions, which can cause the installment financing process to stall and result in losses. In this study, the authors used the Naive Bayes and K-Means method in determining creditworthiness grouping.

Data collection was carried out using document studies by taking German Credit Data UCI machine learning. The required data contains 1000 records to be carried out as training data and testing data. The K-Means algorithm functions in the grouping stage with 3 numeric attributes including age, credit, duration credit.

The results of the K-Means grouping in order of age, credit and duration credit variables were 3, 4 and 3 groups. The Naive Bayes method functions in the probability assessment stage. The results of the research conducted resulted in an accuracy of 75.01% with the use of the Naive Bayes method plus K-means clustering, while the calculation of Naive Bayes with the Gauss function resulted in an average value of 74.91%.

Keyword: Naive Bayes, Credit, Bank

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Bank dikenal sebagai lembaga keuangan yang kegiatannya tidak hanya menghimpun dana dan menyalurkan kembali kepada masyarakat melainkan memberikan berbagai fasilitas perbankan (Imran, 2013). UU Perbankan No. 10 tahun 1998 menerangkan bahwa bank adalah badan usaha yang kegiatannya menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk kredit dan atau bentuk–bentuk lainnya dalam rangka meningkatkan taraf hidup rakyat banyak. Keberadaan bank sangat dibutuhkan tidak hanya bagi masyarakat kalangan bawah tapi hampir semua kalangan.

Meningkatnya perekonomian saat ini membuat banyaknya pengeluaran yang harus ditanggung oleh setiaporang. Kebutuhan yang semakin hari semakin banyak dan membuat setiap orang membutuhkan biaya dalam mencukupi kebutuhan sehari-hari maupun biaya untuk menjalankan usahanya agar tetap berjalan terus. Salah satu upaya yang bisa ditempuh adalah melakukan pinjaman kepada bank untuk mengatasi permasalahan yang ada. Perbankan menyediakan berbagai macam alternatif pinjaman uang bagi nasabah salah satunya adalah melalui pemberian pinjaman berupa kredit kepada nasabah (Hasan, 2017).

Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu berdasarkan persetujuan atau kesepakatan pinjam-meminjam antar bank dengan pihak lain yang mewajibkan pihak peminjam melunasi hutangnya

setelah jangka waktu tertentu dengan pemberian bunga (Marhumi, 2017). Dalam dunia perbankan keberadaan kredit merupakan hal yang tak lepas dari salah satu bentuk usaha yang dijalankan oleh dunia perbankan. Salah satu pendapatan utama bank merupakan adanya fasilitas kredit untuk mengambil keuntungan dari beban bunga yang ditangguhkan kepada nasabah.

Dalam pelaksanaannya, kredit yang bermasalah (kredit macet) sering terjadi akibat analisis kredit yang tidak hati-hati atau kurang cermat dalam proses pemberian kredit, maupun dari karakter nasabah yang tidak baik. Dalam menjalankan bisnisnya penting bagi bank dan lembaga pembiayaan untuk mengevaluasi risiko kredit yang dilakukan oleh. Dalam rangka mencegah terjadinya kredit macet, seorang analisis kredit perbankan harus mampu mengambil keputusan yang tepat untuk menerima ataupun menolak pengajuan kredit. Sebuah model yang baik bagi penilaian kredit akan membantu bank dan lembaga pembiayaan membuat keputusan yang tepat dalam rangka menghindari potensi besarnya risiko (Defu, Stephen, & Zhimei, 2008). Dalam pemberian kredit kepada nasabah, pihak bank mengalami berbagai masalah atau risiko. Salah satu masalah atau risiko yang dialami bank dalam pemberian kredit adalah perilaku nasabah yang tidak membayar angsuran tepat waktu ataupun menunda sampai beberapa bulan pembayaran angsuran yang pada akhirnya menyebabkan kredit macet. Hal ini merupakan masalah yang serius yang perlu diperhatikan oleh pihak bank untuk lebih berhati-hati dalam menentukan nasabah karena dalam pemberian kredit sangat berisiko. Faktor lainnya adalah situasi ekonomi yang terjadi

sehingga mengganggu usaha yang dijalankannya. Dalam arti luas risiko kredit adalah ketidakpastian atau fluktuasi laba dalam kegiatan kredit (Yu, 2007).

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan pola atau hubungan dalam set data yang berukuran besar (Santosa, 2007). Dengan menggunakan teknologi di bidang data mining yang meng optimasi proses pencarian informasi prediksi dalam basis data yang besar, serta menemukan pola-pola yang tidak diketahui sebelumnya. Penggunaan data mining sudah banyak dilakukan untuk memecahkan berbagai kasus yang ada pada permasalahan dunia bisnis. Data mining dapat digunakan pada beberapa kasus yang meliputi ekonomi, bisnis, intelektual yang dapat dikategorikan menjadi 6 bagian task diantaranya *Classification*, *Estimation*, *Prediction*, *Affinity grouping*, *Clustering*, *Description* dan *Profiling* (Berry, 2004).

Metode klasifikasi adalah salah satu metode yang paling sering digunakan dalam Data Mining, salah satu metode yang digunakan untuk mengklasifikasikan data yaitu metode Naive Bayes. Metode Naive Bayes merupakan proses pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class (Kusrini, 2009). Ciri utama dari Naïve Bayes Classifier adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian. Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya seperti Decision Tree dan Neural Network dengan tingkat akurasi yang lebih baik dibanding model classifier lainnya (Xhemali, Hinde, & Stone, 2009). Penggunaan metode K-Means merupakan metode interaktif yang

mudah diinterpretasikan, diterapkan, dan bersifat dinamis pada data yang tersebar (Savitri, Bachtiar, & Setiawan, 2018). Algoritma K-Means akan melakukan perulangan dalam proses pengelompokan dan berhenti dalam kondisi optimum dengan jumlah cluster dan jumlah perulangan lebih kecil dari banyaknya objek (Andayani, 2007). Input dalam algoritma K-Means dapat dikatakan bahwa hanya mengolah data kuantitatif atau dapat dikatakan hanya atribut numerik.

Beberapa peneliti sebelumnya telah melakukan penelitian dengan menggunakan metode Naive bayes dengan berbagai macam studi kasus yang digunakan. Penelitian ini telah mengambil beberapa rujukan dari penelitian sebelumnya sebagai penyempurnaan penelitian yang akan dikembangkan. Rujukan yang diambil adalah penggunaan metode Naive Bayes yang dilakukan oleh Lestaari, dkk (2020) menggunakan metode Naive Bayes dan teknik fold cross validation mendapatkan nilai sebesar 62,33 %, 66,44 %, 69,18 % dan 78,08% dalam setiap proses skenario 4 pengujian dengan tidak memperhitungkan pembagian jenis kuantitas yang sembarangan dalam proses pengelompokannya. Penelitian yang mengambil data yang serupa dengan metode lain dilakukan oleh Hardianto, dkk (2019) yang menggunakan metode Neural Network dengan menghasilkan Klasifikasi yang dihasilkan sangat baik, sehingga nasabah dengan parameter yang ada dapat diprediksi menggunakan pola ini dengan akurasi 98,21 %, dengan data yang digunakan semuanya bersifat kuantitatif, pada hakikatnya tidak semua data yang dijadikan variabel serta merta dapat dirubah dengan di notasi kan dengan variabel angka.

Penelitian tentang data pinjaman dilakukan oleh Lan, dkk (2020) dengan german data credit scoring menghasilkan nilai akurasi 0,676 atau sebesar 67,7% dengan 10 Cross Validation. Penelitian tersebut perlu adanya hubungan antar variabel yang saling berhubungan untuk meningkatkan akurasi. Hubungan antar variabel dapat dilakukan dengan proses pengelompokan dengan clustering. Penelitian yang membahas tentang risiko kredit untuk evaluasi dilakukan oleh Caruso, dkk (2020) dengan menghasilkan penelitian mendapatkan nilai error yang cukup besar dengan 3 pengujian meliputi Huang , Ahmad & D dan Cheung dan J dengan nilai berturut turut 34 %, 33% dan 43. Dalam penelitian yang dilakukan juga didapatkan perlunya teknik untuk mengelompokkan secara dinamis sehingga didapatkan jumlah kelompok yang ideal.

Penggunaan metode yang lain digunakan oleh Harlina (2018) dengan bantuan *Forward Selection* menghasilkan klasifikasi risiko penentuan kelayakan kredit menggunakan algoritma K-NN berbasis Forward Selection telah dilakukan dengan hasil akurasi 73,60% dengan menghilangkan atribut yang tidak relevan. Peneliti berpendapat dapat dikembangkan perlu penelitian lebih lanjut agar peningkatan akurasi dalam penentuan kelayakan kredit lebih meningkat. Penelitian yang dilakukan oleh Wahyuningsih dan Utari (2018) mencoba melakukan perbandingan terhadap 3 metode sehingga menghasilkan metode Decision Tree memiliki tingkat akurasi yang baik yaitu sebesar 92,21% untuk prediksi kelayakan pemberian kredit kepada nasabah, metode K-Nearest Neighbor sebesar 81,82% dan metode Naïve Bayes memiliki akurasi sebesar 81,83%. Salah

satu algoritma tersebut dapat dikembangkan dengan algoritma pengelompokan yang lain yang terdapat dalam data mining seperti K-Means.

Penelitian tentang *improved* K-means dilakukan oleh Mar'i dan Supianto (2019) untuk mengetahui metode klasifikasi yang memiliki nilai akurasi tertinggi dalam memprediksi kelayakan pemberian kredit kepada nasabah. *Improved* K-Means dengan PSO dapat memberikan model cluster yang lebih baik yaitu sebesar 0,3730 dibandingkan K-Means murni yang sebesar 0,3312.

Berdasarkan latar belakang dan penelitian terdahulu, penulis menggunakan metode naive bayes untuk proses klasifikasi penentuan kelayakan pinjaman dengan terlebih dahulu, Metode K-means digunakan untuk mengelompokkan variabel dengan sifat kuantitas ke dalam kelompok yang disediakan dengan menguji jumlah kelompok dengan Metode Elbow untuk mencari jumlah kelompok dengan nilai error terkecil. Hasil dari pengelompokan akan dihitung probabilitas dengan metode Naive Bayes dengan terlebih dahulu dibagi data dibagi ke dalam data training dan data testing menggunakan cross validation, sehingga akan didapatkan hasil nilai klasifikasi yang akan dilakukan proses pengujian menggunakan confusion matrik.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang dikemukakan maka akan dirumuskan beberapa masalah yaitu sebagai berikut :

a. Berapa jumlah distribusi kelompok yang ideal dalam pengelompokan dengan metode K-means pada variabel yang bersifat kuantitas serta jumlah anggota di setiap kelompok?

b. Berapa nilai akurasi yang dihasilkan dari proses klasifikasi dengan metode Naive bayes berdasarkan pengelompokan variabel yang dilakukan dengan algoritma K-Means dibandingkan dengan penggunaan Naive Bayes tanpa pengelompokan K-means?

1.3 Batasan Masalah

Untuk melakukan penelitian ini, maka harus ada Batasan masalah agar masalah atau penelitian tidak meluaskan dari pembahasan yang ditetapkan dan penelitian ini lebih terarah. Adapun Batasan masalah tersebut adalah:

a. Implementasi menggunakan bahasa pemrograman PHP dan basis data MySQL.
b. Data yang digunakan didapatkan dari UCI Machine Learning Dataset German Data dengan total 1000 data.
c. Variabel yang akan digunakan dalam proses klasifikasi meliputi Status of existing checking account, Duration in month, Credit history, Purpose, Credit amount, Savings account, Present employment since, Installment of disposable income, Personal status n sex, Other debtors/guarantors, Present residence since, Property, Age, Other installment plans, Housing, Existing credits at this bank, Job, Number of people being liable to provide maintenance for, Telephone, Foreign work.
d. Metode klasifikasi yang digunakan menggunakan Naive Bayes.
e. Metode pengelompokan pada variabel data yang bersifat kuantitas menggunakan K-Means *Clustering*.

f. Eksperimen pengelompokan dilakukan dari 2 kelompok sampai dengan 9 kelompok dengan metode K-Means *Clustering*

g. Proses pengujian *cluster* dalam menggunakan algoritma K-Means menggunakan metode Elbow.

h. Proses pembagian data ke dalam dataset training dan testing menggunakan K-Fold cross validation dengan fold 3.

i. Pengujian akurasi dilakukan dengan metode *confusion matrik* untuk melihat nilai akurasi pengujian yang dilakukan.

j. Keluaran yang dihasilkan berupa tingkat akurasi metode naive bayes yang didahului dengan proses pengelompokan dengan algoritma K-Means untuk variabel bersifat kuantitas.

1.4 Tujuan Penelitian

Pada bagian ini berisikan solusi yang ditawarkan untuk menangani permasalahan yang dihadapi, hal – hal yang ingin dicapai melalui kegiatan penelitian ini dan tujuan berdasarkan judul penelitian. Adapun tujuan tersebut adalah:

a. Mengetahui tingkat akurasi dari metode klasifikasi Naive Bayes dalam proses pemberian kelayakan kredit, disertai penggunaan metode K-Means dalam proses penentuan kelompok variabel terlebih dahulu.

b. Mengetahui komposisi kelompok yang ideal untuk menghasilkan nilai akurasi yang tinggi sehingga dapat dijadikan acuan model dalam analisis data kelayakan kredit nasabah.

1.5 Manfaat Penelitian

Penelitian yang baik akan memberikan ilmu dan manfaat yang dapat diterapkan di kehidupan nyata. Adapun manfaat dari penelitian tersebut adalah sebagai berikut:

a. Sebagai bahan referensi untuk menyempurnakan metode naive bayes dalam menanggulangi data yang bersifat numerik dengan solusi pengelompokan dahulu dengan metode K-Means.
b. Mengkaji besarnya nilai akurasi yang dihasilkan dari metode naive bayes dengan pengelompokan metode K-means dibandingkan dengan metode Naive Baeys tanpa metode K-means.
c. Bagi peneliti lain adalah memperkaya literatur mengenai bidang data mining khususnya dalam klasifikasi data dengan metode Naive Bayes, serta pengelompokan data dengan metode K-Means.

BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Penelitian ini telah mengambil beberapa rujukan dari penelitian sebelumnya sebagai penyempurnaan penelitian yang akan dikembangkan. Rujukan yang diambil adalah penggunaan metode Naive Bayes yang dilakukan Lestari, dkk (2020) menggunakan metode Naive Bayes dan teknik fold cross validation mendapatkan nilai sebesar 62,33 %, 66,44 %, 69,18 % dan 78,08% dalam setiap proses skenario 4 pengujian dengan tidak memperhitungkan pembagian jenis kuantitas yang sembarangan dalam proses pengelompokannya. Penelitian yang mengambil data yang serupa dengan metode lain dilakukan oleh Hardianto, dkk (2019) yang menggunakan metode Neural Network dengan menghasilkan Klasifikasi yang dihasilkan sangat baik, sehingga nasabah dengan parameter yang ada dapat diprediksi menggunakan pola ini dengan akurasi 98,21 %, dengan data yang digunakan semuanya bersifat kuantitatif.

Penelitian tentang data pinjaman dilakukan oleh Lan, dkk (2020) dengan menghasilkan german data credit scoring menghasilkan nilai akurasi 0,676 atau sebesar 67,7% dengan 10 Cross Validation. Penelitian tersebut perlu adanya hubungan antar variabel yang saling berhubungan untuk meningkatkan akurasi. Hubungan antar variabel dapat dilakukan dengan proses pengelompokan dengan clustering. Penelitian yang membahas tentang risiko kredit untuk evaluasi dilakukan oleh Caruso, dkk (2020) dengan menghasilkan nilai error yang cukup besar dengan 3 pengujian meliputi Huang , Ahmad & D dan Cheung dan J dengan

nilai berturut turut 34 %, 33% dan 43 %. Dalam penelitian yang dilakukan juga didapatkan perlunya teknik untuk mengelompokkan secara dinamis sehingga didapatkan jumlah kelompok yang ideal.

Penggunaan metode yang lain digunakan oleh Harlina (2018) dengan bantuan *forward selection* menghasilkan klasifikasi risiko penentuan kelayakan kredit menggunakan algoritma K-NN berbasis *forward selection* telah dilakukan dengan hasil akurasi 73,60% dengan menghilangkan atribut yang tidak relevan. Peneliti berpendapat dapat dikembangkan penelitian lebih lanjut agar peningkatan akurasi dalam penentuan kelayakan kredit lebih meningkat. Penelitian yang dilakukan oleh Wahyuningsih dan Utari (2018) mencoba melakukan perbandingan terhadap 3 metode. Penelitian ini menghasilkan metode Decision Tree memiliki tingkat akurasi yang baik yaitu sebesar 92,21% untuk prediksi kelayakan pemberian kredit kepada nasabah, metode K-Nearest Neighbor sebesar 81,82% dan metode Naïve Bayes memiliki akurasi sebesar 81,83%. Salah satu algoritma tersebut dapat dikembangkan dengan algoritma pengelompokan yang lain yang terdapat dalam data mining seperti K-Means.

Penelitian tentang improved K-means dilakukan oleh Mar'i dan Supianto (2019) untuk mengetahui metode klasifikasi yang memiliki nilai akurasi tertinggi dalam memprediksi kelayakan pemberian kredit kepada nasabah. Improved K-Means dengan PSO dapat memberikan model cluster yang lebih baik yaitu sebesar 0,3730 dibandingkan K-Means murni yang sebesar 0,3312.

2.2 Keaslian Penelitian

Tabel 2. 1 Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network	Nur Hadianto, Hafifah Bella Novitasari, Ami Rahmawati, Jurnal PILAR Nusa Mandiri Vol. 15, No. 2, 2019	Penulis menggunakan metode Neural Network menentukan klasifikasi pinjaman bagi nasabah yang memiliki peluang tinggi untuk dijadikan nasabah peminjam	Klasifikasi yang dihasilkan sangat baik, sehingga nasabah dengan parameter yang ada dapat diprediksi menggunakan pola ini dengan akurasi 98,21 % menggunakan struktur 12-15-8-1.	Data yang digunakan semuanya bersifat kuantitatif, pada hakikatnya tidak semua data yang dijadikan variabel serta merta dapat dirubah dengan dinotasikan dengan variabel angka.	Penelitian yang dilakukan Nurhadianti ,dkk menggunakan Neural Network dan di analisis dengan Rapidminer sebagai tools uji coba, sedangkan penelitian yang akan dilakukan menggunakn Naive Bayes dengan pengelompokan variabel terlebih dahulu menggunakan metode K-Means dan di implementasi kedalam bentuk program.
2	Implementasi Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Anugerah Bintang Cemerlang	Siti Lestari, Akmaludin, Mohammad Badrul, Jurnal PROSISKO Vol. 7 No. 1, 2020	Penulis menerapkan metode data mining untuk mengklasifikasikan kelayakan nasabah dalam kategori layak dan tidak layak berdasarkan data historis nasabah di masa sebelumnya	Hali akurasi yang diperoleh dengan 4 kali proses pengujian menggunakan teknik k-fold cross validation mendapatkan nilai sebesar 62.33 %, 66.44 %, 69.18 % dan 78.08%.	Penggunaan probabilitas pada variabel awal yang berjenis numerik dimasukan kedalam kategori yang sembarangan, tanpa ada dasar untuk menetapkan jumlah kategori yang digunakan.	Penelitian sri lestari dkk menggunakan pengelompokan yang bersifat pengaktegorian sembaragngan. Penelitian yang akan dilakukan menggunakan algoritma K-Means dalam proses pembagian data kategori kelompok dengan pengujian jumlah kategori menggunakan metode Elbow.

Tabel 2.1 Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Data Mining Pada Penentuan Kelayakan Kredit Menggunakan Algoritma K-NN Berbasis Forward Selection	Sitti Harlina, Creative Communication and Innovative Technology (CCIT) ISSN: 1978 -8282, 2019	Menerapkan algoritma K-nn yang diterapkan pada data konsumen yang menggunakan jasa keuangan kredit dengan bantuan Forward Selection.	Klasifikasi risiko penentuan kelayakan kredit menggunakan algoritma K-NN berbasis Forward Selection telah dilakukan dengan hasil akurasi 73,60% dengan menghilangkan atribut yang tidak relevan. Penggunaan fitur Forward Selection dalam pemrosesan data akan mempengaruhi hasil pencapaian akurasi yang didapatkan	Masih perlu penelitian lebih lanjut agar peningkatan akurasi dalam penentuan kelayakan kredit lebih meningkat	Penelitian siti harlina menggunakan seleksi forward selection untuk menghilangkan atribut yang tidak relevan dan menggunakan metode K-NN. Penelitian yang akan dilakukan menggunakan metode K-Mean untuk pengelompokan variabel yang bersifat numerik dan proses klasifikasi menggunakan algoritma Naive bayes
4	Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit	Sri Wahyuningsih, Dyah Retno Utari, Konferensi Nasional Sistem Informasi, 2018	Untuk mengetahui metode klasifikasi yang memiliki nilai akurasi tertinggi dalam memprediksi kelayakan pemberian kredit kepada nasabah	Metode Decision Tree memiliki tingkat akurasi yang baik yaitu sebesar 92,21% untuk prediksi kelayakan pemberian kredit kepada nasabah, metode K-Nearest Neighbor sebesar 81,82% dan metode Naïve Bayes memiliki akurasi sebesar 81,83%.	Mengkombinasikan lebih banyak metode dalam Analisa data dan penyelesaian masalah. Dapat dikembangkan dengan algoritma klasifikasi yang lain yang terdapat dalam data mining seperti K-Means.	Pnelitiansriwahyuningsih dkk membandingkan 3 metode ununtuk dicarkan nilai akurasi yang terbaik tanpa adanya pengelompokan terlebih dahulu. Penelitian yang akan dilakukan menggunakan algoritma K-Means untuk pengelompokan variabel terlebih dahulu.

Tabel 2.1 Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	Clustering Credit Card Holder Berdasarkan Pembayaran Tagihan Menggunakan Improved K-Means Dengan Particle Swarm Optimization	Farhanna Mar'i1 , Ahmad Afif Supianto, Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 2018	Menentukan pengelompokan nasabah dengan kelompok yang layak diberikan kredit dan kelompok yang tidak layak diberikan kredit.	*Improved K-Means* dengan PSO dapat memberikan model *cluster* yang lebih baik yaitu sebesar 0,3730 dibandingkan K-Means murni yang sebesar 0,3312	Meningkatkan hasil yang lebih baik dapat dalam proses penentuan jumlah cluster yang ada pada setiap variabel dalam setiap iterasinya sehingga menghasilkan fungsi yang lebih baik.	Penelitian ayng dilakukan farhana dan ahmad menggunkan Improveed K-Means untuk menghasilkan akurasi dengan algoritma PSO. Penelitian yang akan dilakukan dengan mengambil algoritma K-Means dalam proses pengelompokan variabel probabilitas dan melakukan eksperimen dengan pengelompkkan berbagai macam variabel yang sejenis.
6	Multivariable data imputation for the analysis of incomplete credit data.	Qiujun Lan, Xuqing xu, Hoijie Ma, Gang Li, Expert Systems With Applications, 2020	Penelitian ditujukan untuk proses analisis mengisi data probabilitas yang masih belum terisi dengan pendekatan probabilitas.	Hasil penelitian dengan data german data credit scoring menghasilkan nilai akurasi 0,676 atau akurasi sebesar 67,7% dengan 10 Cross Validation.	Pengujian yang dilakukan perlu adanya hubungan antar variabel yang saling berhubungan untuk meningkatkan akurasi.	Penelitian yang dilakukan Qiuju dkk menganalisis untuk pemberian data probabilitas dengan distribusi normal dan dilalukan pengujian terhadap hasil yang baru. Penelitian yang akan dikerjakan mengelompkna variabel probabilitas dengan jenis yang serupa dengan algoritma K-Means terlebih dahulu.

Tabel 2.1 Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
7	Cluster Analysis for mixed data: An application to credit risk evaluation ation	G. Caruso; S.A. Gattone; F. Fortuna; T. Di Battista, Socio-Economic Planning Sciences, 2020	Untuk Melakukan uji coba dengan indikator kualitatif dengan indikator kuantitatif pada data denga cluster analisis..	Dari hasil penelitian yang dilakukan didapatkan nilai error yang cukup besar dengan 3 pengujian meliputi Huang , Ahmad & D dan Cheung dan J dengan nilai berturut turut 34 %, 33% dan 43 % .	Perlunya teknik untuk mengelompokkan secara dinamis sehingga didapatkan jumlah kelompok yang ideal..	Penelitian batista dkk melakukan penelitian untuk mencoba menggabungkan indikator kualitatif dan kuantitatif. Penelitian yang akan dilakukan mengambil dataset nasabah untuk prediksi kelayakan kredit, menggunakan algoritma K-Means untuk proses pengelompokan variabel awal dengan jumlah cluster yang dihitung terlebih dahulu dengan metode Elbow. Setelah itu dilanjutkan untuk pengujian untuk klasifikasi dengan metode Naive bayes
8	A Naive Bayes approach to fraud prediction in loan default	I O Eweoya;A A Adebiyi; A A Azeta; F Chidozie;, F O Agono; B Guembe, International Conference on Science and Sustainable Development, 2019	Penggunaan machine learning dengan naive bayes untuk menguji dataset dengan 9 variabel probabiltas.	Hasil pengujian nilai 78% dengan cross validation sebesar 25 % untuk testing.	Data yang digunakan semua bersifat kualitatif sehingga tidak perlu dimodelkan untuk transformasi terlebih dahulu.	Penelitian Eweoya dkk, melakukan pengujian dataset dengan metode naive bayes tanpa perlu proses tranformasi terlebih dahulu. Penelitian yang akan dilakukan mencoba melakukan transformasi data dengan K-Means dan menggabungkan variabel yang sejenis selanjutnya dilakukan proses klasifikasi dengan naive bayes.

2.3 Landasan Teori

2.3.1 Kredit

Kemampuan untuk melaksanakan suatu pembelian atau mengadakan suatu pinjaman dengan suatu janji, pembayaran akan dilaksanakan pada jangka waktu yang telah disepakati (Astiko, 1996). Pengertian kredit yang lebih mapan untuk kegiatan perbankan di Indonesia telah dirumuskan dalam Undang – Undang Pokok Perbankan No. 7 Tahun 1992 yang menyatakan bahwa kriteria adalah penyediaan uang/tagihan yang dapat dipersamakan dengan itu berdasarkan persetujuan/kesepakatan pinjam meminjam antara pihak bank dengan pihak lain yang mewajibkan pihak peminjam untuk melaksanakan dengan jumlah bunga sebagai imbalan. Dalam praktek sehari – hari pinjaman kredit dinyatakan dalam bentuk perjanjian tertulis baik di bawah tangan maupun secara materiil. Dan sebagai jaminan pengaman, pihak peminjam akan memenuhi kewajiban dan menyerahkan jaminan baik bersifat kebendaan maupun bukan kebendaan.

2.3.2 Data Mining

Data mining merupakan kegiatan penemuan pola-pola yang menarik dari data berukuran besar yang disimpan dalam basis data, data warehouse, atau sarana penyimpanan yang lain. Data mining dapat diklasifikasikan menjadi dua kategori: descriptive data mining dan predictive data mining (Hermadi, 2007).Data mining sering disebut sebagai *Knowledge Discovery in Database* (KDD) yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu algoritma yang spesifik (Wirdasari & Calam, 2011). Dari beberapa pengertian tersebut dapat ditarik kesimpulan data mining merupakan suatu teknik menggali

informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (database) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui. Data mining itu sendiri merupakan usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar oleh karena itu data mining memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), machine learning, statistik dan database (Wirdasari & Calam, 2011). Tahapan untuk melakukan data mining seperti berikut:

a. Pembersihan data data (*data cleaning*) merupakan proses menghilangkan noise dan data yang tidak konsisten atau yang tidak relevan.
b. Integrasi data (*data integration*) merupakan penggabungan data dari berbagai database ke dalam satu database baru.
c. Seleksi data (*data selection*) merupakan data yang tidak sesuai atau yang tidak perlu dianalisis tidak diambil sedangkan data yang sesuai untuk dianalisis yang akan diambil di database.
d. Transformasi data (*data transformation*) merupakan data yang akan diubah atau digabungkan ke dalam format yang sesuai untuk diproses dalam data mining.
e. Proses mining merupakan suatu proses utama saat metode diterapkan untuk menentukan pengetahuan berharga dan tersembunyi dari data.
f. Evaluasi pola (*pattern evaluation*) merupakan pengidentifikasian pola-pola menarik ke dalam knowledge based yang ditemukan.

g. Presentasi pengetahuan (*knowledge presentation*) merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.3.3 Metode Naïve Bayes

Metode Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decesion tree dan neural network. Bayesian Classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar (Kusrini, 2009).

Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Persamaan dari teorema *Bayes* adalah :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (2.1)$$

Keterangan :

X	: Data dengan class yang belum diketahui
H	: Hipotesis data *X* merupakan suatu class spesifik
P(H|X)	: Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori
P (H)	: Probabilitas hipotesis *H* (prior probability)
P(X|H)	: Probabilitas *X* berdasarkan kondisi pada hipotesis *H*
P(X)	: Probabilitas *X*

Untuk menjelaskan teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut.

2.3.4 K-Means

K-Means merupakan algoritma yang sederhana dan cukup efektif dalam proses pengelompok kan data berdasarkan kemiripannya atau disebut *clustering*. Penentuan kemiripan data dihitung menggunakan rumus *Euclidean Distance* (Anggodo, Cahyaningrum, & Fauziyah, 2017). K-Means termasuk dalam salah satu algoritma yang bersifat *unsupervised* yaitu sebuah algoritma yang tidak membutuhkan proses *training* atau pembelajaran terlebih dahulu. Setiap *cluster* dibentuk berdasarkan titik *centroid* atau titik pusat yang ditentukan. Jumlah titik *centroid* pada *cluster* menggambarkan jumlah K atau *cluster* itu sendiri.

Pada algoritma *K-Means* penentuan awal titik *centroid* didapatkan secara random, dan pada iterasi berikutnya titik *centroid* didapatkan dari perhitungan jarak rata-rata dari anggota *cluster* terhadap titik *centroid* awal. Adapun algoritma *K-Means* adalah sebagai berikut:

1. Tentukan Nilai K terlebih dahulu. Nilai K adalah jumlah *cluster* yang ingin dibuat.
2. *Generate* titik *centroid* awal secara *random*, apabila K=2 maka diperlukan 2 titik *centroid* awal yang didapatkan secara acak dari dataset.
3. Hitung jarak antara data x_i dengan titik *centroid* menggunakan rumus *Euclidean Distance* sebagai berikut.

$$D((X_i, C_j) = \sqrt{\sum_{i=1}^{p}(X_i - C_j)^2} \quad (2.2)$$

Keterangan

X_i = data ke-i

C_j = titik pusat *cluster* atau *(centroid)*

P = dimensi data

4. Menentukan setiap objek data masuk kedalam *cluster* dengan jarak *euclidean* terdekat atau terkecil.
5. Melakukan perhitungan ulang titik centroid pada iterasi berikutnya dengan menghitung rata-rata jarak antara semua objek dalam cluster.
6. Ulangi langkah ke-3 hingga nilai *centroid* tidak berubah atau telah dalam kondisi berhenti *(stopping condition)*.

2.3.5 Metode Elbow

Metode Elbow merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik.

Berikut ini tahapan algoritma metode Elbow dalam menentukan nilai k pada K-Means (Rahman, Wiranto, & Anggrainingsih, 2017)

1. Menginisialisasi awal nilai k.
2. Menaikan nilai k.
3. Menghitung hasil sum of square error dari tiap nilai k.
4. Analisis hasil sum of square error dari nilai k yang mengalami penurunan secara drastis.
5. Cari dan tentukan nilai k yang berbentuk siku.

Pada metode Elbow nilai cluster terbaik yang akan diambil dari nilai Sum of Square Error (SSE) yang mengalami penurunan yang signifikan dan berbentuk

siku.Untuk menghitung SSE menggunakan rumus sebagai berikut (Irwanto, 2012).

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_k} \|X_i - C_j\|^2 \quad (2.3)$$

Sum of Square Error (SSE) merupakan rumus yang digunakan untuk mengukur perbedaan antara data yang diperoleh dengan model perkiraan yang telah dilakukan sebelumnya. SSE sering digunakan sebagai acuan penelitian terkait dalam menentukan optimal cluster

2.3.6 Cross Validation

Cross validation adalah prosedur untuk memperkirakan generalisasi kinerja dalam sebuah metode pemodelan. Cross-validation adalah metode yang paling sering digunakan untuk evaluasi kinerja prediktif dari sebuah model, yakni model yang diberikan sebelumnya atau model yang telah dikembangkan oleh prosedur pemodelan (Yadav & Shukla, 2016). Data biasanya akan dibagi menjadi dua bagian pada bagian pertama dilakukan pelatihan sementara pada bagian lainnya dilakukan uji kinerja, skema pelatihan dan pengujian bekerja dengan baik pada model klasifikasi di dalam machine learning, beberapa record dalam dataset dijadikan data training untuk dilatih sementara record lainya di dalam dataset digunakan sebagai data testing, hal tersebut adalah prinsip dasar dari cross validation, karena hal tersebutlah cross-validation sangat diterima dalam komunitas data mining dan machine learning dan berfungsi sebagai prosedur standar untuk pemilihan model atau pemilihan prosedur pemodelan (Yadav & Shukla, 2016).

BAB III

METODE PENELITIAN

3.1 Jenis, Sifat dan Pendekatan Penelitian

Penelitian ini menggunakan jenis penelitian eksperimental, karena dilakukan dengan serangkaian tindakan uji coba untuk menghasilkan akurasi yang terbaik. Tingkat akurasi akan dihitung dengan metode *confusion matrik*. Sifat penelitian ini dilakukan secara mandiri dengan menggunakan metode deskriptif dan dilakukan pendekatan kuantitatif dalam perhitungan tingkat akurasi klasifikasi data nasabah yang sudah melakukan peminjaman.

Objek yang diteliti yaitu dataset yang sudah disediakan dalam proses pemberian pinjaman dengan adanya pembagian status yang lancar dan yang tidak lancar. Proses pengujian akan dilakukan dengan pembagian kedalam data training yang digunakan untuk pembelajaran dan dilakukan pengujian dengan data testing. Pembagian jenis data ini dilakukan dengan cross validation dengan menguji coba data kedalam nilai yang dirubah untuk mencari nilai akurasi yang terbaik.

3.2 Metode Pengumpulan Data

Metode yang digunakan dalam penelitian ini menggunakan studi dokumen. Studi dokumen dilakukan dengan mengkaji dokumen-dokumen terkait penelitian dalam hal ini merupakan dataset UCI Machine Learning german data. Data yang digunakan merupakan data sekunder. Data yang didapatkan merupakan data yang bersifat publik karena disediakan dan siapapun dapat mengunduh data yang dibutuhkan. Pada penelitian ini dilakukan proses klasifikasi data terhadap data – data history nasabah yang sudah melakukan peminjaman dengan dibagi

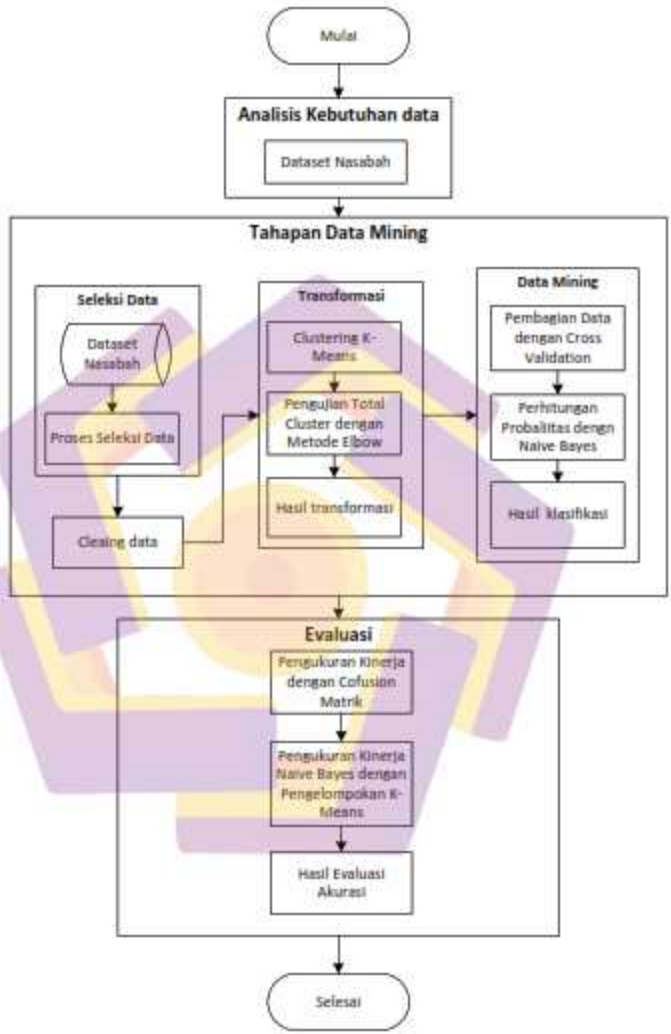
menjadi data training dan data testing. Kegiatan ini bertujuan untuk mengevaluasi dan mencari nilai akurasi terhadap proses yang dilakukan. Data yang dibutuhkan adalah data history pinjaman dari nasabah dengan sudah diketahui hasil macet atau lancar proses kredit yang sudah berjalan. Data diambil dari UCI Machine learning database german data yang meliput 1000 record data.

3.3 Metode Analisis Data

Pada bagian analisis data akan dilakukan suatu proses analisis data yang terdiri dataset berjumlah 1000 data yang terlebih dahulu dilakukan proses penyesuaian format dalam excel untuk dapat diolah kedalam aplikasi yang akan dibangun. Analisis yang dilakukan dalam penelitian dengan dataset tersebut adalah sebagai berikut :

1. Untuk mengetahui jumlah kelompok yang terbaik dalam dataset dengan variabel bersifat numerik menggunakan algoritma K-Means terlebih dahulu, hasil dari beberapa uji coba akan di tes akurasi jumlah kelompok yang terbaik dilihat dari selisih error dari jarak centroid ke data yang ada dengan metode Elbow.
2. Untuk mengetahui hasil akurasi dalam hasil penelitian ini dilakukan perhitungan error dengan confusion. Pembagian ke 2 jenis data ini menggunakan metode Cross Validation dengan skenario uji coba dengan memasukan nilai yang mendekati hasil terbaik.
3. Untuk mengetahui tingkat keberhasilan yang tinggi dilihat dari nilai akurasi dengan cara pengelompokan variabel dengan algoritma K-Means dibandingkan dengan perhitungan dengan metode Naive Bayes saja.

3.4 Alur Penelitian



Gambar 3. 1 Alur Penelitian

Pada Gambar 3.1 menjelaskan 4 tahapan alur penelitian yang akan dilakukan yang dapat dijelaskan sebagai berikut:

a. Analisis kebutuhan data

Dalam tahapan analisis kebutuhan data mengambil data dari dataset yang di dapat dari UCI Machine Learning dengan total jumlah data set 1000 data. Dalam data yang didapatkan akan dikategorikan ke dalam 2 klasifikasi yaitu yang layak atau dengan kata lain kredit lancar dan yang tidak layak atau kredit macet.

B. Tahapan data mining

Dalam tahapan data mining akan dibagi menjadi 4 tahapan secara garis besar dengan penjabaran sebagai berikut:

1. Seleksi data

   Dalam tahapan seleksi data data didapatkan dari dataset nasabah untuk dipilih variabel yang akan digunakan dalam probabilitas. Dalam penelitian ini akan diambil 12 variabel probabilitas dengan dilengkapi 1 tabel keputusan.

2. Cleaning data

   Proses cleaning data berfungsi untuk proses pembersihan data yang dapat mempengaruhi kinerja dari data mining. Cleaning data akan menghilangkan data yang tidak dapat dikenali seperti data yang bersifat kosong.

3. Transformasi

Tahapan transformasi data dilakukan dengan proses penentuan kelompok dengan algoritma K-means. Dalam proses transformasi akan dilakukan skenario pengujian data dimulai dari 2 sampai dengan 10 kelompok menyesuaikan rentang data yang ada. Proses pengujian jumlah kelompok yang terbaik dalam setiap variabel akan menggunakan metode elbow sehingga didapat jumlah kelompok dengan kualitas yang terbaik. Hasil proses transformasi selanjutnya akan dilakukan dengan proses data mining.

4. Data mining

Dalam tahapan data mining akan dilakukan proses pembagian data ke dalam 2 jenis dengan data training dan data testing. Skenario pengujian alokasi data akan dilakukan dengan *Crosss Validation* untuk mencari nilai akurasi yang terbaik. Proses naive bayes dilakukan perhitungan kedua jenis data sehingga dihasilkan informasi klasifikasi pada data testing.

c. Evaluasi

evaluasi dilakukan dengan *Confussion Matrik* dengan melakukan perhitungan kinerja naive bayes yang terlebih dahulu menggunakan pengelompokan variabel dengan K-Means sehingga didapatkan hasil akurasi setiap skenario pengujian. Hasil pengujian akan dibandingkan dengan kelas awal dan dibandingkan dengan target kelas.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1 Analisis Kebutuhan Data

Tahapan analisis kebutuhan data merupakan tahapan awal ketika penelitian akan dilakukan. Dalam tahapan ini data didapat dari german credit data dengan jumlah record 1000 data, dengan total 20 variabel yang digunakan dan 1 variabel tujuan dengan 2 kelas bernilai good dan bad.

4.1.1 Variabel Data

Variabel data yang ada dalam german credit data berjumlah totoal 20 dengan keterangan yang dapat dilihat dalam Tabel 4.1

Tabel 4. 1 Variabel Data

No	Atribut	Tipe
1	Status of existing checking account	Kategori
2	Duration in month	Numerik
3	Credit history	Kategori
4	Purpose	Kategori
5	Credit amount	Numerik
6	Savings account	Kategori
7	Present employment since	Kategori
8	Installment of disposable income	Numerik
9	Personal status n sex	Kategori
10	Other debtors/guarantors	Kategori
11	Present residence since	Numerik
12	Property	Kategori
13	Age	Numerik
14	Other installment plans	Kategori
15	Housing	Kategori

Tabel 4.1 Variabel Data (Lanjutan)

No	Atribut	Tipe
16	Existing credits at this bank	Numerik
17	Job	Kategori
18	Number of people being liable to provide maintenance for	Numerik
19	Telephone	Kategori
20	Foreign work	Kategori

4.1.2 Dataset

Dataset yang digunakan merupakan data yang terdiri dari 1000 record dengan 20 variabel dengan 7 bersifat numerik dan 13 kategori, dan 1 variabel target. Dalam data set terbagi kedalam 2 kelompok variabel target dengan 300 dengan keterangan bad credit dan 700 good credit. Tabel sample dataset dapat dilihat dalam Tabel 4.2

Tabel 4. 2 Sample Dataset

No	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	Kelas
1	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
2	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	2
3	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
4	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1
5	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	2
6	A14	36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A912	A201	1
7	A14	24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A201	1
8	A12	36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A201	1
9	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	1
10	A13	30	A34	A40	5234	A61	A71	4	A94	A101	2	A123	28	A143	A152	2	A174	1	A191	A201	2
11	A11	12	A32	A40	1295	A61	A72	3	A92	A101	1	A123	25	A143	A151	1	A173	1	A191	A201	2
12	A12	48	A32	A49	4308	A61	A72	3	A92	A101	4	A122	24	A143	A151	1	A173	1	A191	A201	2
13	A14	12	A32	A43	1567	A61	A73	1	A92	A101	1	A123	22	A143	A152	1	A173	1	A192	A201	1
14	A14	24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	2
15	A12	15	A32	A40	1403	A61	A73	2	A92	A101	4	A123	28	A143	A151	1	A173	1	A191	A201	1
16	A11	24	A32	A43	1282	A62	A73	4	A92	A101	2	A123	32	A143	A152	1	A172	1	A191	A201	2
17	A12	24	A34	A43	2424	A65	A75	4	A93	A101	4	A122	53	A143	A152	2	A173	1	A191	A201	1
18	A12	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1

4.1.3 Informasi Label Variabel

Dalam penelitian ini menggunakan 20 variabel yang digunakan dalam perhitungan dengan 1 kelas tujuan. Informasi mengenai informasi variabel dapat dilihat dalam Tabel 4.3

Label V1 merupakan data dari *Status of existing checking accoun*t mempunyai 4 Label yang terdiri dari A11, A12, A13 dan A14 yang dapat dilihat informasi keterangannya pada Tabel 4.3

Tabel 4. 3Status of existing checking account

Label	Keterangan
A11	< 0 DM
A12	0 <= ... < 200 DM
A13	... >= 200 DM / salary assignments for at least 1 year
A14	no checking account

Label V3 merupakan data dari *Credit history* yang mempunyai 5 label terdiri dari A30, A31, A32, A33 dan A34 yang dapat dilihat informasi keterangannya pada Tabel 4.4

Tabel 4. 4 Credit history

Label	Keterangan
A30	no credits taken/ all credits paid back duly
A31	all credits at this bank paid back duly
A32	existing credits paid back duly till now
A33	delay in paying off in the past
A34	critical account/ other credits existing (not at this bank)

Label V4 merupakan data dari variabel *purpose* mempunyai 11 label terdiri dari A40, A41, A42, A43, A44, A45, A46, A47, A48, A49 dan A410 yang dapat dilihat informasi keterangannya pada Tabel 4.5

Tabel 4. 5Purpose

Label	Keterangan
A40	Car (new)
A41	Car (used)
A42	Furniture/Equipment
A43	Radio/Television
A44	Domestic Applicances
A45	Repairs
A46	Education
A47	(Vacation – Does not exist?)
A48	Retrainning
A49	Business
A410	Others

Label V6 merupakan keterangan dari *Savings account/bonds* dengan mempunyai 5 label yang terdiri dari A61, A62, A63, A64 dan A65 yang dapat dilihat informasi keterangannya pada Tabel 4.6

Tabel 4. 6 Savings account/bonds

Label	Keterangan
A61	... < 100 DM
A62	100 <= ... < 500 DM
A63	500 <= ... < 1000 DM
A64	.. >= 1000 DM
A65	unknown/ no savings account

Label V7 merupakan keterangan dari *employment since* yang mempunyai 5 label yang terdiri dari A71, A72, A73, A74 dan A75 yang dapat dilihat informasi keterangannya pada tabel 4.7

Tabel 4. 7 Present employment since

Label	Keterangan
A71	unemployed
A72	... < 1 year
A73	1 <= ... < 4 years
A74	4 <= ... < 7 years
A75	.. >= 7 years

Label V9 merupakan keterangan dari *status and sex since* dengan 5 label yang terdiri dari A91, A92, A93, A94 dan A95 yang dapat dilihat informasi keterangannya pada Tabel 4.8

Tabel 4. 8Personal status and sex since

Label	Keterangan
A91	male: divorced/separated
A92	female: divorced/separated/married
A93	male: single
A94	male: married/widowed
A95	female: single

Label V10 merupakan keterangan dari *Other debtors/guarantors* Mempunyai 3 label yang terdiri dari, A101, A102 dan A103 yang dapat dilihat informasi keterangannya pada Tabel 4.9

Tabel 4. 9 Other debtors / guarantors

Label	Keterangan
A101	None
A102	Co-applicant
A103	Guarantor

Label V12 merupakan keterangan dari *property* Mempunyai 4 label yang terdiri dari, A121, A122, A123 dan A124 yang dapat dilihat informasi keterangannya pada Tabel 4.10

Tabel 4. 10 Property

Label	Keterangan
A121	real estate
A122	if not A121: building society savings agreement/ life insurance
A123	if not A121/A122: car or other, not in attribute 6
A124	unknown / no property

Label V14 merupakan keterangan dari *Other installment plans* dengan 3 label yang terdiri dari, A141, A142 dan A143 yang dapat dilihat informasi keterangannya pada Tabel 4.11

Tabel 4. 11 Other installment plans

Label	Keterangan
A141	Bank
A142	Stores
A143	None

Label V15 merupakan keterangan dari *Housing* dengan 3 label yang terdiri dari, A151, A152 dan A153 yang dapat dilihat informasinya keterangannya pada Tabel 4.12

Tabel 4. 12 Housing

Label	Keterangan
A151	Rent
A152	Own
A153	For free

Label V17 merupakan keterangan dari Job dengan 4 label yang terdiri dari A171, A172, A173 dan A174 yang dapat dilihat informasi keterangannya pada Tabel 4.13

Tabel 4. 13 Job

Label	Keterangan
A171	unemployed/ unskilled - non-resident
A172	unskilled - resident
A173	skilled employee / official
A174	management/ self-employed/ highly qualified employee/ officer

Label V19 merupakan keterangan dari Telephone dengan 2 label yang terdiri dari A191 dan A192 yang dapat dilihat informasinya keterangannya pada Tabel 4.14

Tabel 4. 14 Telephone

Label	Keterangan
A191	None
A192	yes, registered under the customers name

Label V20 merupakan keterangan dari Foreign Worker Mempunyai 2 label yang terdiri dari A201dan A202yang dapat dilihat informasinya keterangannya pada Tabel 4.15

Tabel 4. 15 Foreign Worker

Label	Keterangan
A201	yes
A202	no

4.2 Seleksi Data

Tahapan seleksi data bermaksud untuk memilih variabel data yang dilakukan dalam tahap transformasi. Seleksi data diambil dari variabel yang bersifat numerik dengan mempertimbangkan jarak data terkecil dengan terbesar kemungkinan yang ada. Data numerik yang dapat dilakukan dengan seleksi data dapat dilihat dalam Tabel 4.16

Tabel 4. 16 Variabel data Numerik.

No	Variabel	Max	Min	Variasi
1	Duration in month	72	4	69
2	Credit amount	18424	250	18175
3	Installment rate in percentage of disposable income	4	1	4

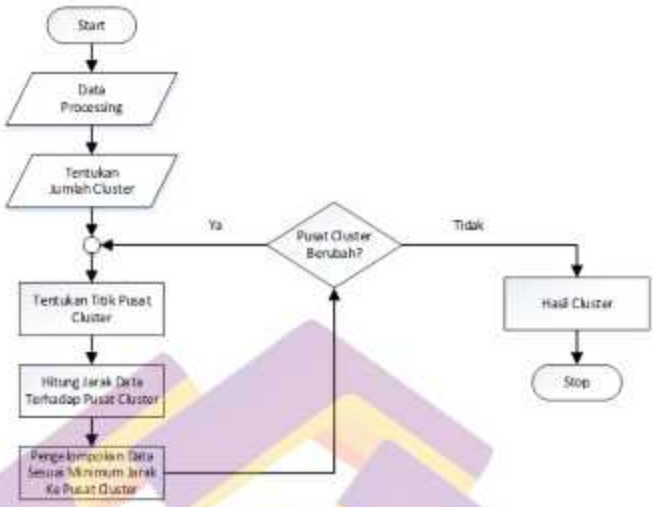
Tabel 4.16 Variabrl data Numerik (Lanjutan)

No	Variabel	Max	Min	Variasi
4	Present residence since	4	1	4
5	Age in years	75	19	57
6	Number of existing credits at this bank	2	1	2
7	Number of people being liable to provide maintenance for	2	1	2

Dari Tabel 4.3 dapat dilihat data yang mempunyai variasi data paling signifikan terdapat dalam 3 variabel yaitu duration in month, credit amount dan age in years, maka proses transformasi data dapat dilakukan dengan mengambil 3 variabel itu.

4.3 Transformasi data

Tahapan transformasi data berfungsi untuk merubah data yang bersifat numerik ke dalam data bersifat kategori sehingga bisa dihitung probabilitas dengan algoritma Naive Bayes. Dalam tahapan transformasi data menggunakan algoritma K-Means dengan dilakukan proses uji coba dari cluster 2 sampai dengan cluster 9 untuk dicari pengelompokan yang paling baik dengan metode Elbow. Flowchart algoritma K-Means dapat dilihat dalam gambar 4.1



Gambar 4. 1 Flowchart Algoritma K-means

Dalam gambar 4.1 tentang flowchart Algoritma K-Means dapat dijelaskan dengan dimulai dari start kemudian dilakukan dengan data pemrosesan yaitu dilakukan prosesing data jika diperlukan. Tahapan selanjutnya menentukan jumlah cluster atau kelompok. Setelah ditentukan jumlah kelompok maka dicari titik pusat cluster/kelompok yang bisa disebut dengan nama centroid. Tahapan selanjutnya akan dicari jarak data terhadap titik cluster atau centroid, untuk dicari titik yang paling kecil dan dimasukkan ke dalam kelompok yang mempunyai selisih paling kecil. Proses selanjutnya dihitung rata - rata titik pusat cluster setiap kelompok apakah berubah dari titik pada percobaan perhitungan sebelumnya atau tidak, jika berubah maka akan dilakukan perhitungan lagi, jika tidak berubah maka ditemukan hasil cluster setiap data yang di kelompokan.

4.3.1 Transformasi Variabel Credit in Month

Variabel credit in month mempunyai data yang cukup bervariasi dengan data paling kecil bernilai 250 dan data yang paling besar bernilai 18.424. Tahapan pengelompokan akan dilakukan dengan percobaan 2 sampai dengan 9 clutser untuk dicari distribusi data dan dicari yang paling optimal pengelompokannya

Membangkitkan *centroid* awal. *Centroid* awal dapat diperoleh secara acak, atau dengan rumus tertentu dengan jumlah centroid sebanyak cluster yang akan dibuat. *Centroid* awal merupakan titik pusat *cluster* pertama atau awal pusat *cluster*. Pembentukan centroid awal yang dilakukan dengan random akan memicu inkosistensi data, maka dengan itu penelitian kali ini proses perhitungan centroi dilakukan dengan rumus. Controid awal dapat dihitung dengan persamaan sebagai berikut (Kusrini, 2015).

$$c\text{-i} = \min + \frac{(i-1)*(\max-\min)}{n} + \frac{(\max-\min)}{2*n} \qquad (4.1)$$

Dalam kasus ini kita coba menghitungan Centroid 1 pada data credit in month dengan pengelompokan sejumlah 2 cluster.

$$C\,[credit][1] = 1 + \frac{(1-1)*(3-1)}{3} + \frac{(3-1)}{2*1}$$

$$C\,[credit][1] = 4793.5$$

Implementasi kode program dapat dilihat dalam potongan kode program yang ditunjukan pada Listing 4.1

```
1.  <?php
2.  $variabel='credit';
3.  $cluster=array();
4.  $min=caridata($mysqli,"select min($variabel) from
    tb_transformasi");
5.  $max=caridata($mysqli,"select max($variabel) from
    tb_transformasi");
6.  for ($i=1;$i<=1000;$i++) {
7.    $clusterawal[$i]="C0";}
```

```
8.  for ($i=1;$i<=$jumlahcluster; $i++) {
9.     $cluster[$i]=$min+((($i-1)*($max-
    $min))/($jumlahcluster))+(($max-$min)/(2*$jumlahcluster));}
10. echo "Variabel : ".$variabel;
11. echo "<br>Jumlah Cluster : ".$jumlahcluster;
12. echo "<br>Nilai Min : ".$min;
13. echo "<br>Nilai Min : ".$max;
14. echo "Jumlah Akurasi $i = ".$row[0];
15. ?>
```

Listing 4.1 Kode Program penentuan centroid awal

Pada Listing program 4.1 dapat dijelaskan kode no 1 dapat dijelaskan kode no 2 untuk mengambil data variabel credit. kode no 4 untuk mencari nilai terkecil. Kode no 5 untuk mencari nilai terbesar dari variabel credit. Kode no 8 sampai dengan kode no 9 digunakan untuk mencari nilai centroid pada titik yang akan dicari. Kode no 10 sampai dengan kode no 14 untuk menampilkan informasi data.

Dari hasil perhitungan dengan formula d iatas maka ditemukan nilai centroid pada 2 cluster yang dapat dilihat dalam Tabel 4.17

Tabel 4. 17 Nilai Centroid Awal variabel credit dengan 2 Cluster

Titik Centroid	Centroid
1	4.793,5
2	13.880,5

Dari tabel 4.4 dapat dilihat bahwa nilai centroid awal pada titik pusat 1 sebesar 4793,5 dan titik centroid 2 sebesar 13880,5. Setelah ditemukan titik centroid maka Menghitung distance space data ke masing-masing centroid pada variabel credit in month terhadap masing - masing centroid. Pada perhitungan akan dicontohkan data dengan nilai credit in month sebesar 1169.

$$C1 = \sqrt{(4793{,}5 - 1169)^2}$$

$$C1 = \sqrt{13137000{,}25} = 3624{,}5$$

$$C2 = \sqrt{(13880{,}5 - 1169)}$$

$$C2 = \sqrt{161582232{,}3} = 12311{,}5$$

Dari hasil perhitungan di atas data diketahui nilai selisih dari data yang dihitung C1 sebesar 3624,5 dan C2 12311,5 . Berdasarkan perhitungan di atas nilai C1 lebih kecil dibanding nilai C2 maka data dengan nilai 1169 di masukan ke dalam C1. Proses perhitungan dilakukan dari data pertama sampai dengan data ke 1000.

Dalam tahapan ini akan dibuat cluster baru dimana kita mempunyai 2 cluster makan dapat kita hitung rata – rata dari setiap cluster dengan menjumlahkan total nilai ke setiap cluster dan dibagi jumlah data distribusi yang memenuhi hasil pada cluster tersebut, sehingga bisa dituliskan hasil sebagai berikut:

$$C1\ [credit] = \frac{(2698164)}{953} = 2831{,}23$$

$$C2\ [credit]: = \frac{(573093)}{47} = 12193{,}46$$

Dikarenakan nili centorid sekarang masih berbeda dibandingkan centroid sebelumnya maka dilakukan proses perhitungan ulang, dengan proses perhitungan eugene distance, menghitung nilai centroid akhir pada setiap perulangan. Dalam proses centroid akhir ditemukan nilai yang sama pada proses iterasi ke 8 dan 9 yan sudah mempunyai nilai yang sama sehingga ditetapkan nilai centroid akhir pada cluster 2 dapat dilihat dalam Tabel 4.18.

Tabel 4. 18 Hasil akhir perhitungan variabel credit dengan 2 cluster

Cluster	Centroid Akhir	Jumlah Anggota
1	2193,3059	827
2	8424.2428	173

Dari tabel 4.18 dapat dilihat nilai dari centroid akhir variabel credit dengan 2 cluster adalah 2193,3059 pada titik 1 dan 8424,2428 pada titik 2 dengan jumlah distribusi data pada cluster 1 sebanyak 827 dan 173 pada cluster 2.

Implementasi proses perhitungan kmeans dalam kode program dapat dilihat sebagai berikut.

```
1.  <?php
2.  $status="false";
3.  $loop=0;
4.  while($status=='false') {
5.  $sql="select * from tb_transformasi";
6.  $result=$mysqli->query($sql);
7.  $x=0;
8.  while ($data=mysqli_fetch_assoc($result)) {
9.  extract($data);
10. for ($i=1;$i<=$jumlahcluster;$i++) {
11. $hasilc[$i]=sqrt(pow($credit-$cluster[$i],2));}
12. $hasil=array_search(min($hasilc), $hasilc);
13. $clusterakhir[$x+=1]=$hasil;
14. mysqli_query($mysqli,"UPDATE tb_transformasi SET
    c_$variabel='C$hasil' where id='$id'");}
15. $loop+=1;
16. for ($i=1;$i<=$jumlahcluster; $i++) {
17. $cluster[$i]=caridata($mysqli,"select avg($variabel) from
    tb_transformasi where c_$variabel='C$i'");}
18. $status='true';
19. for ($i=1;$i<=1000; $i++) {
20. if($clusterawal[$i]!=$clusterakhir[$i]){
21. $status='false'; //Jika  Masih Ada Yang belum sama, maka akan
    diulang,}}
22. if($status=='false'){
23. $clusterawal=$clusterakhir;}
24. echo $loop.' => '.batas($cluster);
25. $sql="select c_$variabel,count(c_$variabel) as jumlah from
    tb_transformasi
26. group by c_$variabel";
27. $result=$mysqli->query($sql);
28. $x=0;
29. while ($data=mysqli_fetch_assoc($result)) {
30. extract($data);
```

```
31.  echo $c_credit."->".$jumlah;
32.  echo "<br>";}
33.  if($status=='true'){
34.  echo "Centroid Akhir : <br>";
35.  echo $loop.' => '.batas($cluster);}
36.  mysqli_query($mysqli,"delete from tb_centroid where
     number='$jumlahcluster'");
37.  foreach ($cluster as $key => $value) {
38.  $centro="C".$key;
39.  mysqli_query($mysqli,"INSERT INTO
     tb_centroid(number,centro,value)
     values('$jumlahcluster','$centro','$value')");
40.  ?>
```

Listing 4.2 Kode Program perhitungan Algoritma K-Means

Pada keterangan kode program 4.2 dapat dijelaskan kode no 1 untuk inisilaliasi status perulangan. Kode no 5 untuk mencari data yang akan dilakukan proses transformasi. Kode np 10 untuk untuk uji jumlah data cluster yang disediakan. Kode no 11 untuk mencari nilai selisih dat dengan nilai centroid yang ada. Kode no 13 memasukkan ke dalam cluster yang paling kecil.

Kode no 20 untuk proses pengecekan apakah centroid pada data sebelumnya sudah sama dengan centroid pada perhitungan akhir. Kode no 21 untuk menegaskan perulangan akan dilakukan kembali jika nilai centroid akhir masih belum sama.

Tahapan selanjutnya dilakukan dengan perhitungan dari cluster 3 sampai dengan cluster 9. Dari hasil perhitungan didapatkan nilai error setiap data dihitung dari selisih data dan nilai titik centroid pada kelompok tersebut. Nilai error ini yang akan digunakan dalam tahapan perhitungan nilai Kelompok terbaik dengan metode Elbow.

Proses terakhir adalah mengevaluasi nilai optimal dengan metode elbow yang dapat dilakukan dengan formula 4.2

$$SSE = \sum_{K=1}^{K} \sum_{X_i \in S_k} \|X_i - C_j\|^2 \quad (4.2)$$

Keterangan

SSE = jumlah nilai selisih error

k = Data pada urutan k

Xi = Nilai data pada variabel

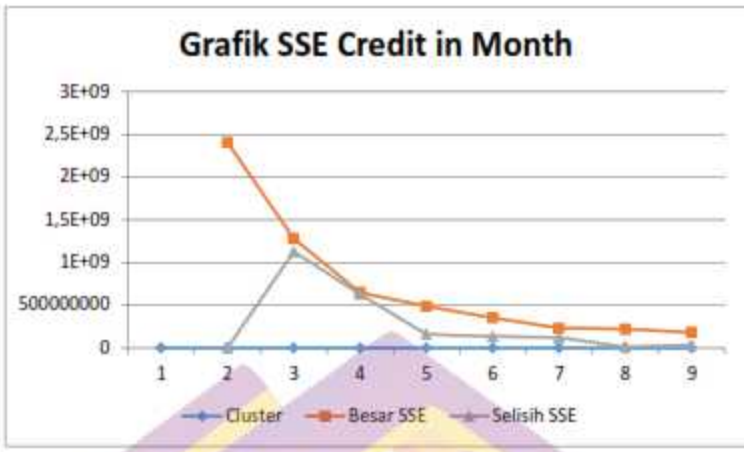
Cj = Nilai titik pusat cluster

Dari proses perhitungan yang dilakukan dilakukan perhitungan dari cluster 2 sampai dengan cluster 9 sehingga dapat dilihat dalam Tabel 4.19

Tabel 4. 19 Hasil perhitungan metode Elbow variabel Credit

Cluster	Besar SSE	Selisih SSE
2	2.405.204.997	0
3	1.277.977.146	1.127.227.850
4	649.788.469	628.188.677
5	488.701.148	161.087.320
6	354.736.854	133.964.294
7	232.489.592	122.247.262
8	220.622.138	11.867.453
9	183.160.442	37.461.696

Dari tabel 4.19 dapat dilihat pergeseran nilai terbesar pada cluster ke 3 dengan selisih 1.127.227.850,8005. Untuk mempermudah dapat dilihat dalam gambar 4.2 pada titik cluster ke 3 terjadi nilai SSE tertinggi dan pergeseran elbow paling signifikan.



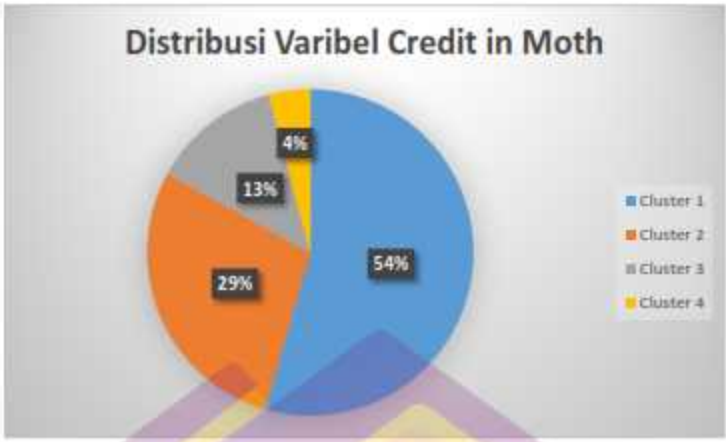
Gambar 4. 2 Grafik SSE Credit in Month

Dari Gambar 4.2 dapat dilihat bahwa titik yang paling mendekati siku terdapat pada cluster 4, sehingga dapat dikatakan cluster yang paling optimal didapatkan pada 4 kelompok. Dari hasil perhitungan yang dilakukan dengan percobaan 7 kali perulangan didapatkan hasil seperti Tabel 4.20

Tabel 4. 20 Hasil akhir perhitungan variabel credit dengan 4 cluster

Cluster	Jumlah Anggota
1	545
2	285
3	129
4	41

Dari Tabel 4.20 dapat dilihat distribusi data pada cluster 1 sebanyak 545, cluster 2 sebesar 285, cluster 3 sebanyak 129, Cluster 4 sebanyak 41. Dari tabel 4.9 dapat dilihat diagram distribusi data yang dapat dilihat dalam Gambar 4.3



Gambar 4. 3 Diagram Distribusi data Credit in Month

Dari Gambar 4.3 dapat dilihat bahwa cluster 1 memiliki distribusi data 55%. Cluster 2 sebesar 28 %, cluster 3 sebesar 13 %, cluster 4 sebesar 4 %.

4.3.2 Transformasi Variabel Duration

Variabel duration in month mempunyai data yang cukup bervariasi dengan data paling kecil sampai paling besar. Tahapan pengelompokan akan dilakukan dengan percobaan 2 sampai dengan 9 clutser untuk dicari distribusi data dan dicari yang paling optimal pengelompokannya.

Hasil Perhitungan selanjutnya dilakukan pada variabel duration yang telah dilakukan dapat dilihat dalam Tabel 4.21

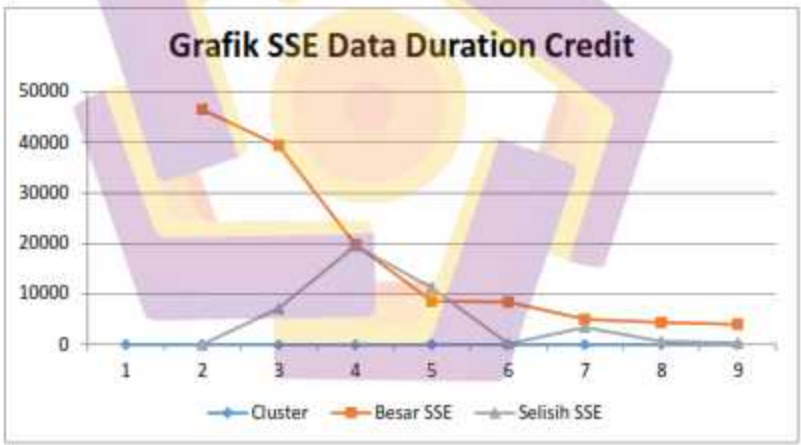
Tabel 4. 21 Hasil perhitungan metode Elbow variabel Duration

Cluster	Besar SSE	Selisih SSE
2	46.538	0
3	39.355	7.182
4	19.823	19.532

Tabel 4. 21 Hasil perhitungan metode Elbow variabel Duration (Lanjutan)

Cluster	Besar SSE	Selisih SSE
5	8.572	11.250
6	8.420	152
7	5.024	3.397
8	4.378	646
9	4.030	348

Dari tabel 4.21 dapat dilihat pergeseran nilai terbesar pada cluster ke 4 dengan selisih 19.532. Untuk mempermudah dapat dilihat dalam Gambar 4.4 pada titik cluster ke 4.



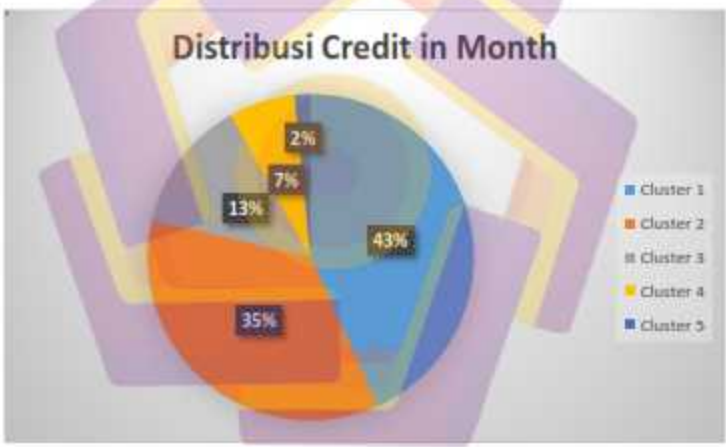
Gambar 4. 4 Grafik SSE duration Credit

Dari hasil Gambar 4.4 dapat dilihat titik siku yang paling dominan terdapat pada cluster 5 sehingga dapat dikatakan cluster 5 merupakan paling optimal. Hasil distribusi data dari cluster 5 dapat dilihat dalamTtabel 4.22 sedangkan untuk hasil

distribusi data dapat dilihat dalam bentuk grafik yang ditunjukan dalam Gambar 4.5

Tabel 4. 22 Hasil akhir perhitungan variabel credit in month dengan 5 cluster

Cluster	Jumlah Anggota
1	433
2	354
3	132
4	65
5	16



Gambar 4. 5 Diagram Distribusi data Credit in Month

Dari Tabel 4.22 dapat dilihat distribusi cluster pada variabel duration dengan distribusi cluster 1 sebesar 433, cluster 2 sebesar 354, cluster 3 sebesar 132, cluster 4 sebesar 65 dan cluster 5 seebsar 16. Gambar 4.5 menampilkan distribusi data dalam satuan persen dengan cluster 1 sebesar 43 %, cluster 2 seebesar 35 %, cluster 3 sebesar 13 %, cluster 4 sebesar 7%, cluster 5 sebesar 2%.

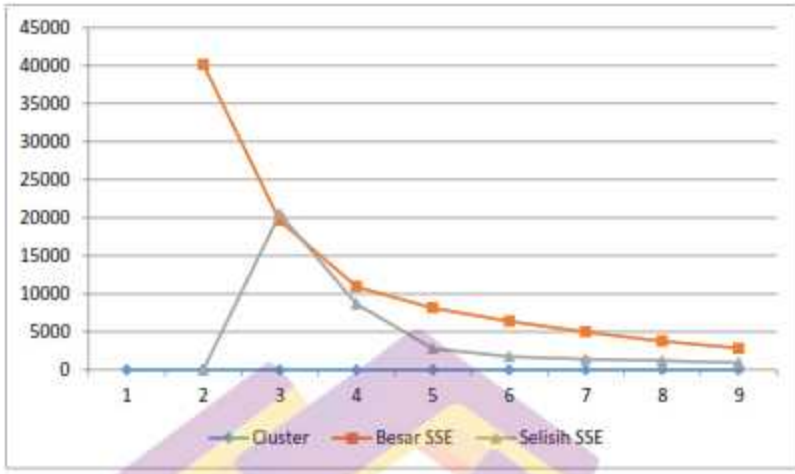
4.3.3 Transformasi Variabel Age

Variabel credit in month mempunyai data yang cukup bervariasi dengan data paling kecil sampai yang paling besar. Tahapan pengelompokan akan dilakukan dengan percobaan 2 sampai dengan 9 cluster untuk dicari distribusi data dan dicari yang paling optimal pengelompokannya. Hasil nilai SSE dapat dilihat dalam Tabel 4.23.

Tabel 4. 23 Hasil perhitungan SSE variabel age

Cluster	Besar SSE	Selisih SSE
2	40.155	0
3	19.617	20.538
4	10.935	8.683
5	8.133	2.802
6	6.393	1.740
7	4.983	1.410
8	3.782	1.201
9	2.847	935

Dari tabel 4.23 dapat dilihat pergeseran nilai terbesar pada cluster ke 3 dengan selisih 20.538. Untuk mempermudah dapat dilihat dalam gambar 4.6 pada titik cluster ke 3 terjadi nilai SSE tertinggi dan pergeseran elbow paling signifikan.



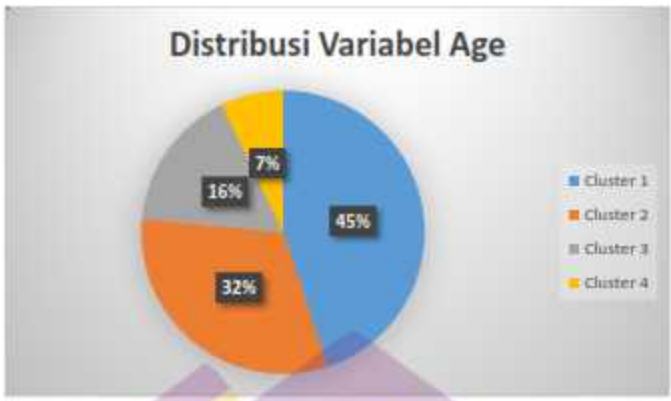
Gambar 4. 6 Grafik SSE Age

Dari Gambar 4.6 dapat dilihat titik yang paling optimal untuk mendekati siku pada nilai besar SSE terdapat pada cluster 4, sehingga dapat dikatakan cluster 4 dikatakan sebagai yang paling optimal. Hasil distribusi data dapat dilihat dalam Tabel 4.24

Tabel 4. 24 Hasil akhir perhitungan variabel age dengan 4 cluster

Cluster	Jumlah Anggota
1	449
2	316
3	164
4	71

Dari Tabel 4.24 dapat dilihat bahwa cluster 1 senilai 449, cluster 2 sebesar 316, cluster 3 sebesar 164 dan cluster 4 sebesar 71. Hasil presentase distibusi dapat dilihat pada Gambar 4.7



Gambar 4. 7 Diagram Distribusi Varianel Age

Dari Gambar 4.7 dapat dilihat distribusi cluster 1 sebesar 45%, cluster 2 sebesar 32 %, cluster 3 sebesar 16 %, cluster 4 sebesar 7%.

4.3.4 Cluster Optimal dengan Metode Elbow

Pengukuran nilai optimal cluster dengan metode Elbow dilakukan dengan 3 kali percobaan dengan 3 kali variabel tunggal. Hasil distribusi data secara keseluruhan dapat dilihat dari tabel 4.25

Tabel 4. 25 Hasil distribusi kelompok cluster paling optimal

Variabel	Jumlah Cluster	Distribusi Data
Credit in Month	4	Cluster 1: 545 Cluster 2: 285 Cluster 3: 129 Cluster 4: 41
Duration Credit	5	Cluster 1: 433 Cluster 2: 354 Cluster 3: 132 Cluster 4: 65 Cluster 5: 16

Tabel 4.25 Hasil distribusi kelompok cluster paling optimal (Lanjutan)

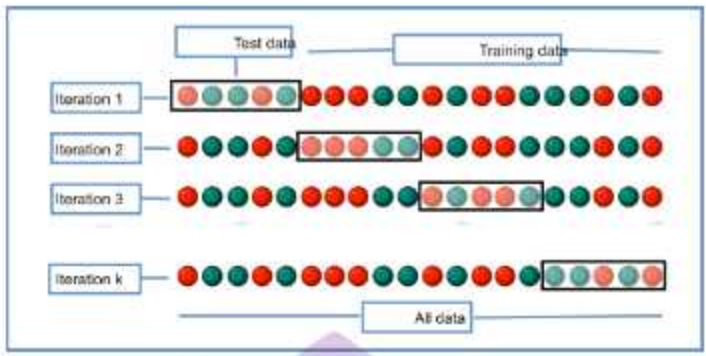
Variabel	Jumlah Cluster	Distribusi Data
Age	4	Cluster 1: 449 Cluster 2: 316 Cluster 3: 164 Cluster 3: 71

4.4 Implementasi Naive Bayes

Penggunaan Algoritma Naive Bayes bertujuan untuk mencari klasifikasi data berdasarkan variabel yang ada dengan terlebih dahulu dicari probabilitas masing - masing variabel. Variabel credit, duration, dan age sudah dapat digunakan untuk pencarian probabilitas dikarenakan nilainya sudah dikelompok kan dengan algoritma K-Means. Hasil dari perhitungan bayes kemudian akan dibandingkan dengan kelas yang sudah didefinisikan dengan hasil perhitungan apakah sesuai atau tidak.

4.4.1 Penggunaan Cross Validation

Pengujian data dilakukan dengan cross validation. Cross validation berfungsi untuk membagi data ke dalam data training da testing. Data training digunakan sebagai sumber data untuk mencari probabilitas sedangkan data testing berfungsi untuk dilakukan pengujian. Penggunaan cross dapat di gambar pada Gambar 4.8



Gambar 4. 8 Contoh Skenario Cross Validation

Dalam Gambar 4.8 dapat dijelaskan skenario dengan contoh pengujian dengan 4 fold, keseluruhan data dibagi menjadi 4 golongan. Pengujian dilakukan dengan 4 kali percobaan dengan setiap percobaan data testing yang digunakan pada kelompok yang berbeda. Implmentasi cross fold dapat dilihat dalam kode program no 4.3

```
1. <?php
2. require_once 'koneksi.php';
3. $jumlahdata=1000;
4. $fold=3;
5. $pembagian=ceil($jumlahdata/$fold);
6. for ($i=0;$i<$fold;$i++) {
7. echo "<br>Iterasi Ke ".($i+1)."<br>";
8. $nilaiawal=(($i*$pembagian)+1);
9. $nilaiakhir=(($i+1)*$pembagian);
10.    if($nilaiakhir>=$jumlahdata)
11.    $nilaiakhir=$jumlahdata;
12.    echo "".$nilaiawal." - ".$nilaiakhir;}
13.    ?>
```

Listing 4.3 Kode Program penentuan cross validation

Pada Listing kode program no 4.3 dapat dijelaskan kode no 3 untuk mengambil jumlah data yang ada 1000. Kode no 4 untuk mengambil nilai fold

sejumlah 10. Kode no 5 untuk menghitung distribusi data. kode no 6 sampai dengan kode 12 berfungsi untuk mencari distribusi data.

4.4.2 Perhitungan Naive Bayes Manual

Dalam contoh perhitungan naive bayes akan dicontohkan perhitungan 1 klasifikasi data dengan uji 500 sebagai daya training, dan data yang dihitung merupakan data pertama yang digunakan dengan atribut sebagai berikut :

Var1: A11

Var2: C1

Var3: A33

Var4: A43

Var5: C1

Var6: A65

Var7: A75

Var8: 4

Var9: A93

Var10 : A101

Var11: 4

Var12: A121

Var13: C3

Var14: A143

Var15: A152

Var16: 2

Var17: A173

Var18: 1

Var19: A192

Var 20: A201

Kelas : 1

Klasfikasi : ?

Tahapan dalam proses perhitungan *Naive Bayes* sebagai berikut:

1. **Menghitung Jumlah Class/ Label.**

   P(Y=1) = 336/500

   P(Y=2) = 164/500

2. **Menghitung jumlah kasus yang sama dengan class yang sama.**

   P(Var1= A1| Y=1) = 65/336

   P(Var1 = A1| Y=2) = 81/164

   P(Var2= A1| Y=1) = 177/336

   P(Var2 = A1| Y=2) = 75/164

   P(Var3= A1| Y=1) = 119/336

   P(Var3 = A1| Y=2) = 25/164

   P(Var4= A1| Y=1) = 111/336

   P(Var5 = A1| Y=2) = 30/164

   P(Var5= A1| Y=1) = 254/336

   P(Var5= A1| Y=2) = 109/164

   P(Var6= A1| Y=1) = 80/336

   P(Var6 = A1| Y=2) = 15/164

   P(Var7= A1| Y=1) = 93/336

P(Var7 = A1| Y=2) = 31/164

P(Var8= A1| Y=1) = 159/336

P(Var8 = A1| Y=2) = 82/164

P(Var9= A1| Y=1) = 188/336

P(Var9 = A1| Y=2) = 85/164

P(Var10= A1| Y=1) = 305/336

P(Var10 = A1| Y=2) = 148/164

P(Var11= A1| Y=1) = 137/336

P(Var11 = A1| Y=2) = 74/164

P(Var12= A1| Y=1) = 101/336

P(Var12 = A1| Y=2) = 34/164

P(Var13= A1| Y=1) = 37/336

P(Var13 = A1| Y=2) = 12/164

P(Var14= A1| Y=1) = 280/336

P(Var14 = A1| Y=2) = 122/164

P(Var15= A1| Y=1) = 256/336

P(Var15 = A1| Y=2) = 95/164

P(Var16= A1| Y=1) = 123/336

P(Var16 = A1| Y=2) = 49/164

P(Var17= A1| Y=1) = 212/336

P(Var17 = A1| Y=2) = 107/164

P(Var18= A1| Y=1) = 278/336

P(Var18 = A1| Y=2) =139/164

P(Var19= A1| Y=1) = 144/336

P(Var19 = A1| Y=2) = 52/164

P(Var20= A1| Y=1) = 323/336

P(Var20= A1| Y=2) = 162/164

3. Mengalikan semua hasil variabel setiap klasifikasi.

Dari perhitungan hasil pencarian probabilitas sebelumnya maka didapatkan hasil nilai ditunjuk kan pada Tabel 4.26

Tabel 4. 26 Hasil perhitungan Probabilitas setiap kelas.

Variabel	Probabilitas Kelas 1	Probabilitas Kelas 2
Var1	0.1935	0.4939
Var2	0.5268	0.4573
Var3	0.3542	0.1524
Var4	0.3304	0.1829
Var5	0.7560	0.6646
Var6	0.2381	0.0915
Var7	0.2768	0.1890
Var8	0.4732	0.5000
Var9	0.5595	0.5183
Var10	0.9077	0.9024
Var11	0.4077	0.4512
Var12	0.3006	0.2073
Var13	0.1101	0.0732
Var14	0.8333	0.7439
Var15	0.7619	0.5793
Var16	0.3661	0.2988
Var17	0.6310	0.6524
Var18	0.8274	0.8476
Var19	0.4286	0.3171

Tabel 4.26 Hasil perhitungan Probabilitas setiap kelas (Lanjutan).

Variabel	Probabilitas Kelas 1	Probabilitas Kelas 2
Var20	0.9613	0.9878
Probabliitas Kelas	0.6720	0,3280
Nilai Akhir	6,4757E-8	0,8475E-8

Dari hasil tabel 4.26, nilai probabilitas tertinggi ada pada kelas (1) dengan nilai 6,4757E-8 sehingga dapat disimpulkan bahwa calon nasabah masuk dalam kategori good. Implementasi perhitungan naive bayes dapat dilihat dalam Listing kode program 4.4

```
1. <?php
2. $p[1]=All($mysqli,"1");
3. $p[2]=All($mysqli,"2");
4. $sql="SELECT * FROM v_data where jenis='Testing'";
5. $result=$mysqli->query($sql);
6. while ($data=mysqli_fetch_assoc($result)) {
7. $kriteria[1]=1;
8. $kriteria[2]=1;
9. for ($i=0;$i<20;$i++) {
10.    $variabel='var'.$i;
11.    $kriteria[1]*=variabel($mysqli,'1',$variabel,$data[$variabel
   ]);
12.    $kriteria[2]*=variabel($mysqli,'2',$variabel,$data[$variabel
   ]);}
13.    $kriteria[1]*=$p[1];
14.    $kriteria[2]*=$p[2];
15.    if($kriteria[1]>$kriteria[2]){
16.    $target='1';
17.    }else{
18.    $target='2';}
19.    $id=$data['id'];
20.    mysqli_query($mysqli,"update tb_master set target='$target'
   where id='$id'");}
21.    ?>
```

Listing 4.4 Kode Program perhitungan Naive Bayes

Dalam Listing kode 4.4 dapat dijelaskan kode no 1 dan 2 digunakan untuk mencari nilai probabilitas kelas 1 dan kelas 2. Kode 4 untuk menseleksi data testing yang akan dilakukan perhitungan. Kode no 9 sampai dengan no 12

digunakan untuk melakukan proses probabilitas variabel. Kode no 13 dan 14 digunkan untuk perhitungan probabilitas akhir. Kode no 15 sampai dengan no 18 digunakan untuk proses pencarian nilai terbesar ke kelas.

4.5 Evaluasi Kinerja Naive Bayes

Evaluasi kinerja Naive Bayes dilakukan dengan confusion matrik. Dalam penggunaan confusion matrik akan dicari data dari data training dan testing terdahulu dengan cross validation. Hasil dari perhitungan testing kemudian dicari hasil yang sesuai dengan target kelas. Ketentuan confusion matrik dapat dililah dalam Tabel 4.27

Tabel 4. 27 Ketentuan Kelas Confusion Matrix

No	Kelas	Keterangan
1	True Positive (TP)	Data dengan kelas good dan menghasilkan perhitungan good
2	True Negative (TN)	Data dengan kelas good dan menghasilkan perhitungan bad
3	False Positive (FP)	Data dengan kelas bad dan menghasilkan perhitungan good
4	False Negative (FP)	Data dengan kelas bad dan menghasilkan perhitungan bad

4.5.1 Pengujian Naive Bayes Tanpa Pengelompokan Variabel

Pengujian perhitungan dengan fold 3 data dibagi kedalam 3 fold dan dilakukan pengujian dengan 3 iterasi. Data pengujian setiap iterasi akan dibagi kedalam data trainig dan testing yang dapat dijabarkan pada Tabel 4.28

Tabel 4. 28 Skenario Pengujian 3 Fold

Iterlasi	Fold	Range Data	Keterangan
1	1	1-334	Data Testing
1	2	335-668	Data Training
1	3	669-1000	Data Training
2	1	1-334	Data Training
2	2	335-668	Data Testing
2	3	669-1000	Data Training
3	1	1-334	Data Training
3	2	335-668	Data Training
3	3	669-1000	Data Testing

Proses pengujian Naive Bayes tanpa K-means digantikan dengan Fungsi Gauss untuk mengantisipasi variabel yang bersifat numerik. Perhitungan dengan fungsi Gauss terlebih dahulu mencari nilai mean dan standar deviasi, kemudian proses pencarian nilai Gauss dengan data training yang digunakan untuk dicari nilai probabilitas terhadap data testing. Proses perhitungan pencarian Gauss dapat dilihat dalam Listing Program 4.5

```
1.  <?php
2.  function fungsi_gaus($variabel,$mean){
3.  $phi=pi();
4.  $etha=" 2.7183";
5.  for ($i=0;$i<2;$i++) {
6.  if ($mean[$i][2]>0){
7.  $pembagi=1/((sqrt(2*$phi))*$mean[$i][2]);
8.  $hmean=pow(($variabel-$mean[$i][1]),2);
9.  $hdeviasi=2*(pow($mean[$i][2],2));
10. $pemangkat=-1*($hmean/$hdeviasi);
11. $hasil=$pembagi*(pow($etha,$pemangkat));
12. $hasilakhir[$i]=$hasil;
13. }else{
14. $hasilakhir[$i]=0;}}
15. return $hasilakhir; }
16. ?>
```

Listing 4.5 Kode Program fungsi perhitungan gaussians

Dari Listing 4.5 dapat dijelaskan kode nomer 2 untuk pembuatan fungsi yang akan dipanggil dalam proses probabilitas. Kode no 3 untuk mendefinisikan nilai pi, kode nomer 7 untuk menghitung nilai pembagi. Kode 8 untuk menghitung nilai pemangkat. Kode no 11 untuk menghitung nilai hasil akhir dari probabilitas Gaus.

Proses pengujian dilakukan dengan 3 iterasi di setiap fold yang diuji. Hasil pengujian dengan 3 fold pada iterasi pertama dapat dilihat confusion matrik pada Tabel. 4.29

Tabel 4. 29 Confusion Matrik Pengujian pada fold 3 iterasi ke 1

Target / Kelas	**Positive**	**Negative**
True	TP (205)	TN (48)
False	FP (37)	FN (44)

Dari tabel 4.29 dapat dihitung nilai akurasi pada pengujian pada fold 3 dan iterasi ke 1 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (205 +48) / (205+48+37+44) * 100 %

Akurasi = 75,75 %

Dari hasil perhitungan nilai akurasi pada fold 3 iterasi ke 1 didapatkan hasil 75,75 %. Perhitungan selanjtunya akan dicari confusion matrik pada fold ke 3 iterasi 2 yang dapat dilihat dalam Tabel 4.30

Tabel 4. 30 Confusion Matrik Pengujian pada fold 3 iterasi ke 2

Target / Kelas	Positive	Negative
True	TP (198)	TN (44)
False	FP (24)	FN (68)

Dari tabel 4.59 dapat dihitung nilai akurasi pada pengujian pada fold 3 dan iterasi ke 1 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (198+44) / 198+44+24+68) * 100 %

Akurasi = 72,46 %

Dari hasil perhitungan nilai akurasi pada fold 3 iterasi ke 2 didapatkan hasil 72,46 %. Perhitungan Selanjutnya akan dicari confusion matrik pada fold ke 3 iterasi 3 yang dapat dilihat dalam Tabel 4.31

Tabel 4. 31 Confusion Matrik Pengujian pada fold 3 iterasi ke 2

Target / Kelas	Positive	Negative
True	TP (203)	TN (51)
False	FP (33)	FN (45)

Dari Tabel 4.31 dapat dihitung nilai akurasi pada pengujian pada fold 2 dan iterasi ke 1 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (203 + 51) / (203+51+33+45) * 100 %

Akurasi = 76,51 %

Dari hasil perhitungan nilai akurasi pada fold 3 iterasi ke 3 didapatkan hasil 76,51 %. Proses pengujian dapat diimplementasikan secara kode program pad a Listing 4.6

```
1. <?php
2. $tp=caridata($mysqli,"select count(*) from tb_master where
   jenis='Testing' and kelas='2' and target='2'");
3. $fp=caridata($mysqli,"select count(*) from tb_master where
   jenis='Testing' and kelas='1' and target='2'");
4. $fn=caridata($mysqli,"select count(*) from tb_master where
   jenis='Testing' and kelas='2' and target='1'");
5. $tn=caridata($mysqli,"select count(*) from tb_master where
   jenis='Testing' and kelas='1' and target='1'");
6. echo "<br>TP + TN : ".$tp."+".$tn;
7. echo "<br> TP + TN + Fp + FN : ".$tp."+".$tn."+".$fp."+".$fn;
8. $query="SELECT(
9. (SELECT COUNT(*) FROM tb_master WHERE jenis='Testing' AND
   kelas=target)/
10.    (SELECT COUNT(*) FROM tb_master WHERE jenis='Testing'))
11.    AS hasil";
12.    $row = $mysqli->query($query)->fetch_array();
13. echo "Jumlah Akurasi $i = ".$row[0];?>
```

Listing 4.6 Kode Program perhitungan akurasi

Dari listing 4.6 dapat dijelaskan kode no 2 untuk mencari nilai true prositif. kode no 2 untuk mencari nilai false positif. Kode no 3 untuk mencari nilai false positif. kode no 5 untuk mencari nilai true negatif. kode no 8 sampai dengan no 13 untuk mencari nilai akurasi.

4.5.2 Pengujian Naive Bayes dengan Pengelompokan K-Means

Pengujian dengan 3 fold data dibagi ke dalam 3 kelompok dengan distribusi masing - masing yang hampir seimbang. Pengujian dengan pengelompokan K-Means dilakukan terlebih dahulu dari pengelompokan variabel yang sudah dilakukan sehingga setiap nilai numerik sudah berubah menjadi label kategorial untuk dapat dihitung probabilitasnya. Hasil pengujian dengan 3 fold pada iterasi 1 dapat dilihat confusion matrik pada Tabel. 4.32

Tabel 4. 32 Confusion Matrik Pengujian pada fold 3 iterasi ke 1

Target / Kels	**Positive**	**Negative**
True	TP (207)	TP (49)
False	FP (45)	FN (43)

Dari Tabel 4.32 dapat dihitung nilai akurasi pada pengujian pada fold 2 dan iterasi ke 1 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (207 + 49) / (207+49+35+43) * 100 %

Akurasi = 76,65%

Dari hasil perhitungan nilai akurasi pada fold 3 iterasi ke 1 didapatkan hasil 76,65 %. Perhitungan selanjutnya akan dicari confusion matrik pada fold ke 3 iterasi 2 yang dapat dilihat dalam Tabel 4.33

Tabel 4. 33 Confusion Matrik Pengujian pada fold 3 iterasi ke 2

Target / Kelas	**Positive**	**Negative**
True	TP (190)	TN (49)
False	FP (32)	FN (63)

Dari tabel 4.33 dapat dihitung nilai akurasi pada pengujian pada fold 2 dan iterasi ke 1 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (190 + 49) / (190+49+32+63) * 100 %

Akurasi = 71,56%

Dari hasil perhitungan nilai akurasi pada fold 3 iterasi ke 2 didapatkan hasil 71,56 %. Perhitungan selanjutnya akan dicari confusion matrik pada fold ke 2 iterasi 3 yang dapat dilihat dalam tabel 4.34

Tabel 4. 34 Confusion Matrik Pengujian pada fold 3 iterasi ke 3

Target / Kelas	**Positive**	**Negative**
True	TP (202)	TN (53)
False	FP (34)	TN (43)

Dari tabel 4.34 dapat dihitung nilai akurasi pada pengujian pada fold 3 dan iterasi ke 3 dengan cara ;

Akurasi = (TP + TN) / (TP+TN+FP+FN) * 100 %

Akurasi = (202+53) / (202+53+34+43) * 100 %

Akurasi = 76,81 %

Dari hasil pengujian pada fold 3 iterasi 3 didapatkan nilai akurasi sebesar 76,81%.

4.5.3 Hasil Evaluasi Pengujian

Dari skenario pengujian yang dilakukan dapat dihitung nilai hasil rata - rata pengujian akhir dengan cara menjumlah nilai setiap rata - rata iterasi pada fold3 dan dibagi dengan sejumlah pengujian yang dilakukan seperti pada Tabel 4.35

Tabel 4. 35 Perbandingan Hasil Uji

Iterasi	Naive Bayes	Naive Bayes + K-Means
Iterasi 1	75,75 %	76,65 %
Iterasi 2	72,46 %	71,56 %
Iterasi 3	76,51 %	76,81 %
Rata - Rata	74,91 %	75,01 %

Dari hasil pengujian pada Tabel 4.35 yang dilakukan pada iterasi pertama lebih besar akurasinya Naive Bayes dengan K-Means dengan nilai 76,65% dibandingkan dengan Naive Bayes dengan nilai akurasi 75,75%. Pengujian pada iterasi kedua pengujian dengan Naive Bayes dan K-Means lebih kecil nilai nya dibandingkan dengan nilai Naive Bayes, dengan perbandingan Nilai 71,56% dibandingkan dengan 72,46%. Pengujian pada iterasi ke 3 nilai uji Naive Bayes dan K-Means menghasilkan nilai 76,81% lebih besar dibanding dengan metode Naive Bayes saja dengan nilai 76,51%. Nilai akurasi keseluruhan/rata-rata akhir menunjukan metode Naive Bayes dengan pengelompokan K-means menghasilkan nilai yang lebih besar dengan nilai 75,01% dibandingkan dengan perhitungan Naive Bayes dengan fungsi gauss dengan nilai 74,91%.

BAB V

PENUTUP

5.1 Kesimpulan

Kesimpulan dari hasil penelitian yang dilakukan yaitu:

1. Hasil pengujian pencarian cluster terbaik dengan metode elbow yang dilakukan dengan skenario uji coba 2 cluster sampai dengan 9 didapatkan hasil optimal untuk variabel *credit in month* terbentuk 4 cluster . Variabel *duration of credit* dengan 5 cluster. Variabel *age* terbentuk dengan 4 cluster.
2. Hasil pengujian akurasi naive bayes dengan pengelompokan terlebih dahulu dengan algoritma K-means menghasilkan nilai akurasi sebesar 75,01 %, sedangkan pengujian dengan naive bayes tanpa pengelompokan terlebih dahulu menghasilkan akurasi 74,91 %.
3. Hasil akurasi antara metode naive bayes dibandingkan dengan metode naive bayes dengan pengelompokan terlebih dahulu dengan K-means clustering pada dataset German Credit Data, menunjukan penggunaan naive bayes dengan pengelompokan K-means lebih baik dengan adanya peningkatan akurasi.

5.1 Saran

Saran penelitian untuk penelitian selanjutnya adalah:

1. Perlunya evaluasi terhadap variabel yang berpengaruh ke dalam akurasi. Sehingga dapat meningkatkan kinerja akurasi yang lebih baik, dalam hal ini bisa dilakukan di dalam tahapan *preprocesing* dengan seleksi variabel yang ketat sehingga bisa dilakukan optimasi.
2. Penggunaan metode naive bayes dengan pengelompokan K-means dapat dilakukan pada dataset yang berbeda apakah menghasilkan nilai akurasi yang lebih baik atau tidak, sehingga bisadijadikan alternatif dalam proses penyempurnaan metode yang sudah ada.

DAFTAR PUSTAKA

PUSTAKA BUKU

Astiko. (1996). Manajemen Perkreditan. Yogyakarta: Andi Offset .

Kusrini. (2009). Algoritma Data Mining. Yogyakarta: Andi Ofset.

Santosa, B. (2007). Data mining: teknik pemanfaatan data untuk keperluan bisnis. Yogyakarta: Graha Ilmu.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Anggodo, y., Cahyaningrum, w., & Fauziyah, a. (2017). Hybrid K-Means Dan Particle Swarm Optimization Untuk Clustering Nasabah Kredit. *urnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 104-110.

Caruso, G., Gattone, S., Fortuna, F., & Battista, T. D. (2020). Cluster Analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, 1-20.

Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., Chidozie, F., F. O., & Guembe, B. (2019). A Naive Bayes approach to fraud prediction in loan default. *International Conference on Science and Sustainable Development*, 1-4.

Hadianto, N., Novitasari, H. B., & Rahmawati, A. (2019). Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network . *Jurnal PILAR Nusa Mandiri* , 163-170.

Harlina, S. (2018). Data Mining On Credit Feasibility Determination Using K-Nn Algorithm Based On Forward Selection . *ISSN : 1978 -8282* , 236-244.

Hasan, M. (2017). Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naive Bayes Berbasis Forward Selection. *ILKOM Jurnal Ilmiah*, 317-324.

Hermadi, I. (2007). Clustering Menggunakan Self Organizing Maps (Studi Kasus: Data PPMB IPB). *Jurnal SAINTIKOM*.

Imran, K. (2013). Determinants of bank credit in Pakistan: A supply side approach. *Economic Modelling 35*, 384–390.

Irwanto. (2012). Optimasi Kinerja Algoritma Klasterisasi K-Means untuk Kuantisasi Warna Citra. *Jurnal Teknik ITS Vol. 1, No. 1*.

Lan Yu, G. C. (2007). Application and Comparison of Classification Techniques in Controlling Credit Risk. *World Scientific* , 111.

Lan, Q., xu, X., Ma, H., & Li, G. (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems With Applications*, 1-12.

Lestari, S., Akmaludin, & Badrul, M. (2020). Implementasi Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Anugerah Bintang Cemerlang . *Jurnal PROSISKO*, 8-16.

M. J. Berry, G. S. (2004). *Data Mining Techniques For Marketing,Sales, and Customer Relationship Management*. Indiana: Wiley Publishing.

Mar'i, F., & Supianto, A. A. (2019). Clustering Credit Card Holder Berdasarkan Pembayaran Tagihan Menggunakan Improved K-Means Dengan Particle

Swarm Optimization. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*.

Marhumi, S. (2017). Analisis Manajemen Perkreditan untuk Meningkatkan Profitabilitas pada Bank BNI Wilayah VII Makassar. *Jurnal Perspektif*, 145-153.

Pujianto, U., Hidayat, M. F., & Rosyid, H. A. (2019). Text Difficulty Classification Based on Lexile Levels Using K-Means Clustering and Multinomial Naive Bayes. 163-170.

Rahman, A. T., Wiranto, & Anggrainingsih, R. (2017). Coal Trade Data Clusterung Using K-Means. *ITSMART: Jurnal Ilmiah Teknologi dan Informasi*, 24-31.

Tripathi, A., Yadav, S., & Rajan, R. (2019). Naïve Bayes Classification Model for the Student Performance Prediction . *International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 1548-1553.

Wahyuningsih, S., & Utari, D. R. (2018). Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit. *Konferensi Nasional Sistem Informasi* , 619-623.

Wirdasari, D., & Calam, A. (2011). Penerapan Data Mining Untuk Mengelola Data Penempatan Buku Di Perpustakaan SMK TI PAB 7 Lubuk Pakam Dengan Metode Association Rule. *Jurnal SAINTIKOM. Vol. 10 / No. 2* .

Xhemali, D., hinde, C. J., & Stone, R. G. (2009). Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *IJCSI International Journal of Computer Science Issues*, 16-23.

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. Proceedings - 6th International Advanced Computing Conference. *IACC*, 78-83.