2. LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 *Customer churn*

Customer churn adalah istilah untuk pelanggan yang berhenti menggunakan produk atau jasa suatu perusahaan, dapat dikarenakan pelanggan berpindah ke pesaing yang lain atau alasan lainnya. *Customer* churn sering diasosiasikan dengan *customer relationship* dan kepuasan pelanggan. Semakin puas seorang pelanggan terhadap layanan atau produk maka kemungkinan churn akan semakin kecil. (Jahromi et al., 2010 ; Chandar et al., 2006). Namun tidak menutup kemungkinan terdapat faktor - faktor lain di dalam perusahaan yang menyebabkan churn. Faktor ini tidak bisa dipatenkan untuk setiap bidang usaha, namun harus disesuaikan lagi dengan tipe perusahaan dan jenis data yang ada.

Customer churn menjadi salah satu masalah yang dialami perusahaan di bidang apapun karena kehilangan *customer* tentunya menyebabkan kerugian. Untuk mencegah turunnya pendapatan perusahaan perlu untuk mencari *customer* baru sehingga dapat menggantikan pelanggan yang hilang itu. Mencari *customer* baru tentunya mengeluarkan biaya dan tenaga yang lebih bila dibandingkan dengan biaya dan tenaga untuk mempertahankan pelanggan. Nyatanya perusahaan membutuhkan 15 kali biaya untuk mencari *customer* baru daripada biaya mempertahankan *customer* (Amin, *et al*., 2017). Maka dari itu *Customer Relationship Management* mengutamakan pencegahan *customer* churn ini.

2.1.2 Prediksi *Customer Churn* untuk Mempertahankan Pelanggan

Prediksi adalah kegiatan memperkirakan kemungkinan terjadinya sesuatu secara sistematis. Dalam menjalankan proses ini terdapat kemungkinan bahwa hasilnya salah, sehingga untuk meminimalisir kesalahan, prediksi dilakukan berdasarkan bantuan informasi - informasi yang lalu. Prediksi tidak harus memberikan jawaban secara pasti kejadian yang akan terjadi, melainkan berusaha untuk mencari jawaban sedekat mungkin yang akan terjadi (Narassati, 2022).

Prediksi *customer churn* adalah proses memperkirakan kemungkinan terjadinya perpindahan *customer* atau *churn* dengan informasi yang ada. Informasi yang digunakan untuk melakukan prediksi ini adalah tingkah laku dari *customer* perusahaan selama ini. (Staircase, 2023). Prediksi ini sangat berguna bagi perusahaan terutama pada sisi *customer relationship* agar dapat mengetahui pelanggan mana yang memiliki potensi untuk pindah. Dengan begitu perusahaan dapat membuat keputusan yang dapat membuat pelanggan tersebut bertahansur.

Prediksi *customer churn* biasanya dilakukan dengan bantuan metode machine learning, dan sudah banyak dilakukan sebelumnya dengan berbagai metode serta dengan objek yang beragam. Dalam 5 tahun terakhir juga peneliti berusaha untuk meningkatkan akurasi dari model yang dibuat. Peningkatan akurasi ini dilakukan dengan menyeleksi atribut, melakukan data preprocessing, maupun menambahkan data. Hasil dari prediksi ini beragam, meskipun menggunakan metode yang sama tetapi bila data yang diolah berbeda akan menghasilkan prediksi dengan keakuratan berbeda (Kim & Lee, 2022). Maka dapat disimpulkan dalam prediksi *customer churn* ini tidak ada satu metode dan cara yang mutlak paling baik karena semuanya perlu dicocokan lagi dengan data yang diolah.

Mempertahankan pelanggan merupakan salah satu tantangan yang dialami oleh banyak perusahaan. Semakin hari semakin banyak kompetitor yang muncul dan hal itu menimbulkan potensi berpindahnya pelanggan. Maka dari itu perusahaan perlu melakukan usaha untuk mempertahankan pelanggan agar pendapatan perusahaan tidak menurun. Usaha mempertahankan pelanggan dapat didukung dengan prediksi *customer churn* karena hasil prediksi ini adalah pelanggan - pelanggan mana yang berpotensi untuk pindah. Informasi inilah yang menjadi kunci dalam melakukan usaha mempertahankan, karena perusahaan dapat lebih fokus pada pelanggan tersebut dan hasilnya lebih maksimal. Usaha yang dapat dilakukan dapat berupa meningkatkan *customer relationship management*. (Moscato et al., 2019)

2.1.3 PT X

PT X adalah perusahaan yang bergerak di bidang data center sejak 12 tahun yang lalu. Fasilitas yang ditawarkan oleh perusahaan ini adalah *colocation* dan *cable management*. *Colocation* sendiri memiliki arti sebagai layanan penitipan server pada pihak ketiga. Server dari perusahaan lain akan diletakkan di rak yang berada di ruangan khusus. Perusahaan dapat menyewa rak sesuai dengan

paket yang dipilih. Data center ini menerapkan sistem dimana pelanggan melakukan transaksinya secara tahunan.

PT X menjadi objek dari penelitian prediksi *customer churn* ini sehingga data yang digunakan adalah data *customer* dan data transaksi dari perusahaan. Data yang akan digunakan untuk penelitian adalah data dari tahun 2017 - 2023. Atribut yang digunakan adalah sebagai berikut :

a. Customer id
b. Jenis *customer*

   Customer dari PT X adalah perusahaan, jadi yang dimaksud dengan jenis *customer* adalah apakah perusahaan tersebut dikategorikan sebagai skala besar atau kecil. Data ini didapatkan dari data *customer* yang diisi saat kontrak di awal.

c. Jangka waktu menjadi *customer*

   Data ini didapatkan dengan menjumlahkan jumlah tahun *customer* telah menggunakan jasa PT X

d. Total belanja

   Data total belanja didapatkan dengan menjumlahkan semua total transaksi yang pernah dilakukan oleh *customer*

e. Jumlah fasilitas

   Jumlah fasilitas maksudnya adalah total fasilitas yang pernah digunakan oleh *customer*. Data ini didapatkan dari pencatatan transaksi yang dilakukan oleh PT X

f. Jumlah komplain

   Komplain yang dilakukan oleh *customer* akan tercatat dalam sistem PT X sehingga data ini didapatkan dengan menjumlahkan komplain yang pernah dilakukan oleh *customer*

g. Hasil Survey Customer

   PT X melakukan survey kepuasan pelanggan per tahunnya terhadap tiap pelanggannya.

h. Label

Label adalah keterangan bahwa *customer* tersebut telah berhenti menggunakan jasa PT X karena berpindah ke kompetitor. Data ini didapatkan dari survey yang dilakukan PT X saat pelanggan memilih tidak melanjutkan kontrak.

2.1.4 Naive Bayes

Naive bayes adalah salah satu metode data mining yang bersifat klasifikasi. Metode yang diciptakan oleh Thomas Bayes ini biasa digunakan untuk menghitung probabilitas atau peluang dari suatu kejadian didasarkan dari kejadian - kejadian sebelumnya. Ciri khas dari metode Naive bayes adalah asumsi yang kuat terhadap independensi dari masing - masing kondisi (Cmcbinus, 2021). Asumsi inilah yang mengakibatkan Naive Bayes memiliki kelebihan yaitu tidak memerlukan jumlah data yang banyak untuk pelatihan. Dengan menggunakan atribut independen, menentukan klasifikasi dilakukan hanya varians dari atribut dalam kelas tanpa perlu keseluruhan matriks kovarians. Kelebihan lain yang dimiliki Naive Bayes adalah perhitungannya yang cepat dan efisien, serta mudah dipahami. Naive bayes juga memiliki kekurangan yaitu akurasi dapat berkurang karena adanya asumsi atribut yang independen sedangkan bisa saja terdapat korelasi antara atribut - atribut tersebut (*What Are Naive Bayes Classifiers? | IBM*, n.d.).

Rumus yang digunakan Naive Bayes adalah :

$$P(Y|X) = \frac{P\ (X \mid Y\ )\ P(Y)}{P(X)}$$

Persamaan 2.1

Keterangan :

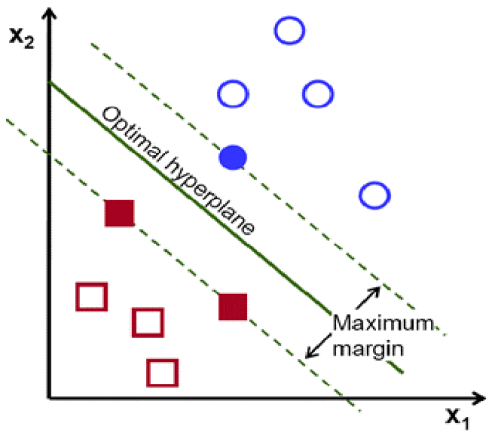
a. P(Y|X) adalah probabilitas posterior kelas (Y,target) yang diberikan prediktor (X,atribut)
b. P(Y) adalah probabilitas kelas sebelumnya
c. P(X|Y) adalah kemungkinan yang merupakan probabilitas dari kelas prediktor tertentu
d. P(X) adalah probabilitas prior dari prediktor.

Naive Bayes memiliki 3 macam algoritma umum yaitu Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes. Gaussian Naive Bayes adalah metode Naive Bayes yang menggunakan distribusi gaussian atau distribusi normal. Sedangkan Bernoulli Naive Bayes adalah metode yang digunakan untuk data binary. Metode Multinomial naive bayes adalah metode yang digunakan bila data yang diolah bersifat diskrit (Ha, 2023).

Dalam proses klasifikasi menggunakan Naive Bayes terdapat kemungkinan probabilitas yang dihasilkan adalah nol. Hasil ini dapat membuat proses klasifikasi error. Maka dari itu untuk menghindari terjadinya hal ini maka dilakukan Laplace Smoothing. Metode ini dilakukan dengan menambahkan nilai positif terkecil (Noto., 2022).

2.1.5 Support Vector Machine

Support Vector Machine atau yang biasa disingkat SVM adalah model data mining yang diawasi dan menggunakan algoritma klasifikasi untuk memisahkan data menjadi dua grup. Dalam algoritma SVM ini, tiap data di plot sebagai titik dalam ruang n-dimensi (n adalah jumlah fitur). Setelah itu, dilakukan klasifikasi dengan mencari hyperplane yang optimal. Hyperplane adalah fungsi yang digunakan untuk pemisah kelas. Hyperplane yang optimal adalah hyperplane yang bisa membagi secara jelas kedua kelas tersebut dan memiliki margin / jarak maksimum antara titik data dari kedua kelas.(Samsudiney, 2019).



Gambar 2.1 Ilustrasi SVM

Jarak margin yang maksimal memberikan beberapa penguatan sehingga titik data uji dapat diklasifikasikan dengan baik. Istilah support vector adalah data terluar yang paling dekat dengan

hyperplane. Object support vector adalah yang paling sulit diklasifikasikan karena posisinya yang dekat dengan kelas lain. Support vector inilah yang menjadi acuan mencari hyperplane paling optimal.

Prinsip dasar dari SVM adalah linear classifier, namun bisa dikembangkan lagi agar dapat menyelesaikan masalah yang non linear dengan konsep kernel. Konsep kernel mengubah bidang hyperplane pemisah non linear di ruang fitur asli menjadi hyperplane linear pemisah dalam dimensi yang lebih tinggi (Aini, 2023). Teknik kernel memiliki 3 jenis yaitu Kernel Linear, Kernel RBF, dan Kernel Polynomial.

Kernel Linear adalah jenis fungsi yang paling sederhana dan digunakan saat memiliki data yang telah terpisah secara linear. Fungsi ini paling cocok digunakan ketika data memiliki banyak fitur dikarenakan pemetaan ke ruang dimensi yang lebih tinggi tidak memberikan kenaikan kinerja yang signifikan (Ningrum, 2018). Kernel RBF adalah kernel yang digunakan bila data yang dimiliki memiliki distribusi normal atau gaussian. Kernel Polynomial adalah adalah metode yang dapat digunakan apabila data tidak terpisah secara linear.

Metode SVM memiliki 2 parameter yang mempengaruhi model dan hasil prediksi. Parameter yang pertama adalah C atau Cost, yang berfungsi untuk untuk mengoptimalkan hasil prediksi agar tidak banyak terjadi klasifikasi yang salah. Bila nilai C terlalu besar maka akan meminimalisir kesalahan dalam pelatihan dan menghasilkan margin yang lebih kecil. Margin yang kecil dapat meningkatkan resiko overfitting, yang berarti model terlalu sesuai dengan data training dan tidak meninggalkan ruang untuk data yang baru sehingga hasil prediksi pada data baru kurang maksimal. Namun sebaliknya, nilai C yang kecil mengakibatkan margin yang lebih besar. Hal ini mengakibatkan meningkatnya misklasifikasi yang akan terjadi karena margin semakin besar. Margin yang besar mengakibatkan underfitting dimana model gagal mengklasifikasi pola data (Science & Science, 2023). Maka dari itu penting untuk menentukan nilai C yang optimal agar hasil prediksi juga optimal.

Parameter lainnya adalah gamma yang berfungsi untuk mendefinisikan sejauh mana pengaruh suatu contoh pelatihan dapat dicapai, dengan nilai rendah berarti ‘jauh’ dan nilai tinggi berarti ‘dekat’. Nilai gamma tinggi menunjukan bahwa titik yang berada di sekitar hyperplane yang akan diperhitungkan. Sedangkan nilai gamma rendah berarti titik yang jauh juga diperhitungkan. Nilai gamma juga mempengaruhi hasil dari metode SVM kernel, sehingga harus ditentukan nilai yang tepat untuk mendapatkan hasil yang sesuai (Al-Mejibli, 2020).

2.2 Tinjauan Studi

2.2.1 *A Support Vector Machine Approach for Churn Prediction in Telecom Industry* (Rodan et al., 2014)

Algoritma Support Vector machine diterapkan untuk memprediksi *churn* pada pelanggan perusahaan yang bergerak pada bidang telekomunikasi. Sebanyak 5000 data dianalisis dan sebanyak 11 atribut digunakan untuk memprediksi *churn*. Atribut tersebut adalah 3G, total consumption, calling fees, local sms fees, international sms fees, international calling fees, local sms count, international MOU (Minutes of Use), total MOU, on net MOU,dan *churn*. Model SVM dibuat dengan metode RBF Kernel. Setelah dilakukan penelitian, hasilnya dibandingkan dengan metode - metode lain seperti c4.5, KNN, dan lain - lain. Hasilnya menunjukan bahwa akurasi SVM terbaik yaitu 98,7% dan *customer churn* adalah 94,3%, sedangkan untuk hit rate dan lift metode c4.5 yang menghasilkan skor lebih tinggi (Rodan et al., 2014).

Penelitian ini dilaksanakan sekitar 9 tahun yang lalu, sehingga banyak perubahan dan library yang dapat digunakan untuk memperbarui penelitian ini agar lebih akurat. Dalam penelitian tersebut, tidak dilakukan seleksi atribut yang berpengaruh terhadap *churn* sehingga atribut tersebut dapat mengganggu hasilnya prediksi. Perbedaan yang akan dilakukan pada skripsi ini adalah menggunakan seleksi atribut yang berpengaruh dan juga menggunakan metode kernel yang berbeda yaitu fungsi kernel linear.

2.2.2 Implementasi Data Mining untuk Memprediksi *Customer Churn* Menggunakan Algoritma Naive Bayes (Novendri, R.,2021)

Toko yang berbasis pada e - commerce juga memiliki permasalahan mengenai *customer churn*. Hal ini diteliti dengan metode Naive Bayes. Dalam penelitian, *customer* segmentasi dilakukan menggunakan RFM (Recency, Frequency, Monetary) dan menghasilkan 10 kategori yaitu *champions, loyal customers, potential loyalist, recent customers, promising, customer needing attention, about to sleep, at risk, can't lose them*, dan hibernating. Data yang digunakan adalah data perusahaan yang berjumlah 32000, yang kemudian dibagi menjadi data training dan data testing. Pembagian ratio data testing dan data training beragam untuk mengetes kombinasi ratio mana yang menghasilkan nilai tertinggi. Hasil prediksi berbagai ratio memiliki nilai terbaik yaitu akurasi 97,27%, precision 100% dan recall sebesar 96,98%. Dari penelitian ini dapat dilihat bahwa metode Naive Bayes berhasil menghasilkan nilai yang cukup tinggi. (Novendri, R.,,2021)

Model Naive Bayes diimplementasikan terhadap data yang berjumlah besar dan menghasilkan hasil yang bagus. Maka patut dicoba dilakukan pengujian terhadap data yang jauh lebih kecil dan tidak dilakukan segmentasi. Jurnal ini juga mengatakan bahwa metode ini memiliki kelebihan tidak memerlukan dataset yang besar untuk melatih, maka pada skripsi ini dibandingkan dengan metode SVM untuk mencari tahu mana metode yang berhasil dengan baik memprediksi *customer churn* dengan data yang kecil.

2.2.3 Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes (Yulianto & Firmansyah, 2021)

Naive Bayes juga diterapkan untuk memprediksi customer churn pada bisnis retail. Di dalam penelitian ini data diambil dari 10 toko yang random, dan data yang digunakan adalah data customer yang memiliki kartu berlangganan atau membership. Total data yang diolah adalah 258 dengan 6 atribut sebagai berikut : jenis kelamin, grade point, kepemilikan kartu kredit, rentang usia, nilai rata-rata transaksi, lama member dan class churn. Data tersebut kemudian dibagi menjadi 2 yaitu data training sebanyak 90% dan data testing sebanyak 10 %. Hasil uji menggunakan confusion matrix adalah akurasi mencapai 80% dan precision mencapai 100%.(Yulianto & Firmansyah, 2021)

Penelitian ini menunjukan bahwa dengan data yang tergolong kecil pun Naive bayes masih dapat menghasilkan akurasi yang cukup tinggi. Namun memang jumlah data yang lebih banyak bisa dapat meningkatkan akurasi dari penelitian ini. Selain itu, juga perlu diterapkan atribut selection untuk memastikan bahwa atribut yang digunakan tidak berkorelasi tinggi satu sama lain yang dapat berpengaruh pada prediksi.