K-MEANS DAN XGBOOST UNTUK ANALISIS PERILAKU PEMBAYARAN REKENING LISTRIK PELANGGAN (STUDI KASUS : PLN ULP PANAKKUKANG)

Raditya Hari Nugraha[1), 2)], Diana Purwitasari [3),*] Agus Budi Raharjo[3)]

[1)] Analitika Bisnis, Magister Manajemen Teknologi, Institut Teknologi Sepuluh Nopember
Jl. Cokroaminoto No.12A, DR. Soetomo, Kec. Tegalsari, Kota SBY, Jawa Timur
[2)] PLN Unit Pelaksana Pelayanan Pelanggan Makassar Selatan
Jl. Hertasning no.99 Makassar, Sulawesi Selatan
[3)] Teknik Informatika, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya
e-mail: raditya.hn@gmail.com, diana@if.its.ac.id, agus.budi@its.ac.id
*penulis korespondensi

ABSTRAK

Percepatan pendapatan dari piutang rekening listrik merupakan salah satu upaya perusahaan energi untuk dapat menjaga aliran kas sehingga dapat menjalankan kegiatan operasional serta melaksanakan kegiatan investasi pengembangan aset perusahaan. Faktor yang mempengaruhi perilaku pembayaran rekening listrik antara lain lokasi pelanggan, besar nilai tagihan, fasilitas payment point yang ada di sekitar rumah pelanggan, pemanfaatan teknologi digital sebagai sarana pembayaran, serta kesadaran dan pemahaman pelanggan terkait batas waktu pembayaran rekening listrik. Oleh karena itu perlu dilakukan analisis sehingga perusahaan dapat menentukan strategi khusus untuk pelanggan yang berpotensi menunggak rekening listrik. Dalam mengelompokkan perilaku pembayaran rekening listrik, beberapa penelitian terdahulu menggunakan berbagai macam metode klasifikasi pemodelan pembelajaran mesin seperti random forest, naïve bayes, SVM, CART dll untuk mendapatkan tingkat akurasi terbaik. Pada penelitian ini digunakan metode klaster dengan teknik k-means serta menggabungkannya dengan metode klasifikasi eXtreme Gradient Boosting (XGBOOST) berdasarkan data karakteristik pembayaran rekening listrik pelanggan. Pada penelitian ini digunakan juga penyetelan hyperparameter dengan teknik hillclimbing, random search, dan bayesian untuk meningkatkan nilai akurasi dari model. Simulasi model yang dilakukan pada tesis ini memberikan hasil bahwa kombinasi klaster k-means dengan klasifikasi XGBoost serta dengan melakukan penyetelan hyperparameter teknik bayesian memiliki tingkat akurasi model jauh lebih baik dengan nilai 89,27% dan nilai AUC sebesar 0,92 jika dibandingkan dengan metode gradient boosting yang tingkat akurasinya hanya 74,76% dan nilai AUC sebesar 0,75. Berdasarkan hasil simulasi pada data pelanggan ULP Panakkukang didapatkan bahwa kelompok pelanggan kategori subsidi dan pelanggan yang sering mengalami pemadaman aliran listrik memiliki kecenderungan untuk menunggak piutang rekening listrik.

***Kata Kunci**: Piutang, Rekening Listrik, K-means, eXtreme Gradient Boosting (XGBOOST), Hyperparameter.*

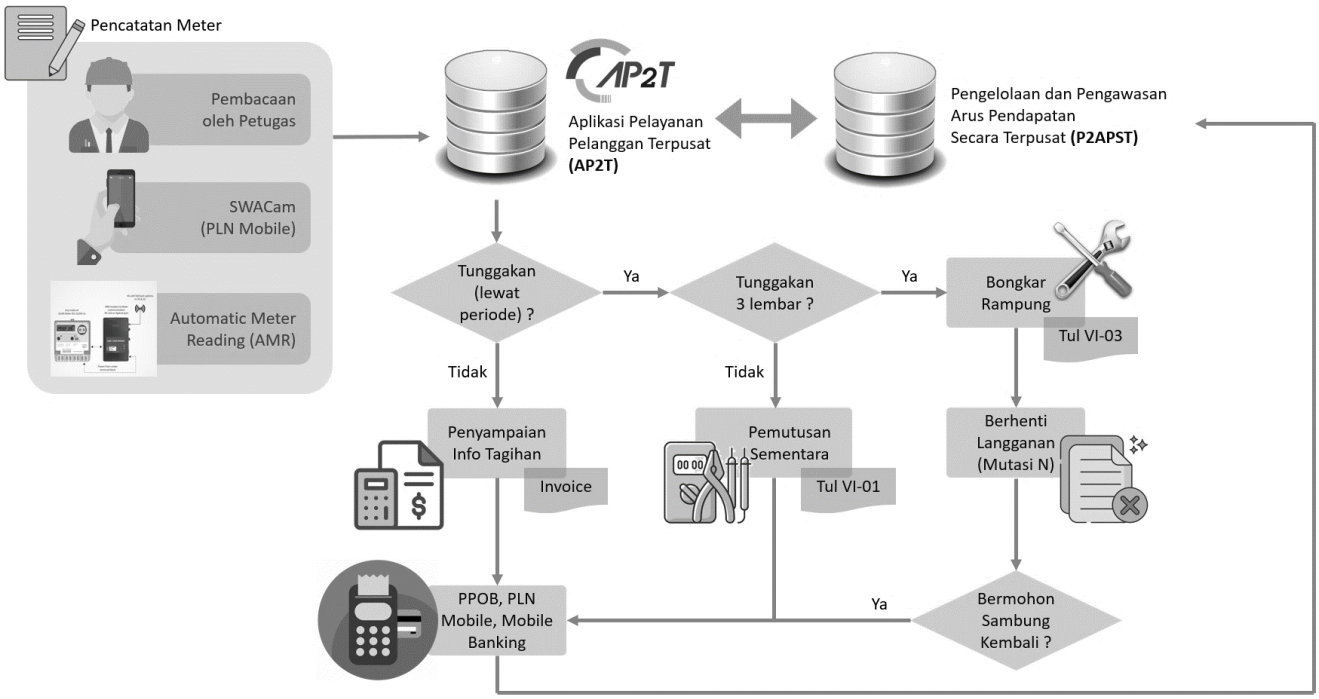
ABSTRACT

Revenue Acceleration from electricity account receivables is one of the energy companies' efforts to maintain cash flow so that they can carry out operational activities and carry out investment activities to develop company assets. Factors that influence electricity bill payment behavior include the location of consumers, the amount of the bill, payment point facilities located around consumers' homes, the use of digital technology as a media of payment, as well as consumer awareness and understanding regarding the time limit for paying electricity bills. Therefore, it is necessary to conduct an analysis so that the company can determine a special strategy for customers who have the potential to be in arrears in electricity bills. To get the characteristic of electricity bill payments, several previous studies have used various classification methods of machine learning such as random forest, nave bayes, SVM, CART, etc. to get the best accuracy. In this research, to increase the accuracy of the model, author using the cluster method with the k-means technique and combining it with the eXtreme Gradient Boosting (XGBOOST) classification method based on data on the characteristics of consumer electricity bill payments. In this study also used hyperparameter adjustment with hillclimbing, random search, and bayesian techniques to increase the accuracy of the model. The model simulation carried out in this thesis gives the result that the combination of the k-means cluster with the XGBoost classification and by adjusting the bayesian technique hyperparameters has a much better model accuracy rate with a value of 89.27% and an Area Under Curve (AUC) value of 0.92 when compared to gradient boosting method with an accuracy rate of only 74.76% and an AUC value of 0.75. Based on the simulation results on ULP Panakkukang customer data, it was found that the subsidy category customer group and customers who often experience power outages have a tendency to be in arrears on electricity bills.

***Keywords**: Account Receivables, Electricity Payment, K-means, eXtreme Gradient Boosting (XGBOOST).*

I. PENDAHULUAN

Memastikan penagihan rekening pada pelanggan menjadi hal penting dalam akitvitas operasional rutin sehingga perusahaan listrik dapat tumbuh dan mendapatkan keuntungan [1] [2] [3]. Monitoring efektif terhadap faktor-faktor yang dapat mempengaruhi perilaku pelanggan untuk melaksanakan kewajiban membayar rekening listrik menjadi perhatian bidang pemasaran dan pelayanan [4]. Secara umum disepakati bahwa



Gambar 1. Diagram Alir Piutang Rekening Listrik

piutang dapat menimbulkan sumber kesulitan keuangan bagi perusahaan ketika mereka tidak dikelola secara efisien [3]. Pengalaman di beberapa industri menunjukkan bahwa manajemen piutang yang efektif dan keseluruhan kinerja keuangan perusahaan berkorelasi positif. Dalam hal konsumen melaksanakan kewajiban membayar tagihan rekening listrik nya, maka perusahaan harus bertindak proaktif dan berkomunikasi dengan baik untuk mempercepat pembayaran kewajiban tepat waktu dan tidak terlambat [5]. Berdasarkan penelitian yang dilakukan oleh Appin dan Irman dengan menggunakan metode *Interpretive Structural Modeling* (ISM) terdapat 10 unsur faktor penyebab tunggakan tinggi, yakni kebiasaan pelanggan menunda pembayaran, pembayaran listrik belum menjadi prioritas, kondisi ekonomi pelanggan, denda keterlambatan rendah, kurangnya SDM PLN untuk menagih, belum ada sosialisasi yang masif, dan tidak ada sanksi yang tegas kepada vendor *billing management* [6].

Di dalam penelitian yang dilakukan oleh Zeng dan timnya membahas masalah pengurangan luar biasa piutang melalui perbaikan dalam strategi penagihan bahkan menghasilkan penghematan waktu penagihan sampai dengan empat kali dengan menggunakan model pembelajaran mesin yang telah dibuat [7].

Penggunaan pembelajaran mesin untuk mengelola data piutang dapat dilakukan dengan berbagai macam metode, namun beberapa langkah penelitian yang harus diperhatikan antara lain pembersihan data dan pemrosesan, analisis statistik, serta membuat model pembelajaran mesin, dan mengevaluasi kinerja model seperti yang dijelaskan pada penelitian yang dilakukan oleh Weikun [8]. Penelitian yang dilakukan oleh Weikun membahas pendekatan menangani data yang tidak seimbang, teknik pengambilan sampel, pengukuran kinerja dan ansambel algoritma. Sebelum mengelola data yang cukup banyak dan kompleks tersebut maka juga harus diperhatikan data tersebut sudah memiliki karakteristik data yang baik.

Ana Paula [9] beserta timnya melakukan penelitian terkait dengan metode pembelajaran mesin yang digunakan dalam pengelolaan piutang dapat melakukan penghematan biaya dan waktu di segala lini kegiatan kolektor piutang setiap bulannya. Penelitian tersebut menghadirkan sebuah prototipe yang dapat mendukung kolektor dalam memprediksi pembayaran faktur dengan data yang kompleks. Beberapa model pembelajaran mesin yang digunakan pada penelitian tersebut memberikan tingkat akurasi model yang sangat baik khususnya pada metode *eXtreme Gradient Boosting* (XGBoost), *Random Forest* dan *Deep Neural Network* (DNN) [9].

Alur proses pembayaran rekening listrik hingga pelanggan dinyatakan memiliki piutang dapat ditunjukkan pada gambar 1 diatas. Pada gambar tersebut kita dapat lihat terdapat beberapa proses yang memiliki data histori atas pelanggan sehingga dapat dilihat kecenderungan perilaku pelanggan terhadap kewajiban membayar rekening listrik. Pada penelitian ini digunakan data primer di PLN UP3 Makassar Selatan ULP Panakkukang selama kurun waktu bulan April sampai dengan bulan September tahun 2021 atas perilaku pembayaran rekening listrik pelanggan. Selain itu terdapat data-data pendukung lainnya seperti data induk langganan (DIL), serta histori gangguan yang di rasakan pelanggan tiap tahun. Dengan mempertimbangkan beberapa aspek disisi pelanggan seperti kategori pelanggan subsidi atau non subsidi, jarak rumah pelanggan dengan *payment point online bank*

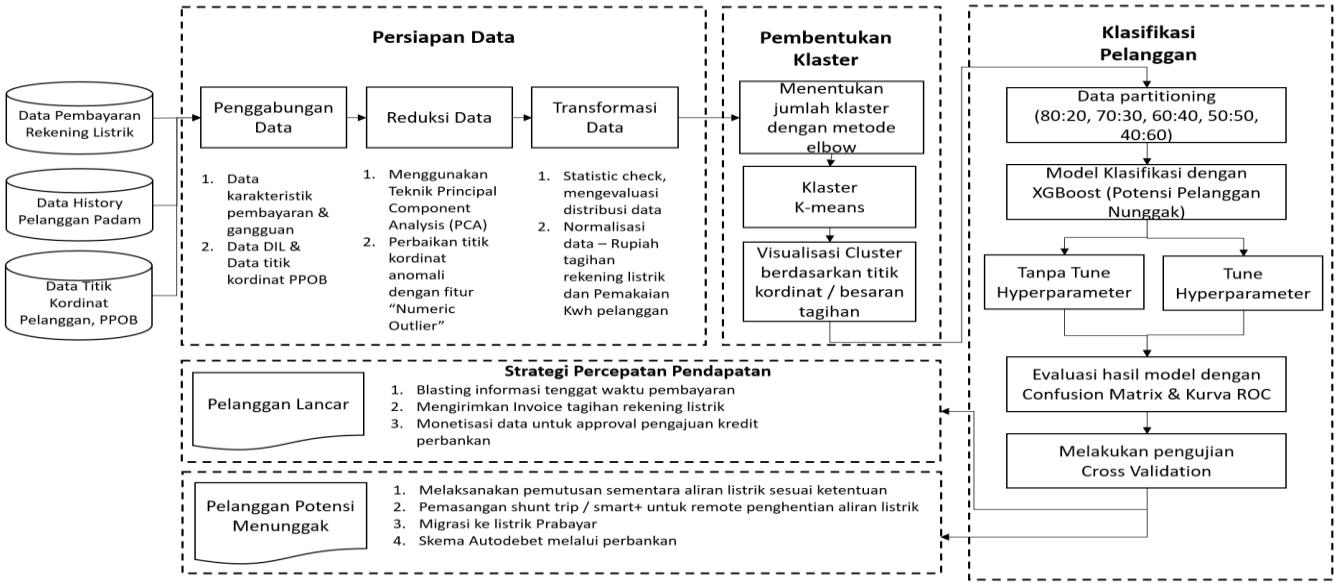
(PPOB) terdekat yang akan mempengaruhi percepatan aliran kas maka korelasi perilaku pelanggan dalam menyelesaikan kewajibannya berupa membayar tagihan rekening listrik akan didapatkan [10].

Sanksi yang akan dikenakan kepada pelanggan bila tidak menyelesaikan tagihan rekening listriknya adalah berupa biaya keterlambatan, dimana akhir waktu pembayaran adalah pada tanggal 1-20 setiap bulannya. Biaya keterlambatan ini dihitung sesuai dengan golong tarif dan berapa bulan tunggakan rekening yang terjadi [11]. Selain itu, PLN akan melaksanakan kegiatan pemutusan sementara terhadap listrik pelanggan bila pelanggan belum melakukan pelunasan rekening listriknya. Sedangkan bila dalam jangka waktu 60 hari sejak dilaksanakannya pemutusan sementara pelangan belum melunasi pembayaran rekening listriknya maka akan dilakukan pembongkaran titik transaksi berupa kwh meter dan assesorisnya serta akan diberhentikan sebagai langganan PLN [12]. Analisis karakteristik pelanggan dalam pembayaran rekening listrik akan memudahkan divisi pelayanan pelanggan untuk melakukan pendekatan atau penerapan strategi sehingga dapat tercipta budaya membayar rekening listrik tepat waktu serta potensi penempatan *payment point online bank* (PPOB) yang efektif. Pada penelitian yang dilakukan oleh nurul dan edi, adalah dengan melakukan pengelompokkan pelanggan dengan metode klastering *k-means* dengan menggunakan variable pelanggan menunggak, tarif daya, dan jumlah tunggakan sehingga didapatkan data tunggakan terbesar, sedang, dan rendah per daerah (desa) yang kemudian akan dilakukan fokus kegiatan sosialisasi untuk menumbuhkan kesadaran masyarakat terkait pembayaran kewajiban rekening listriknya [13]. Selain itu pengelolaan piutang pelanggan dengan metode klasifikasi juga dilakukan oleh dinda dan teman-temannya yakni dengan memprediksi keterlambatan pembayaran rekening listrik menggunakan komparasi metode klasifikasi *decision tree* dan *support vector machine* (SVM) dengan mempertimbangkan karakteristik pelanggan dalam membayar rekening listrik antara lain penghasilan, jumlah tanggungan, tanggal penerimaan gaji, proses pembayaran, kualitas kwh, biaya denda [14]. Hasilnya prediksi pelanggan dengan metode klasifikasi *decision tree* memiliki tingkat akurasi yang lebih tinggi jika dibandingkan dengan metode SVM. Metode lainnya yang digunakan dalam upaya mempermudah proses penekanan piutang rekening listrik adalah dengan metode *geolocation tagging* atau *geotag* yang dilakukan oleh yessy dan teman-temannya, sehingga dengan demikian diharapkan petugas lapangan akan mudah dalam menjalankan aktivitas penagihan kepada pelanggan dengan adanya support sistem tersebut [15].

Saat ini penelitian-penelitian yang telah dilakukan berkaitan dengan pemetaan piutang rekening listrik adalah hanya sebatas pengelompokkan pelanggan berdasarkan klaster tertentu yang kemudian akan digunakan sebagai pemetaan strategi eksekusi. Namun terhadap klaster tersebut belum pernah dilakukan prediksi lanjutan untuk mendapatkan pelanggan yang berpotensi lancar dan menunggak dalam pembayaran rekening listrik. Selain itu metode klasifikasi XGBoost dengan memanfaatkan *hyperparameter* sebagai peningkatan akurasi model masih belum pernah digunakan pada kasus data pelanggan di perusahaan listrik terkait karakteristik pembayaran rekening pelanggan. Beberapa hasil penelitian terdahulu terkait dengan pengelolaan piutang yang telah dijelaskan sebelumnya memberikan inspirasi dan beberapa metode nya dapat diterapkan pada penelitian saat ini dengan topik pembahasan pengelolaan piutang rekening listrik berdasarkan karakteristik pembayaran yang dilakukan oleh pelanggan.

Penelitian ini bertujuan untuk mendapatkan model perilaku pembayaran rekening listrik pelanggan dengan menggunakan kombinasi metode *k-means* dan metode *eXtreme Gradient Boosting* (XGBoost) sehingga hasil dari model akan digunakan sebagai penentuan strategi khusus terhadap kelompok pelanggan-pelanggan tertentu yang akan berdampak terhadap percepatan pendapatan (*cash-in*) perusahaan. Selain itu untuk mendapatkan model dengan tingkat akurasi yang baik berdasarkan *confusion matrix, Area Under Curve* (AUC), serta kurva ROC maka dilakukan juga penyetelan terhadap beberapa parameter yang tersedia. Selain itu juga akan dilakukan pengujian terhadap model dengan menggunakan teknik *cross validation*.

II. METODOLOGI PENELITIAN



Gambar 2. Diagram Alir Penelitian

Sesuai dengan model yang dikembangkan dalam penelitian ini seperti yang ditunjukkan pada gambar 2 diatas maka analisis data dan simulasi yang digunakan adalah *unsupervised learning* menggunakan teknik *k-means* dikombinasi dengan *supervised learning* menggunakan teknik *eXtreme Gradient Boosting* (XGBOOST) yang dioperasikan melalui aplikasi KNIME. Penggunaan pembelajaran mesin pada pengelompokkan piutang rekening listrik telah dilakukan oleh beberapa peneliti terdahulu antara lain oleh Hetul Shah [16], Anna Paula [9] dikarenakan merupakan suatu hal yang paling utama oleh sebuah perusahaan listrik untuk dapat mengelola piutangnya sehingga dapat memberikan keuntungan terhadap perusahaan atas bisnis yang dilakukan. Sedangkan metode kombinasi klaster *k-means* dan klasifikasi XGBoost digunakan untuk mendapatkan hasil pengelompokan yang memiliki akurasi terbaik seperti yang dilakukan oleh Pan Tang [17] pada kasus prediksi pelanggan *churn* di perusahaan telekomunikasi dan Duan Ran [18] pada prediksi kecelakaan lalu-lintas. Penelitian lainnya yang dilakukan oleh Joao Henriques yang berfokus untuk mendapatkan model dengan akurasi tinggi dengan jumlah data yang sangat banyak dengan melakukan teknik kombinasi klaster *k-means* dan XGBoost pada deteksi anomali data log [19]. Hal lain yang dibahas oleh Zeeshan pada prediksi personality MBTI dengan menggunakan kombinasi *k-means* dan XGBoost yakni teknik tersebut memiliki hasil model yang jauh lebih baik dibandingkan dengan metode seperti naïve bayes dan algoritma lainnya [20]. Dengan terlebih dahulu melakukan pengelompokkan menggunakan metode klaster maka akan dapat meningkatkan akurasi dari model, selain itu dengan menggunakan metode klasifikasi XGBoost maka hasil akurasi dapat ditingkatkan melalui penyetelan *hyperparameter* yang tersedia. Hal ini menjadi salah satu pertimbangan peneliti menggunakan metode kombinasi klaster dan klasifikasi untuk menentukan pengelompokkan pelanggan tertib membayar rekening listrik dan pelanggan yang berpotensi menunggak.

A. Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data primer dari pelanggan perusahaan listrik di Indonesia selama kurun waktu 6 bulan yakni bulan April sampai dengan September tahun 2021. Data tersebut merupakan data yang berasal dari pemakaian listrik pelanggan setiap bulan serta data induk langganan pelanggan berupa titik kordinat yang dapat diakses pada Aplikasi Pelayanan Pelanggan Terpadu (AP2T). Selain itu data histori waktu pembayaran rekening listrik pelanggan yang berupa tanggal pembayaran merupakan data yang diakses dari aplikasi Pengelolaan dan Pengawasan Arus Pendapatan Secara Terpusat (P2APST) yakni sejak bulan April sampai dengan September tahun 2021. Data lain yang ikut dimanfaatkan adalah data jumlah padam per tahun tiap pelanggan yang juga akan dikorelasikan dengan perilaku pembayaran rekening listrik pelanggan. Total jumlah pelanggan yang akan dianalisis perilaku nya adalah sebanyak 58.086 pelanggan.

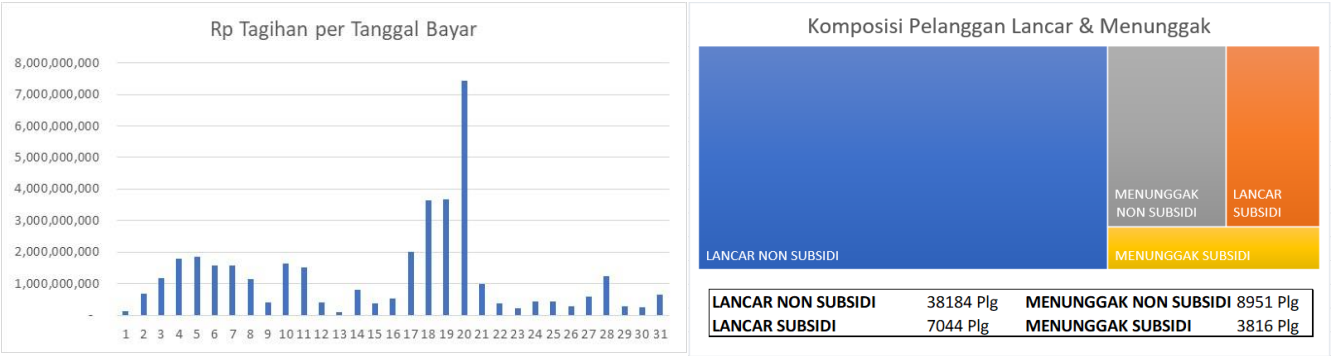
Pada Tabel I, Telah dipetakan struktur *dataset* yang akan digunakan pada penelitian ini. Beberapa fitur utama yang akan digunakan yakni pada tahapan klaster digunakan fitur jarak PPOB ke lokasi pelanggan yang akan direduksi dimensi nya dari 34 dimensi menjadi 2 dimensi. Sedangkan pada tahapan klasifikasi fitur utama yang digunakan antara lain tanggal pembayaran rekening lsitrik, jenis pelanggan, jam nyala rata-rata pelanggan, jumlah trip pelanggan/tahun. Pada penelitian ini tahapan klasifikasi yang akan dilakukan adalah untuk mendapatkan target/kelas pelanggan berpotensi lancar membayar rekening listrik dan pelanggan berpotensi menunggak rekening listrik.

TABEL I
STRUKTUR DATASET KARAKTERISTIK PEMBAYARAN REKENING LISTRIK

No	Nama Field	Satuan	Tipe data	Keterangan
1	IdPel	-	*String*	Id tiap pelanggan
2	Tarif	-	*String*	Kategori tarif pelanggan
3	Jenis Plg	-	*String*	Subsidi/Non subsidi
4	Daya	VA	*String*	Kapasitas listrik pelanggan
5	Rekening Listrik	Kwh	*Numerical*	Pemakaian listrik bln Maret
6	Tanggal Transaksi	waktu	*Numerical*	Tanggal pembayaran
8	Rupiah Tagihan	Rp	Numerical	Besar nilai tagihan reklist
9	Rupiah BK	Rp	Numerical	Nilai biaya keterlambatan
7	Jam Nyala Rata-rata 3 Bln	JN	*Numerical*	Rata-rata pemakaian plg dalam kurun waktu 1 bulan
8	Jumlah Trip	kali	*Numerical*	Frekuensi padam pelanggan dalam 1 tahun
9	Long	-	*Numerical*	Titik kordinat pelanggan
10	Lat	-	*Numerical*	Titik kordinat pelanggan

Selanjutnya berdasarkan Gambar 3 dibawah ini juga dapat terlihat bahwa jika dilakukan pemetaan terhadap data karakteristik pembayaran rekening listrik pelanggan ULP Panakkukang bahwa kecenderungan waktu pelanggan melakukan pembayaran rekening listrik yakni pada tanggal 18,19, dan 20 dimana merupakan batas akhir waktu pembayaran rekening listrik di tiap bulannya. Pada tanggal 20 merupakan puncak hari dimana pelanggan melakukan pembayaran rekening listrik yakni sebanyak 6.773 pelanggan dengan total nilai transaksi sebesar Rp 7.419.161.637,-

Jika diamati terhadap data komposisi pelanggan ULP Panakkukang seperti ditunjukkan pada gambar 3 dibawah, dapat terlihat bahwa komposisi pelanggan terbesar adalah pelanggan yang lancar dalam membayar rekening listrik dan merupakan pelanggan dengan kategori non subsidi yakni sebanyak 38.184 pelanggan. Namun komposisi pelanggan menunggak dan masuk kategori non subsidi juga masih cukup besar yakni sebanyak 8.951 pelanggan yang akan menjadi tantangan PLN untuk melakukan edukasi serta aksi dilapangan sehingga piutang tersebut dapat masuk secepatnya kepada keuangan perusahaan. Dengan melakukan pengelompokkan pelanggan lancar dan berpotensi menunggak menggunakan pembelajaran mesin maka diharapkan pelanggan yang membayar rekening listrik diawal waktu akan semakin banyak sehingga arus kas perusahaan akan baik.



Gambar 3. Evaluasi Data Piutang PLN ULP Panakkukang

B. Pra-pemrosesan data

Pra-pemrosesan data merupakan beberapa teknik yang dilakukan sebelum melakukan proses simulasi dan analisis data yang digunakan untuk menghapus *noise*, data yang tidak lengkap, dan data yang tidak konsisten. Pra-pemrosesan data terdiri dari beberapa kegiatan antara lain pembersihan data, pemilahan data, transformasi data, dan pengurangan data. Pra-pemrosesan data ini biasanya digunakan sehubungan pada data primer seringkali ditemukan bahwa data tidak lengkap, data cenderung tidak konsisten, terdapat nilai error serta ditemukan data tidak penting lainnya. Hal ini mengakibatkan hasil pengolahan data menjadi tidak akurat dan tidak dapat dipertanggung-jawabkan hasilnya. Tahapan pra-pemrosesan data harus dilakukan dengan sesuai agar data yang digunakan dapat tepat sesuai kebutuhan serta dapat memberikan nilai akurasi yang baik saat di simulasikan. Selain itu dilakukan juga pengujian signifikansi variabel terhadap respon dengan menggunakan fitur "*correlation*" pada aplikasi KNIME, sehingga dapat diketahui apakah antar variable dependan dan berapa nilai signifikansi nya berdasarkan matrix korelasi.

Pra-pemrosesan data yang dilakukan pada penelitian ini antara lain :

1. Melakukan pemilahan data atas data histori pembayaran rekening listrik pelanggan, data pemakaian energi

listrik pelanggan serta data induk langganan menggunakan fitur "*Value filter*". Data yang digunakan hanya yang memiliki keterkaitan dengan tujuan mendapatkan klaster serta pengklasifikasian pelanggan berdasarkan karakteristik pembayaran.

2. Titik kordinat pelanggan pada Data Induk Langganan (DIL) tidak semuanya akurat bahkan terdapat pelanggan yang belum terdapat titik kordinatnya. Untuk kondisi tersebut maka dilakukan perbaikan data menggunakan fitur "*Missing value*" dengan cara menghilangkan data yang tidak terdapat titik kordinat, serta menggunakan fitur "*Numeric outlier*" untuk melakukan seleksi pelanggan dengan titik kordinat yang terdapat penyimpangan sangat tinggi.
3. Dilakukan penggabungan 3 data primer dengan menggunakan fitur "*joiner*" dengan menjadikan id pelanggan sebagai target penggabungan.
4. Dilakukan pengurangan data pada variabel jarak pelanggan terhadap PPOB dengan merubah satuan longitude/latitude menjadi satuan jarak (km). Selanjutnya dengan menggunakan teknik *Principal Component Analysis* (PCA), jarak pelanggan ke masing-masing titik PPOB sebanyak 34 dimensi dirubah menjadi 2 dimensi. PCA adalah metode yang ideal untuk menyelesaikan permasalahan fitur data yang terlalu banyak dengan cara mengurangi dimensi data dari kumpulan data yang kompleks [21]. Tahapan Teknik PCA yang pertama adalah dengan melakukan standarisasi atas seluruh variable yang akan digunakan. Tahapan kedua adalah dengan melakukan perhitungan matriks kovarian sehingga dapat diketahui hubungan antara variable. Selanjutnya pada tahapan ketiga yakni dilakukan dengan menghitung nilai dan vektor eigen matriks kovarian sehingga komponen utama dapat teridentifikasi. Dan pada tahapan keempat dilakukan dengan vektor fitur dimana komponen yang kurang signifikan dapat dibuang (yang nilai eigennya rendah).

C. Klasterisasi pelanggan menggunakan metode K-means

Metode klaster merupakan proses pengelompokkan data kedalam grup sehingga objek dalam grup memiliki tingkat kesamaan tertentu yang cukup tinggi jika dibandingkan dengan grup lainnya. Metode klaster merupakan salah satu pembelajaran mesin yang sangat bermanfaat dalam proses pengelompokkan karakteristik tertentu dari pelanggan [22, 15]. Pada penelitian ini, metode klaster *k-means* akan digunakan untuk mendapatkan pengelompokan pelanggan berdasarkan rasio tanggal pembayaran rekening rekening listrik dan titik kordinat pelanggan terhadap *payment point online bank* (PPOB) berupa ATM, indomaret, alfamidi, dan alfamart. Metode *k-means* merupakan salah satu teknik klaster *centroid-based* di mana klaster dibentuk dari jarak terdekat antara titik data ke pusat klaster dengan karakteristik cenderung lebih efisien namun sensitif terhadap *outlier*.

Variabel yang digunakan dalam metode *k-means* antara lain nilai rasio tanggal pembayaran rekening listrik pelanggan dan jarak (meter) antara rumah pelanggan ke titik PPOB terdekat. Kedua variabel ini digunakan secara bersamaan sehingga didapatkan klaster khusus yang terkait antara besar nilai tagihan dan lokasi pelanggan. Beberapa tahapan yang dilakukan untuk melakukan pemodelan klaster, antara lain :

1. Sebelum melakukan metode klaster maka akan ditentukan besarnya nilai k dengan menggunakan metode elbow. Metode *elbow* ini dilakukan dengan cara memilih nilai klaster yang memiliki penurunan nilai *Sum of Square Error* (SSE) terbesar serta ditunjukkan pada sudut dalam grafik antara nilai pada klaster pertama dengan nilai klaster kedua. Jumlah klaster menentukan besarnya nilai SSE, semakin besar jumlah klaster maka akan semakin kecil nilai SSE. Berikut adalah cara menghitung nilai SSE seperti ditunjukkan pada persamaan (1) :

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} \|X_i - C_k\|^2 \quad (1)$$

dimana K adalah jumlah kelompok yang digunakan pada algoritma k-means. X_i adalah jumlah data dan C_k adalah banyaknya klasteri pada klaster ke k.

2. Mengukur nilai *silhouette* terbaik dengan nilai yang akan dicapai adalah >0.5. Nilai silhouette diukur dengan menghitung semua obyek pada klaster dengan klaster terdekatnya dengan persamaan (2) sebagai berikut :

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(ci) \quad (2)$$

Dimana :

$sil(k)$: nilai *silhouette* semua klaster
$|k|$: jumlah klaster k
$sil(c)$: rata-rata nilai *silhouette*

3. Titik *centroid* pada teknik klaster *k-means* menggunakan "*random initialization*" dengan jumlah iterasi sebanyak 99 kali.

4. Hasil pemodelan klaster kemudian akan divisualisasikan pada diagram *scatter plot* agar dapat memudahkan analisis lanjutan.

Model klaster digunakan untuk mendapatkan hasil klasifikasi yang lebih tepat sasaran sebagai strategi eksekusi lapangan oleh perusahaan sehingga didapat juga efektifitas SDM dan biaya. Setelah didapatkan klaster tersebut maka akan digunakan sebagai data inputan untuk melakukan metode klasifikasi pelanggan.

D. Klasifikasi menggunakan Metode XGBoost

Pada penelitian ini akan digunakan teknik *eXtreme Gradient Boosting* (XGBoost) yang merupakan salah satu teknik pengklasifikasian yang memiliki tingkat akurasi yang sangat baik [23]. Algoritma XGBoost juga dapat diidentifikasi sebagai pendekatan untuk mengoptimalkan algoritma *Gradient Boosting* dengan menghapus *missing value*, menghilangkan masalah overfitting dengan menggunakan pemrosesan paralel. Selain itu metode XGBoost juga memberikan hasil klasifikasi yang sangat baik pada situasi data yang tidak seimbang [24] [25].

Cara kerja XGBoost sebenarnya merupakan metode ensemble yang didasarkan pada *gradient boosting tree*. *Gradient boosting tree* yang merupakan algoritma pembelajaran mesin yang mencoba dengan akurat untuk memprediksi variable target dengan menggabungkan perkiraan satu set model yang lebih sederhana dan lebih lemah sehingga nantinya akan terbentuk prediksi akhir yang lebih akurat [23]. Setiap pohon regresi akan memetakan titik data input kepada salah satu daun yang nanti akan diinisiasikan pada skor tertentu dan hal tersebut dilakukan secara terus menerus. Pelatihan dilakukan secara berulang, gradien yang dimaksud adalah dengan penambahan pohon baru yang digunakan untuk mendapatkan skor tertentu dengan tetap memprediksi residu/kesalahan pada pohon sebelumnya kemudian dikombinasikan dengan pohon sebelumnya sehingga dapat membuat prediksi akhir. Pada XGBoost, di dalam pohon regresi dimana simpul yang berada pada sisi dalam merupakan nilai untuk tes atribut dan simpul daun mewakili keputusan. Nilai prediksi merupakan nilai yang diprediksi oleh pohon k seperti yang ditunjukkan pada persamaan (3) berikut :

$$\hat{y}_l = \emptyset(x_i) = \sum_{k}^{K} f_k(x_i), f_k \quad \in F \tag{3}$$

$$F = \{f_x = w_{q(x)}\}, \qquad q: R^m \longrightarrow T, w \in R^T$$

dimana k merepresentasikan jumlah dari pohon, x_i dimana i merupakan jumlah iterasi dari proses pelatihan, dan f_k berupa aturan keputusan pohon dan bobot skor daun. Pada persamaan (3), q merepresentasikan struktur pohon, kemudian T adalah jumlah simpul daun, serta w diinisiasi sebagai bobot simpul daun.

Klasifikasi menggunakan metode XGBoost akan dilakukan setelah didapatkan hasil pemodelan klaster *k-means* antara nilai rupiah tagihan pembayaran rekening listrik pelanggan terhadap jarak pelanggan ke titik *Payment Point Online Bank* (PPOB). Selanjutnya dengan menggunakan beberapa variable lain maka akan didapatkan jenis pelanggan yang lancar dalam membayar dan pelanggan yang berpotensi menunggak.

Pada penelitian ini juga akan dilakukan perbandingan nilai akurasi dengan dan tanpa melakukan tahapan *discriminant analysis* terhadap data sebelum dilakukan metode klasifikasi. *Linear Discriminant Analysis* (LDA) adalah metode yang digunakan dalam statistik berupa pengenalan pola, dan pembelajaran mesin untuk menemukan kombinasi linier fitur atau memisahkan dua atau lebih objek atau peristiwa. Kombinasi yang dihasilkan dapat digunakan sebagai pengklasifikasi linier, atau biasa digunakan untuk reduksi dimensi sebelum klasifikasi. LDA bekerja jika pengukuran dilakukan pada variabel bebas untuk setiap pengamatan yang merupakan besaran kontinu.

Pada penelitian ini, akan dilakukan pemodelan teknik klasifikasi XGBoost dengan skema perbandingan data pelatihan terhadap data pengujian yakni 80:20, 70:30, 60:40, 50:50, 40:60. Variabel yang akan dilatih pada metode klasifikasi XGBoost antara lain :

1. Tanggal pembayaran rekening listrik pelanggan
2. Jenis pelanggan, subsidi atau non subsidi
3. Jam Nyala pelanggan (jam/bln), yang menandakan bahwa bangunan tersebut dihuni
4. Rupiah BK (rupiah), dimana merupakan biaya yang dibebankan kepada pelanggan akibat terlambat dalam melakukan pembayaran rekening listrik
5. Kali padam pelanggan (kali/plg/thn)

Pemodelan klasifikasi XGBoost pada klaster pelanggan selanjutnya akan dilakukan dengan melalui beberapa tahapan, antara lain :

1. Menggunakan fitur "XGBoost *Tree Ensemble Learner*" dan "XGBoost *Predictor*"
2. Pada fitur "XGBoost *Tree Ensemble Learner*", diinisiasi variabel target adalah "kategori" yang merupakan klasifikasi pelanggan lancar atau berpotensi menunggak
3. Variabel yang akan digunakan pada feature adalah 5 variable yang sudah dijelaskan sebelumnya

4. Pada metode klasifikasi XGBoost akan dilakukan pengamatan terhadap nilai *boosting round* sebanyak 10 kali
5. Setelah dilakukan proses simulasi model klasifikasi dengan metode XGBoost maka dapat dilihat nilai akurasi menggunakan *confusion matrix* pada fitur "*scorer (Java script)*" dimana data yang dapat dihasilkan adalah *overall accuracy, overall error, cohen's kappa, correctly classified, incorrectly classified*
6. Selain confusin matrix hasil output model klasifikasi XGBoost juga dapat diamati menggunakan fitur "ROC *curve*" yang akan menampilkan output berupa kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under Curve* (AUC).

E. Penyetelan Hyperparameter

Proses ini dilakukan dengan melakukan optimasi terhadap parameter input dari model XGBoost yang sudah dibuat. Dengan melakukan penyetelan *hyperparameter* pada menu "*booster*" maka diharapkan dapat meningkatkan hasil akurasi dari model. Penyetelan *hyperparameter* adalah proses mendefinisikan beberapa perparameter untuk membuat model optimal yang meminimalkan fungsi kerugian yang telah ditentukan pada data independen yang disediakan.

Keunggulan metode XGBoost sebagian besar berada pada efisiensinya dalam melakukan proses pembelajaran, selain itu terdapat banyak parameter yang dapat di set sehingga dapat meningkatkan kecepatan dalam proses pemodelan dan menghasilkan tingkat akurasi yang terbaik. Misalnya, dalam model berbasis pohon seperti XGBoost, parameter yang dapat dipelajari antara lain beberapa variabel keputusan di setiap simpul pohon. Selain itu pada metode klasifikasi XGBoost, *hyperparameter* juga mencakup kedalaman pohon, jumlah pohon, variabel dalam setiap pohon, pengamatan yang digunakan untuk setiap pohon, dll.

Terdapat enam (6) *hyperparameter* untuk XGBoost yang paling penting seperti yang dijelaskan pada penelitian yang dilakukan oleh Kartik [26], yang memiliki probabilitas tertinggi terhadap pembelajaran dari algoritma sehingga menghasilkan hasil paling akurat, tidak bias, cepat dan tidak terdapat *overfitting*.

Model yang diharapkan pada penelitian ini adalah model dengan tingkat akurasi terbaik, sehingga pada saat model dijalankan sebagai strategi eksekusi dilapangan maka peluang terjadinya terhadap potensi pelanggan akan tertib membayar atau menunggak akan lebih akurat sehingga akan dapat menghemat tenaga kerja serta waktu dalam penekanan piutang sehingga hasil akhir berupa percepatan pendapatan perusahaan akan dapat terealisasi. Penyetelan *hyperparameter* pada XGBoost dilakukan juga dengan melakukan pembanding yang digunakan untuk membandingkan antara penyetelan parameter pada nilai *baseline* dengan setelah melakukan *tuning*/penyetelan. Hal ini dilakukan adalah untuk mengetahui peningkatan akurasi atas pembelajaran mesin yang dilakukan dengan menggunakan metode klasifikasi XGBoost. Langkah yang dilakukan dalam menentukan parameter "booster" yang optimal antara lain dengan melakukan penyetelan parameter pada XGBoost learner seperti yang ditunjukkan pada Tabel II dibawah ini [26]:

TABEL II
PARAMETER SETTING PADA XGBOOST

No	Jenis Parameter	Range	Parameter Baseline	Parameter Optimasi	Fungsi
1	*Eta (learning rate)*	0 - 1	0.3	0 - 1	Kecepatan belajar *gradient boosting*, sehingga membuat model lebih kuat dengan mengecilkan bobot pada setiap langkah
2	*Gamma (min split loss)*	0 - ∞	0	0 – 5	Nilai gamma menentukan kerugian minimum yang diperlukan untuk membuat split
3	*Maximum depth*	0 - ∞	6	0 – 10	Jumlah kedalaman tree untuk menghindari *overfitting*
4	*Minimum child weight*	0 - ∞	1	0 – 10	Jumlah bobot minimum dari pengamatan
5	*Alpha*	0 - ∞	0	0 – 100	Digunakan pada kasus dengan dimensi tinggi sehingga algo-ritma berjalan lebih cepat
6	*Lamda*	0 - ∞	1	0 - 100	Digunakan untuk menangani *regularization* pada XGBoost

Penyetelan *hyperparameter* juga dilakukan dengan menggunakan teknik optimasi, antara lain *hillclimbing, random search dan bayesian*. Teknik *hillclimbing* merupakan teknik optimasi dengan cara menentukan langkah kedepannya yakni dengan menentukan node yang berpotensi muncul sedekat mungkin dengan target yang akan disasar [27]. Selanjutnya pada teknik *random search* yakni dilakukan dengan menentukan suatu sasaran/target yang diperoleh secara acak namun tetap layak sebagai solusi terbaik [28]. Selanjutnya proses menentukan solusi optimal dilakukan secara iteratif, dimana tiap iterasi, akan dilakukan evaluasi terhadap solusi acak disekitar solusi terbaik sementara. Sedangkan pada teknik optimasi *bayesian* digunakan model pengganti yang cocok untuk pengamatan model nyata. Optimasi *bayesian* menggunakan strategi desain secara berurutan untuk optimalisasi global terhadap

fungsi yang masih belum diketahui [29]. Tujuan utama optimasi *hyperparameter* adalah menghasilkan model optimal yang meminimalkan *loss function* yang telah ditentukan pada data independen yang timbul akibat proses pelatihan dan pengujian. Berikut adalah hubungan *hyperparameter* dan *loss function* seperti yang ditunjukkan pada persamaan (4) dan (5):

$$p* = argmin_{p \epsilon p} loss(p),$$

dimana p adalah himpunan kombinasi *hyperparameter*, p∗ adalah kombinasi parameter optimal yang diperoleh dari optimasi akhir, dan loss(p) adalah *objective function*.

$$loss(p_j) = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i(p_j) - y_i)^2}{N}}, \quad (5)$$

dimana $p_j$adalah kombinasi *hyperparameter* yang ke-j, y_i adalah nilai sebenarnya, dan $(\hat{y}_i(p_j)$adalah output model yang dihasilkan pada parameter ke j pada kombinasi *hyperparameter*. Dengan melakukan optimasi pada hyperparameter maka diharapkan mendapatkan nilai akurasi tertinggi secara otomatis tanpa harus melakukan input parameter secara manual. Pada penelitian ini jumlah *boosting rounds* diberikan batas sebanyak 10 kali. *Boosting rounds* merupakan banyaknya jumlah pelaksanaan *boosting* (jumlah pohon yang dibangun) yang digunakan untuk membentuk model atau dapat juga diartikan sebagai jumlah iterasi yang akan dilakukan. *Boosting rounds* menentukan berapa kali pohon keputusan harus dihitung berdasarkan algoritma.

F. Pengujian dan Analisis Data

Pada tahapan ini dilakukan proses analisa dan pembahasan dari proses serta output yang sudah dimodelkan. Selanjutnya akan dilakukan analisis terhadap model menggunakan nilai *confusion matrix* yang memberikan output nilai akurasi, nilai error, klasifikasi sesuai, klasifikasi tidak sesuai, serta nilai *cohen's kappa*. Koefisien *cohen's kappa* digunakan untuk mengukur tingkat kesepakatan dari 2 penilaian dalam mengklasifikasikan obyek pada grup tertentu. Rumus dari koefisien *cohen's kappa* ditunjukkan pada persamaan (6) berikut :

$$\mathrm{K} = \frac{\sum_{i=1}^{I} p_{ii} - \sum_{i=1}^{I} p_{i+}p_{+i}}{1 - \sum_{i=1}^{I} p_{i+}p_{+i}} \quad (6)$$

Dimana : $\sum_{i=1}^{I} p_{ii}$ = Total proporsi diagonal utama dari frekuensi observasi
$\sum_{i=1}^{I} p_{i+}p_{+i}$ = Total proporsi marginal dari frekuensi observasi

Kurva ROC dan Nilai *Area Under Curve* (AUC) dari model klasifikasi XGBoost memungkinkan untuk mengkonfirmasi keakuratan prediktor yang digunakan untuk membuat keputusan tentang prediktor mana yang paling akurat dan memiliki tingkat akurasi terbaik. Nilai AUC digunakan untuk mengukur bila terdapat performansi atas beberapa metode yang digunakan. Berikut rumus untuk menghitung nilai AUC : (7)

$$\theta^r = \frac{1}{mn}\sum_{j}^{n} = 1\sum_{i}^{m} = 1\psi(xi^r, xj^r)$$

Dimana,

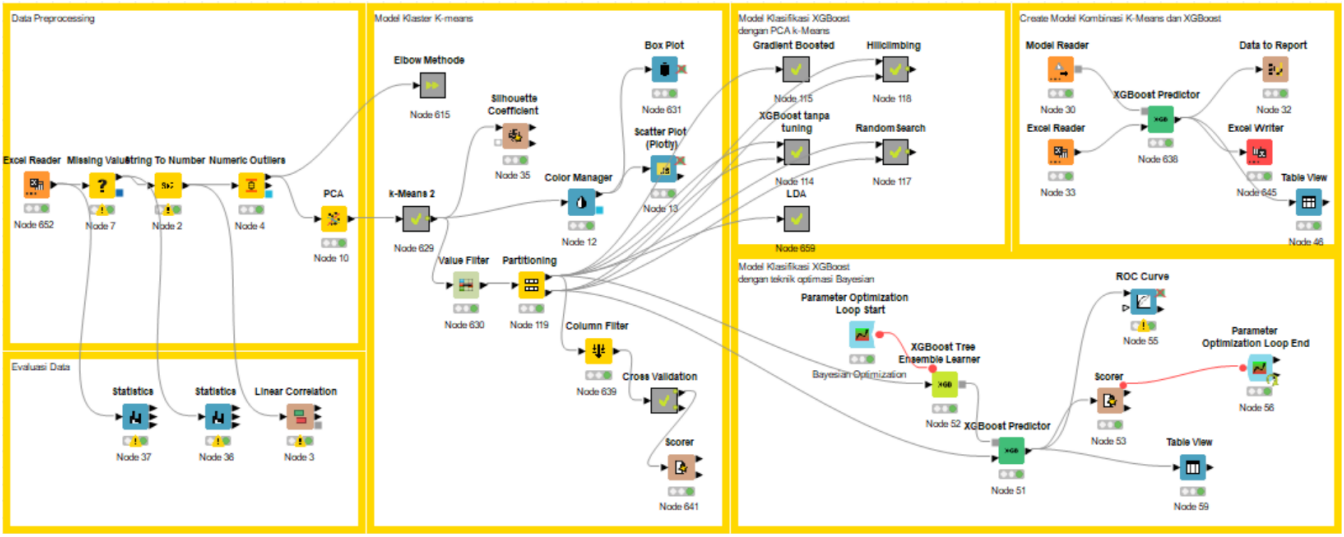
$$\psi(X,Y) = \begin{cases} 1 \; Y < X \\ \frac{1}{2} \; Y = X \\ 1 \; Y > X \end{cases},$$

X = Output Positif, Y = Output Negatif

Selain itu juga akan dilakukan proses uji menggunakan fitur "*cross validation*" untuk mengetahui parameter yang memiliki akurasi tinggi dengan melakukan inisiasi nilai *number of validations* pada jangkauan 0 – 10 fold [26].

III. UJI COBA DAN ANALISA

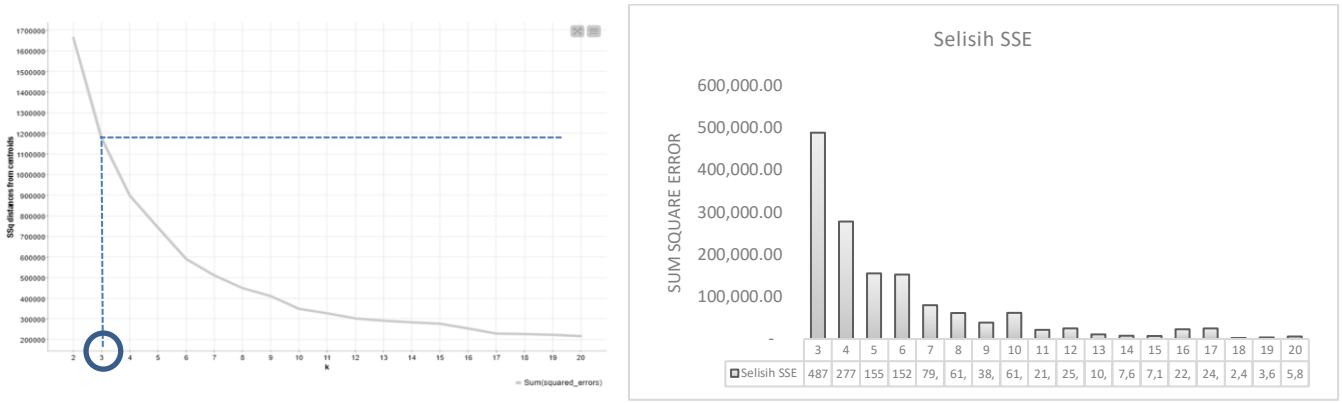
Hasil simulasi model pada penelitian ini dilakukan dengan menggunakan aplikasi KNIME. Simulasi model pembentukan klaster terlebih dahulu dilakukan dengan tujuan mendapatkan pengelompokkan pelanggan yang akan dilakukan proses lanjutan yakni klasifikasi pelanggan lancar membayar rekening listrik dan berpotensi menunggak seperti yang ditunjukkan pada Gambar 4 dibawah ini. Beberapa perbandingan teknik klasifikasi serta teknik optimasi pada penelitian ini dilakukan untuk mendapatkan hasil akurasi terbaik.



Gambar 4. Diagram Model Kombinasi Klaster k-means dan klasifikasi XGBoost pada KNIME

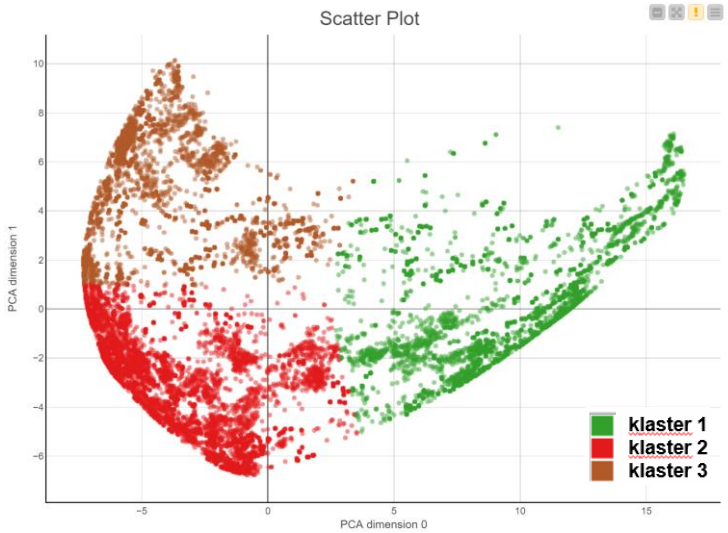
A. *Hasil Klaster Pelanggan menggunakan metode k-means*

Pada tahapan metode klaster *k-means* dilakukan dua tahapan terhadap variabel yang berbeda yakni besar nilai tagihan rekening listrik dan jarak pelanggan terhadap titik *payment point online bank* (PPOB). Pada uji coba data, didapatkan data grafik nilai k dan SSE seperti ditunjukkan pada Gambar 5 dibawah ini. Dari total uji sebanyak 20 klaster maka didapatkan nilai klaster sebanyak tiga memiliki nilai penurunan SSE terbesar sebanyak 487.991 sehingga dapat disimpulkan jumlah nilai k yang akan digunakan pada metode klaster *k-means* adalah k=3.



Gambar 5. Metode *Elbow* untuk Menentukan jumlah K

Metode klaster *k-means* dilakukan untuk mendapatkan pelanggan-pelanggan dengan karakteristik besar nilai rupiah tagihan rekening listrik terhadap jarak rumah pelanggan dengan titik *payment point online bank* (PPOB). Sehingga simulasi dilakukan adalah dengan menggunakan variabel besar nilai tagihan rekening dan data jarak rumah pelanggan ke titik PPOB, serta membandingkan penggunaan teknik PCA dan tanpa PCA. Pada penelitian ini dilakukan reduksi fitur antara lain jarak PPOB 1 sampai dengan PPOB 34 yakni sebanyak 34 fitur dikarenakan nilai korelasi antar fitur yakni dengan nilai 0,9 bahkan terdapat juga nilai korelasi 1 sehingga diperlukan ekstraksi fitur dengan Teknik PCA dikarenakan nilai korelasi sangat kuat (*multikolinieritas*) sehingga kecenderungan hasil menjadi tidak sesuai yang diharapkan. Hasil dari Teknik PCA pada penelitian ini adalah menghasilkan 2 fitur baru yakni PCA 1 dan PCA 2 yang akan digunakan dalam memetakan klaster pelanggan.



Gambar 6. Hasil Simulasi Klaster *k-means* dengan Teknik PCA

Berdasarkan scatter plot pada Gambar 6 diatas, maka dapat diamati bahwa klaster yang terbentuk telah mendekati kerapatan tertentu atau tingkat kemiripan pada klaster yang sama atas variable PCA yang digunakan yakni PCA 1 dan PCA 2. Sehingga dengan demikian klaster yang terbentuk sebanyak 3 klaster dapat dilanjutkan pada proses lanjutan yakni teknik klasifikasi XGBoost untuk mendapatkan pengelompokkan pelanggan lancar dan pelanggan menunggak rekening listrik. Hasil simulasi klaster tanpa dan dengan melakukan proses PCA didapatkan nilai *shillouette score* rata-rata 0.585 yakni lebih baik dengan dilakukan proses PCA. Pada hasil simulasi klaster dengan proses PCA seperti ditunjukkan pada Tabel III maka didapatkan klaster kedua merupakan klaster dengan populasi terbanyak yakni 10.330 pelanggan. Selanjutnya data pelanggan pada klaster ke-2 akan digunakan pada tahapan klasifikasi untuk prediksi pelanggan lancar dan pelanggan menunggak.

TABEL III
HASIL SIMULASI KLASTER K-MEANS DAN SHILLOUETTE SCORE DENGAN PROSES PCA

Klaster	Satuan	K-means		PCA + K-means		Karakteristik Pelanggan Terhadap Klaster
		Jumlah	Skor Silhouette	Jumlah	Skor Silhouette	
Klaster 1	Pelanggan	14352	0.469	8653	0.553	Pelanggan yang berpotensi lancar dalam membayar rekening listrik
Klaster 2	Pelanggan	10176	0.588	**10330**	**0.633**	Pelanggan yang berpotensi terlambat membayar rekening listrik namun belum lewat akhir bulan
Klaster 3	Pelanggan	1168	0.501	6554	0.603	Pelanggan yang berpotensi menunggak rekening listrik melewati akhir bulan

B. Hasil Klasifikasi menggunakan metode XGBoost

Pada simulasi model klasifikasi menggunakan metode XGBoost digunakan tiga uji coba proses simulasi antara data pelatihan dan data pengujian antara lain 80:20, 70:30, 60:40, 50:50, dan 40:60. Penentuan tingkat akurasi juga dilakukan dengan membandingkan metode klasifikasi *Gradient Boosting*, teknik LDA sebelum klasifikasi, XGBoost tanpa penyetelan parameter dan XGBoost dengan penyetelan parameter. Berdasarkan hasil simulasi yang ditunjukkan pada Tabel IV berikut dimana menggunakan metode klasifikasi XGBoost pada data pelatihan sebanyak 80% dan data pengujian 20% serta dengan dilakukan penyetelan parameter memiliki tingkat akurasi terbaik adalah di angka 89,27%.

TABEL IV
PERBANDINGAN METODE KLASIFIKASI DENGAN BEBERAPA SIMULASI DATA PELATIHAN DAN DATA PENGUJIAN

Metode	Data Pelatihan : Data Pengujian									
	80:20		70:30		60:40		50:50		40:60	
	Akurasi	Cohen's kappa	Akurasi	Cohen's kappa	Akurasi	Cohen's kappa	Akurasi	Cohen's kappa	Akurasi	Cohen's kappa
K-means + Gradient Boosting	74,762	0,498	73,968	0,482	75,556	0,517	73,968	0,487	73,730	0,484
K-means + LDA + XGBoost	82,070	0,619	80,655	0,584	81,047	0,599	82,107	0,622	82,179	0,626
K-means + XGBoost tanpa tuning	86,846	0,725	86,998	0,734	86,360	0,720	87,056	0,732	87,731	0,744
K-means + XGBoost dengan tuning	**89,270**	**0,680**	88,423	0,637	88,654	0,644	88,161	0,672	87,770	0,593

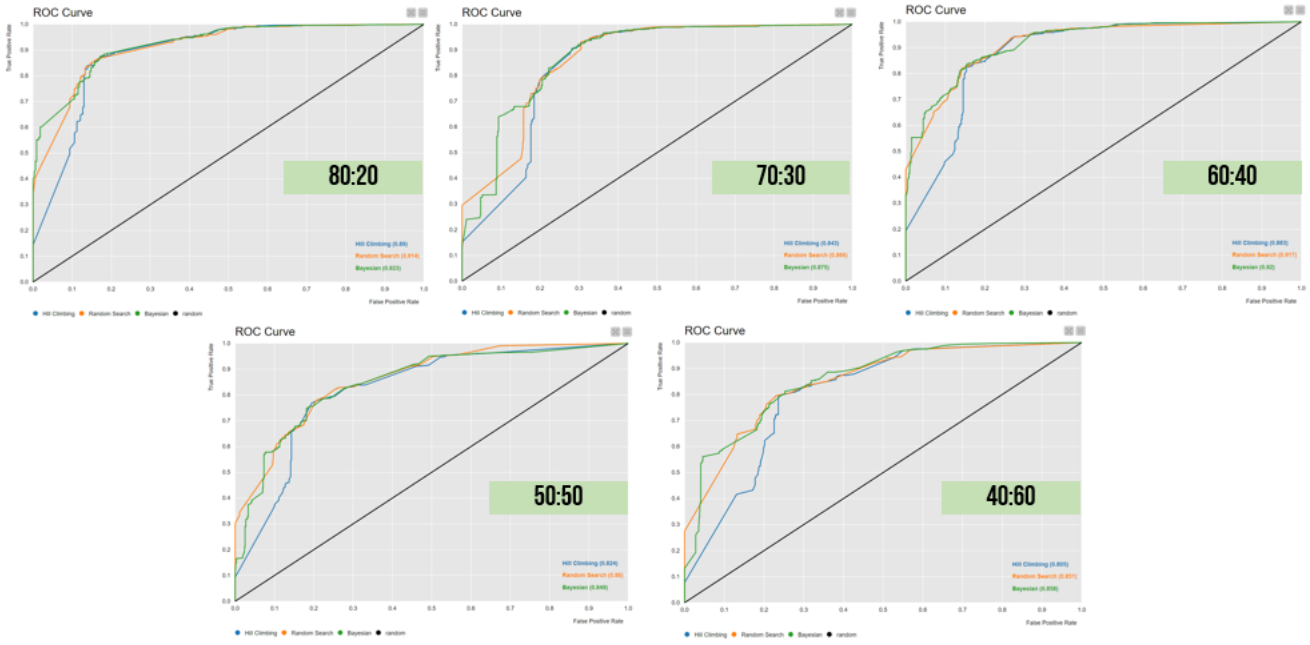
C. Optimasi Hyperparameter pada Klasifikasi XGBoost

Terdapat tiga metode optimasi yang dilakukan simulasi pada penelitian ini antara lain metode *hillclimbing, random search, dan bayesian*. Dengan melakukan pencarian parameter dengan teknik optimasi maka akan didapatkan nilai akurasi terbaik dari model klasifikasi XGBoost. Pada penelitian ini akan dilakukan 10 kali *boosting round* yang akan diamati nilai akurasi serta nilai parameter terhadap akurasi terbaik tersebut.

TABEL V
HASIL SIMULASI BEBERAPA TEKNIK OPTIMASI

Teknik Optimasi	Data Latih (%)	Boosting round Terbaik	Nilai Parameter Terbaik								Klasifikasi Sesuai	Klasifikasi Tidak Sesuai	Akurasi	Kohen's Kappa
			ETA	Gamma	Max. Depth	Min. Child Weight	Col. Sample by tree	Subsample	Alpha	Lamda				
Hill Climbing	80:20	8	0,49	3,42	7,00	7,37	0,41	0,59	39,23	38,72	4529	579	88,67	0,65
	70:30		0,74	0,27	2,00	5,60	0,57	0,65	92,11	47,52	6715	947	87,64	0,62
	60:40		0,25	0,65	5,00	2,52	0,68	0,35	28,45	15,38	9028	1187	88,38	0,63
	50:50		0,64	2,14	6,00	0,31	0,76	0,82	76,41	21,74	11329	1440	88,72	0,66
	40:60		0,20	1,62	3,00	9,28	0,86	0,32	64,14	13,16	13256	2067	86,51	0,53
Random Search	80:20	10	0,66	1,87	8,00	3,82	0,80	0,87	3,08	47,33	4545	563	88,98	0,67
	70:30		0,71	3,60	9,00	1,59	0,89	0,59	71,82	3,31	6669	993	87,04	0,58
	60:40		0,64	1,19	4,00	3,21	0,66	0,76	80,79	3,66	9022	1193	88,32	0,64
	50:50		0,88	0,91	0,00	1,98	0,54	0,86	15,75	56,81	9887	2882	77,43	0,00
	40:60		0,36	1,73	4,00	3,23	0,95	0,61	23,11	35,07	13240	2083	86,41	0,53
Bayesian	80:20	8	0,65	4,17	1,35	2,12	0,57	0,60	5,16	48,30	4560	548	**89,27**	**0,68**
	70:30		0,83	2,83	2,67	0,81	0,75	0,95	31,82	24,45	6709	953	87,56	0,61
	60:40		0,72	4,39	5,56	1,32	0,38	0,98	1,84	90,31	9104	1111	89,12	0,67
	50:50		0,75	0,71	1,97	2,54	0,79	0,57	7,34	88,69	11371	1398	89,05	0,66
	40:60		0,59	3,11	0,47	1,18	0,66	0,75	18,05	58,97	13610	1713	88,82	0,66

Pada Tabel V diatas dapat ditunjukkan bahwa nilai akurasi terbaik didapatkan dengan menggunakan teknik optimasi *bayesian* pada *boosting round* ke-8 dengan nilai 89,272% akurasi. Nilai parameter untuk menghasilkan nilai akurasi tersebut dapat juga terlihat pada tabel diatas antara lain ETA 0,652; *gamma* 4,169; *maximum depth* 1,355; *minimum child* 2,125; *col sample* 0,573; *subsample* 0,6; *alpha* 5,160; *lambda* 48,299. Berdasarkan hasil simulasi model pada Tabel V diatas, dapat juga diketahui bahwa semakin besar persentase data latih terhadap data uji maka kecenderungan akan meningkatkan nilai akurasi dari model.



Gambar 7. Kurva ROC dan Nilai AUC dari 3 teknik optimasi XGBoost

Selain melakukan pengujian dengan menggunakan teknik optimasi sebanyak 10 *boosting rounds*, maka dilakukan juga perbandingan antar teknik optimasi terbaik dengan mengamati nilai kurva ROC dan nilai AUC nya. Berdasarkan Gambar 7 diatas dapat terlihat bahwa teknik optimasi *bayesian* pada data latih 80 % dan data uji 20% memiliki nilai kurva ROC dan nilai AUC terbaik yakni pada angka 0,923 jika dibandingkan dengan teknik *hill-climbing* dan *random search*. Pada uji *cross validation* nilai skor akurasi adalah sebesar 89,63% serta didapatkan besarnya *error* terendah dengan menggunakan k sebanyak 10, yakni pada *fold* ke 9 yakni sebesar 9,696%.

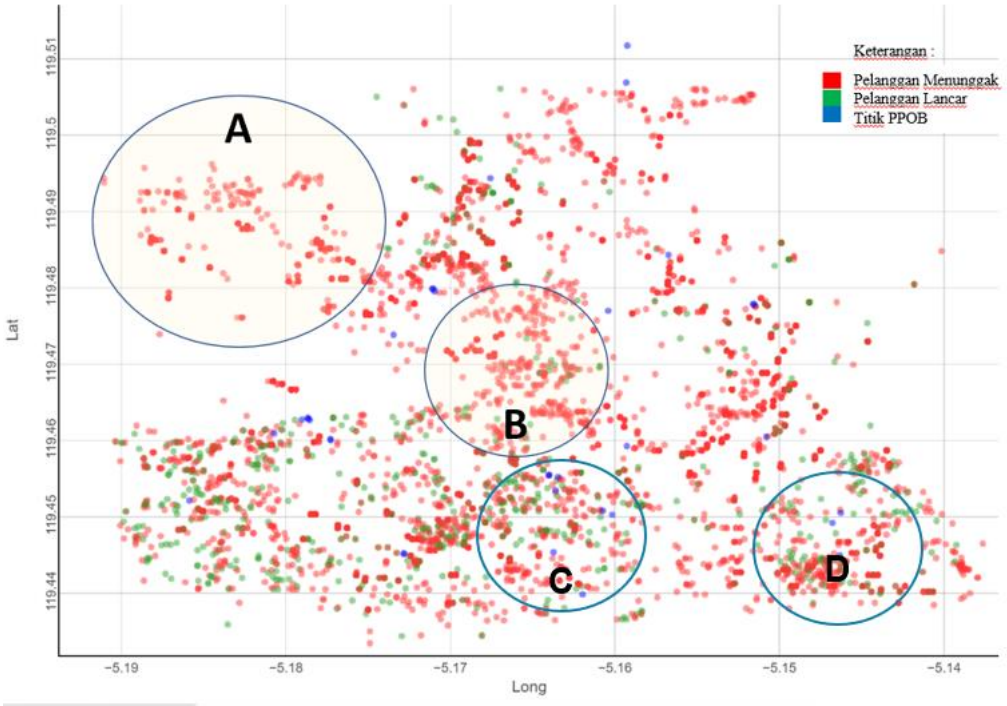
D. Pembahasan

Metode klaster dilakukan dengan tujuan untuk mendapatkan pengelompokkan pelanggan sebanyak 3 klaster sesuai dengan metode *elbow* yang telah dilakukan. Pada metode klaster, penggunaan teknik PCA pada data jarak pelanggan ke titik PPOB memberikan hasil maksimal yakni perbaikan nilai *shilouette score* menjadi 0.585. Berdasarkan hasil simulasi model klaster *k-means* maka didapatkan klaster pelanggan yang berpotensi lancar, berpotensi menunggak tidak lewat bulan dan berpotensi menunggak rekening listrik lewat akhir bulan. Selanjutnya klaster tersebut akan digunakan untuk mendapatkan model yang lebih akurat.

Keunggulan dari klasifikasi XGBoost antara lain memiliki *hyperparameter* yang dapat di atur serta dapat terbukti meningkatkan tingkat akurasi dari model. Teknik optimasi digunakan pada penelitian ini untuk mendapatkan nilai parameter terbaik dan akurasi tertinggi secara otomatis. Nilai akurasi terbukti menjadi lebih baik dengan melakukan perubahan parameter antara lain *ETA, gamma, maximum depth, minimum child weight, col. Sample by tree, subsample, alpha dan lamda*. Jika dibandingkan antara ketiga teknik optimasi yang digunakan maka optimasi *bayesian* memiliki nilai *Area Under Curve* (AUC) serta kurva ROC terbaik yakni pada angka 0.923. Sehingga dapat disimpulkan untuk mendapatkan nilai terbaik pada metode klasifikasi XGBoost dengan melakukan optimasi parameter dapat dilakukan dengan teknik *bayesian*.

Berdasarkan hasil klasifikasi potensi pelanggan lancar dan potensi pelanggan menunggak menggunakan metode XGBoost bahwa terdapat 4573 pelanggan yang prediksi nya sesuai dan 535 pelanggan yang prediksinya tidak sesuai. Dari hasil tersebut jika dipetakan maka terdapat 349 pelanggan menunggak yang prediksinya tidak sesuai dari total 4116 pelanggan menunggak atau setara dengan 8,48%. Sedangkan terdapat juga 186 pelanggan lancar yang prediksinya tidak sesuai dari total 992 pelanggan lancar atau setara dengan 18,75%.

Berdasarkan hasil simulasi model juga dapat diketahui potensi menunggak pelanggan kategori subsidi dan non subsidi, dimana sebesar 2156 pelanggan non subsidi atau setara 70% berpotensi menunggak pembayaran rekening listrik. Sedangkan untuk pelanggan kategori subsidi yakni terdapat sebanyak 1960 pelanggan atau setara 96% pelanggan berpotensi menunggak. Hal ini dapat disimpulkan bahwa potensi pelanggan menunggak pembayaran rekening listrik adalah dari kategori pelanggan subsidi dimana berkaitan erat dengan kondisi keuangan pelanggan.



Gambar 8. Potensi Penempatan Titik PPOB Optimal

Jika melihat peta sebaran pelanggan menunggak, pelanggan lancar, dan titip PPOB yang saat ini terdapat di wilayah kerja PLN Panakkukang seperti yang ditunjukkan pada Gambar 8 diatas maka dapat ditentukan lokasi penempatan titik PPOB baru yang paling optimal untuk menekan angka piutang. Alternatif lokasi pada point A yang ditunjukkan pada Gambar 8 diatas merupakan daerah yang cukup banyak pelanggan menunggak namun

belum tersedia titik PPOB. Kemudian lokasi pada point B juga dapat dijadikan peluang untuk peletakan titik PPOB baru sehubungan daerah yang cukup padat pelanggan menunggak namun belum tersedia titik PPOB yang memadai. Titik C dan titik D merupakan daerah yang telah tersedia cukup banyak PPOB, sehingga dapat mensosialisasikan pembayaran rekening listrik tepat waktu pada PPOB yang telah tersedia di daerah-daerah tersebut. Untuk daerah lainnya dapat digunakan metode penagihan door to door serta mensosialisasikan pembayaran rekening listrik melalui mobile banking. Dengan melihat peta sebaran berupa geolokasi seperti yang ditunjukkan pada Gambar 8 diatas maka dapat dijadikan strategi eksekusi penagihan yang dilakukan oleh mitra billing management dalam bekerja. Selain itu lokasi daerah pelaksanaan kegiatan sosialisasi tertib membayar listrik serta pengalokasian tenaga SDM tambahan untuk melakukan kegiatan pendukung lainnya juga akan menjadi tepat sasaran sehingga dapat efisiensi disisi anggaran dan efektif dalam pelaksanaan kegiatan.

IV. KESIMPULAN DAN SARAN

Pembentukan model karakteristik pelanggan melalui kombinasi antara metode klaster *k-means* dengan metode klasifikasi XGBoost terbukti dapat dilaksanakan dengan tingkat akurasi hasil 89,27% dan nilai AUC 0,923. Hal ini dapat dilaksanakan dengan melakukan tahapan-tahapan pra-pemrosesan data, kemudian membentuk klaster untuk mendapatkan pengelompokkan pelanggan dengan populasi tertinggi dan kemudian digunakan pada tahapan pengklasifikasian pelanggan untuk mendapat prediksi pelanggan lancar dan pelanggan menunggak. Metode klasifikasi XGBoost merupakan metode klasifikasi yang lebih unggul jika dibandingkan dengan metode klasifikasi lainnya dikarenakan memiliki penyetelan *hyperparameter* yang terbukti dapat meningkatkan akurasi dari model, dimana teknik optimasi *bayesian* berdasarkan kurva ROC dan nilai AUC memiliki tingkat akurasi hasil tertinggi jika dibandingkan dengan teknik *hillclimbing* dan teknik *random search*. Berdasarkan hasil simulasi model pada data pelanggan PLN ULP Panakkukang maka didapatkan pemetaan pelanggan lancar dan pelanggan menunggak dimana didapatkan 2 lokasi yang ditentukan sebagai titik optimal pembangunan *Payment Point Online Bank* (PPOB). Dari hasil simulasi pada data PLN ULP Panakkukang maka dapat dianalisis bahwa pelanggan subsidi dan pelanggan yang sering mengalami pemadaman aliran listrik memiliki kecenderungan untuk berpotensi menunggak. Hal ini dapat menjadikan pertimbangan tim PLN untuk peningkatan layanan listrik dan melakukan sosialisasi kepada pelanggan tersebut.

Untuk penelitian selanjutnya peningkatan akurasi hasil klasifikasi XGBoost dapat dilakukan dengan melakukan simulasi optimasi dengan metode lainnya serta dengan jumlah *round boosting* yang lebih banyak. Selain itu perlu dilakukan pengujian berkaitan dengan beberapa data baru pada unit perusahaan listrik sejenis sehingga dapat diketahui seberapa mampu dan akurat model dalam melakukan pengelompokkan pelanggan lancar dan menunggak. Pada penelitian selanjutnya juga dapat dilakukan model kombinasi dengan menggunakan teknik lain seperti *hierarchical*, kNN, dan DBSCAN yang dapat digabungkan dengan metode klasifikasi dengan teknik *random forest*, SVM, *naïve bayes*, dll.

DAFTAR PUSTAKA

[1] A. Darmawan, dan S.P. Bangun. (2016). Electricity Accounts Receivables Billing Procedures. *Journal of Applied Accounting and Taxation. 1(1),* hal. 15-20.

[2] W. Guo, W. Hong, W. Li, dan K. Guo. (2015). Design and Implementation of Electric Charge Arrears Prediction System. *12th Web Information System and Application Conference (WISA),* hal. 309-313.

[3] E. A. Darko, S. Adarkwah, F. Donkor, dan E. Kyei. (2016). Management of accounts receivables in utility companies: A focus on Electricity Company of Ghana (ECG). *International Journal of Academic Research in Business and Social Sciences*. *6*, hal. 486-518.

[4] W. Fu, D. Zhang, Y. Fu, J. Li, dan Y. Xie. (2017). Arrears prediction for electricity customer through Wgan-Gp, *IEEE*. hal. 1667-1670.

[5] M. Bahrami, B. Bozkaya, dan S. Balcisoy. (2020). Using Behavioral Analytics to Predict Customer Invoice Payment. *Big data. 8(1),* hal. 25-37.

[6] A. P. Redaputri, dan I. Apriansyah. (2022). Strategi Pengambilan Keputusan Untuk Meminimalkan Tunggakan Tagihan Listrik Pasca Bayar PT. PLN. *JBMI (Jurnal Bisnis, Manajemen, dan Informatika). 19(1),* hal. 20-33.

[7] S. Zeng, P. Melville, C. Lang, I. Boier, dan C. Murphy. (Agustus 2008). Using predictive analysis to improve invoice-to-cash collection. *International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA.* hal. 1043-1050.

[8] W. Hu. (2016). "Overdue invoice forecasting and data mining,". Massachusetts Institute of Technology. Graduate Thesis.

[9] A. Appel, G. Malfatti, R. Cunha, B. Cardoso and R. de Paula. (Agustus 2020). Predicting Account Receivables with Machine Learning. *KDD (Virtual Conference) MLF '20, San Diego, CA.*

[10] I. Indrayani. (2019). Pengaruh Payment Point Online Bank (PPOB) Terhadap Percepatan Aliran Kas (Studi Kasus di PT. PLN (PERSERO) Provinsi Aceh). *Jurnal Akuntansi dan Keuangan Universitas Malikussaleh. 7(2),* hal. 83-92.

[11] S. Murtaqi. (2011). "Perubahan Sistem Siklis Menjadi Non Siklis,". Peraturan Direksi PT PLN (Persero), Jakarta.

[12] Y. Bambang. (2019). "Pengendalian Piutang," Edaran GM PLN UIW Sulselrabar, Makassar.

[13] R. Nurul dan K. Edi. (2020). Implementasi Metode K-Means Clustering Tunggakan Rekening Listrik pada PT. PLN (Persero) Gardu Induk Kisaran. *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD. 3(1),* hal. 103-117.

[14] D.N Batubara, A.P. Windarto dan E. Irawan. (Februari 2022)Analisis Prediksi Keterlambatan Pembayaran Listrik Menggunakan Komparasi Metode Klasifikasi Decision Tree dan Support Vector Machine. *Jurnal Riset Komputer. 9(1),* hal. 102-108.

[15] Y. Asri, D. Kuswardani dan E. Yosrita. (2021). Clusterization of customer energy usage to detect power shrinkage in an effort to increase the efficiency of electric energy consumption. *Indonesian Journal of Electrical Engineering and Computer Science. 22(1),* hal. 10-17.

[16] S. Shah. (Januari 2019). Customer Payment Prediction in Account Receivable. *International Journal of Science and Research (IJSR). 8(1),* hal. 642-644

[17] P. Tang. (2020). Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm. *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE),* hal. 1128-1131.

[18] D. Ran, H. Jiaxin dan H. Yuzhe. (2020). Application of a Combined Model based on K-means++ and XGBoost in Traffic Congestion Prediction. *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA),* hal. 413-418.

[19] J. Henriques, F. Caldeira, T. Cruz dan P. Simoes. (2020). Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. *Electronics. 9(7),* hal. 1-16.

[20] Z. Mushtaq, S. Ashraf dan N. Sabahat. (2020). Predicting MBTI Personality type with K-means Clustering and Gradient Boosting. *2020 IEEE 23rd International Multitopic Conference (INMIC),* hal. 1-5.

[21] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio dan L. D. F. Costa. (2021). Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Comput. Surv. 54(4),* hal. 1-34.

[22] L. Ye, C. Qiu-ru, X. Hai-xu, L. Yi-jun dan Y. Zhi-min. (2012). Telecom customer segmentation with K-means clustering. *7th International Conference on Computer Science & Education (ICCSE).* hal. 648-651.

[23] T. Chen dan C. Guestrin. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 8,* hal. 785-794.

[24] C. Wang, C. Deng dan S. Wang. (2020). Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Elsevier. 136,* hal. 190-197.

[25] J. Gao, W. Sun dan X. Sui. (2021). Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. *A. Farouk, Ed., Discrete Dynamics in Nature and Society. 2021,* hal. 5080472.

[26] K. Budholiya, S. K. Shrivastava dan V. Sharma. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences. 34,* hal. 4514-4523.

[27] K. Nagarajan. (2018). A Predictive Hill Climbing Algorithm for Real Valued multi-Variable Optimization Problem like PID Tuning. *International Journal of Machine Learning and Computing. 8(1),* hal.14-19.

[28] S.V. Konstantinov, A.I. Diveev, G.I. Balandina dan A.A. Baryshnikov. (2018). Evolutionary Algorithms for the Optimal Control Problem of the Mobile Robot. *13th International Symposium "Intelligent Systems". 1514,* hal. 123-136.

[29] V. H. Nguyen, T. T. Le, H. S. Truong, M. V. Le, V. L. Ngo, A. T. Nguyen dan H. Q. Nguyen. (2021). Applying Bayesian Optimization for Machine Learning Models in Predicting the Surface Roughness in Single-Point Diamond Turning Polycarbonate. *Hindawi. 2021,* hal. 1-16.