Penerapan Data Mining Untuk Memprediksi Prestasi Siswa SMA Pada Dinas Pendidikan Provinsi Jambi

Widya Lastari[1], Jasmir[2]

Pascasarjana, Magister Sistem Informasi, Universitas Dinamika Bangsa, Jambi
Jl. Jend. Sudirman Thehok-Jambi Telp: 0741-35096 Fax : 35093
Email: widialestari8800@gmail.com[1], ijay_jasmir@yahoo.com[2]

Abstract

The implementation of education is one of the important efforts in improving the quality of students. With good education it will be useful in realizing the goals of students. This study utilizes data mining techniques using yahoo K-Nearest Neighbor (K-NN) to predict student achievement. The attributes used in this study are scores: Indonesian, Mathematics, English, Biology, Chemistry, Physics, Sociology, Economics, Geography and the target is the study of students. From the results of the study, the best results were at K = 3, the use of python sklean data mining got an accuracy value 61.9% and error MSE 0.8 by comparison Naive Baiyes with accuracy 58% and error MSE 0.41 , and for Rapid miners using KNN got value accucary51%.

Keywords: Data Mining, Classification, Predicition, K-Nearest Neighbor

Abstrak

Penyelenggaran pendidikan adalah salatu satu upaya penting dalam meningkatkan kualitas siswa. Dengan pendidikan yang baik maka bermanfaat dalam mewujudkan tujuan dari siswa. Penelitian ini memanfaatkan teknik data mining menggunakan algoritma K-Nearest Neighbor (K-NN) untuk memprediksi prestasi siswa. Attibut yang digunakan dalam penelitian ini adalah nilai 6 Mata Pelajaran dari tiap jurusan yang dimiliki: Bahasa Indonesia, Matematika, Bahasa Inggris, Biologi, Kimia, Fisika, Sosiologi, Ekonomi, Geografi. Target dari penelitian ini adalah keberlanjutan studi dari siswa. Dari hasil Penelitian diperoleh hasil terbaik pada K=3, Penggunaan python sklean data minning mendapatkan nilai akurasi 61.9% dan nilai error MSE 0.38 dengan pembanding menggunakan Naïve Baiyes dengan nilai akurasi 58% dan error MSE 0.41 dan untuk Rapid miner dengan menggunakan method KNN mendapatkan nilai sebesar 51%.

Kata kunci: Data Mining, Klasifikasi, Prediksi, K-Nearest Neighbor

1. Pendahuluan

Penggunaan *data mining* yang sudah terbukti mendapat banyak informasi baru. Teknik data mining digunakan dibanyak bidang, seperti : pendidikan, pemasaran, teknik, keuangan, olahraga dan obat-obatan [1]. Baru-baru ini informasi pada data mining dan sistem pendidikan mulai meningkat, hal ini membuat data mining pendidikan menjadi hal baru untuk dilakukan penelitian. Metodologi komputasi yang digunakan adalah EDM (*Educational Data Mining*) untuk mengevaluasi data siswa. Selain itu, EDM adalah sistem yang dikembangkan yang berkonsentrasi pada pengembangan metode untuk menggabungkan sejumlah besar data yang menghasilkan domain pendidikan. Metode ini membantu memahami bagaimana siswa belajar dengan memahami perilaku mereka.

EDM digunakan untuk mengambil hipotesis dan penemuan baru tentang PRESTASI siswa. EDM terus berkembang seiring dengan banyaknya teknik data mining yang digunakan dilingkungan pendidikan. Oleh karena itu, proses teknik EDM dimulai dengan menemukan hubungan antar data dengan menggunakan teknik *description, estimation, prediction, classification, clustering, and association rule mining* [2]. Klasifikasi adalah salah satu teknik *data mining* yang banyak digunakan untuk menempatkan data kedalam kelompok- kelompok dengan tujuan untuk memilih kelas. Klasifikasi merupakan sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep dan kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui . Ada 2 model dalam metode klasifikasi, mode deskriptif dan model prediksi [3].

Peningkatan prestasi siswa dan peningkatan kualitas pendidikan adalah sangat penting bagi semua lembaga pendidikan. Untuk memberikan pendidikan yang berkualitas kepada peserta didik, analisis mendalam dari catatan peserta didik sebelumnya dapat memainkan peran penting. Prestasi siswa, kemajuan siswa dan potensi siswa sangat penting untuk mengukur hasil belajar, pemilihan materi pembelajaran dan kegiatan pembelajaran. Namun, pekerjaan yang ada tidak menyediakan alat analisis yang cukup untuk menganalisis bagaimana prestasi siswa, faktor mana yang akan memengaruhi prestasi mereka, dengan cara apa siswa dapat membuat kemajuan, dan apakah siswa memiliki potensi untuk tampil lebih baik [4].

2. Tinjauan Pustaka

Adapun mempercepat dan mempermudah dalam memprediksi prestasi siswa SMA pada Dinas Pendidikan Provinsi Jambi, maka sangat diperlukan sebuah pembanding dan kesamaan literatur dengan masalah yang diangkat oleh penulis, yaitu:

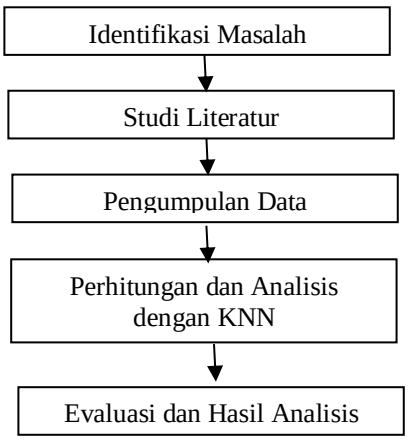
1. Penerapan Algoritma K-Nearest Neighbor (K-NN) Dengan Pencarian Optimal Untuk Prediksi Prestasi Siswa Oleh Yuyun Umaida & Purwantoro, Penelitian ini memanfaatkan teknik data mining menggunakan algoritma K-Nearest Neighbor (K-NN) dengan pencarian K-Optimal menggunakan metode k-fold cross validation untuk memprediksi prestasi siswa. Kriteria yang digunakan adalah: Les Tambahan, Jurusan, Nilai rata-rata rapor mata pelajaran pokok, Nilai rata-rata rapor mata pelajaran penjurusan, Nilai kedisiplinan, Jarak Tempuh, Ekstrakurikuler, Organisasi, dan Prestasi. Metodologi yang digunakan adalah CRISP-DM dan Performa Algoritma dilihat dari nilai accuracy, precision, recall, dan AUC dengan melakukan pemilihan k-fold cross validation (k=2, k=3, k=4, k=5, k=6, k=7, k=8, k=9, k=10). Setelah diperoleh hasil terbaik dari pemilihan k-fold cross validation akan dilakukan pengujian dengan pemilihan klaster k-NN (klaster1, klaster 2, klaster 3, klaster 4 dan klaster 5). Dari penelitian diperoleh hasil terbaik terdapat pada k=5 (5-fold cross validation) pada klaster 2 dengan hasil accuracy = 93.63%, precision=95.77%, recall=96.58% dan AUC=0.782.
2. Penerapan Metode K-Nearest Neighbor Dan Information Gain Pada Klasifikasi prestasi Siswa oleh Tyas Setiyorini dan Rizky Tri Asmoro. dalam hal dimensi vektor yang besar. Penelitian ini bertujuan untuk memprediksi prestasi akademik siswa menggunakan algoritma K-Nearest Neighbor dengan metode seleksi fitur Information Gain untuk mengurangi dimensi vektor. Beberapa percobaan dilakukan untuk mendapatkan arsitektur yang optimal dan menghasilkan klasifikasi yang akurat. Hasil dari 10 percobaan dengan nilai k (1 sampai dengan 10) pada dataset student performance dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Information Gain didapatkan rata- rata akurasi terbesar yaitu 76,553. Dari hasil pengujian tersebut maka dapat disimpulkan bahwa Information Gain mampu mengurangi dimensi vektor, sehingga penerapan K-Nearest Neighbor dan Information Gain dapat meningkatkan akurasi klasifikasi prestasi siswa yang lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja.
3. Klasifikasi prestasi Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor Dengan Seleksi Fitur Information Gain oleh Rizky Adinda Azizah, Fitra A. Bachtiar, dan Sigit Adinugroho. Klasifikasi menggunakan metode Neighbor Weighted K-Nearest Neighbor dengan seleksi fitur Information Gain diterapkan pada penelitian ini untuk membantu klasifikasi prestasi siswa karena metode NW-KNN mempunyai kelebihan memperhitungkan metode pembobotan

kelas dan mengatasi data tidak seimbang. Seleksi fitur dengan Information Gain digunakan agar dapat mengoptimalkan hasil kerja classifier. Berdasarkan pengujian dan analisis penelitian, didapatkan nilai akurasi terbaik sebesar 0,604, dengan nilai precision adalah 0,719, nilai recall sebesar 0,676, dan nilai f-measure diperoleh adalah 0,661. Nilai tersebut dihasilkan saat menggunakan 9 fitur yaitu VisitedResource, StudentAbsenceDay, RaisedHands, AnnouncementsView, Relation, Parents Answering Survey, Discussion, NationalITy, dan Place of Birth dimana fitur tersebut memperoleh nilai Gain tertinggi dari urutan Gain keseluruhan fitur, dengan nilai Gain ≥ 0,1182 dan menggunakan nilai parameter optimal yaitu nilai E = 6, dan nilai K = 45.

4. Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa oleh Esty Purwaningsih dan Ela Nurelasari. Prediksi dilakukan dengan menggunakan metode K-Nearest Neighbor (KNN). K-Nearest Neighbor sering digunakan pada klasifikasi prestasi siswa karena kesederhanaannya juga dapat memberikan hasil yang signifikan dan kompetitif. Hasil dari prediksi tingkat kelulusan siswa dengan metode KNN didapat rata-rata akurasi dengan nilai sebesar 96,49%. Pengolahan data dilakukan dengan menggunakan tools rapid miner. Output dari implementasi pada prediksi tingkat kelulusan dapat dijadikan sebagai acuan bagi siswa untuk meningkatkan prestasi dan predikat studi lanjut dimasa yang akan datang.
5. Penerapan Metode K-Nearest Neighbor Dan Gini Index Pada Klasifikasi prestasi Siswa oleh Tyas Setyorini dan Rizky Tri Asmono. K-Nearest Neighbor merupakan metode yang sederhana untuk klasifikasi prestasi siswa, namun K-Nearest Neighbor memiliki masalah dalam hal dimensi fitur yang tinggi. Untuk menyelesaikan masalah tersebut diperlukan metode seleksi fitur Gini Index dalam mengurangi dimensi fitur yang tinggi. Beberapa percobaan dilakukan untuk mendapatkan arsitektur yang optimal dan menghasilkan klasifikasi yang akurat. Hasil dari 10 percobaan dengan nilai k (1 sampai dengan 10) pada dataset student performance dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Gini Index didapatkan rata-rata akurasi terbesar yaitu 76,516. Dari hasil pengujian tersebut maka dapat disimpulkan bahwa Gini Index mampu mengatasi masalah dimensi fitur yang tinggi pada K-Nearest Neighbor, sehingga penerapan K-Nearest Neighbor dan Gini Index dapat meningkatkan akurasi klasifikasi prestasi siswa yang lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja.

3. Metode Penelitian

Pada Bab tiga ini menjelaskan tentang metodologi yang dipakai untuk melakukan penelitian, meliputi langkah-langkah yang diambil dalam melakukan penelitian. Pembuatan langkah atau alur dimaksudkan agar menghasilkan penelitian yang baik dan tetap sasaran. Adapun alur yang digunakan pada penelitian ini sebagai berikut:



Gambar 1. *Alur Penelitian*

3.1 Identifikasi Masalah

Pada tahap ini penulis melakukan identifikasi masalah terhadap hal apa saja yang menjadi permasalahan data mining mengenai prestasi siswa di beberapa SMA di kota Jambi. Identifikasi ini bertujuan untuk menentukan rencana kerja serta menentukan data apa saja yang akan dibutuhkan dalam penelitian ini.

3.2 Studi Literatur

Pada tahapan ini penulis menambah wawasan guna mendapatkan sebuah topik yang layak diangkat sebagai sebuah penelitian dengan mempelajari dan memahami teori dan konsep dimana penulis banyak melakukan penelitian pada buku, jurnal, *paper*, dan berbagai sumber diantaranya Perpustakaan Universitas Dinamika Bangsa Jambi.

3.3 Pengumpulan Data

Pada tahap ini penulis melakukan pengumpulan data dan informasi yang dibutuhkan dalam penelitian. Pengumpulan data ini dilakukan dengan beberapa metode yaitu :

a. Pengamatan Langsung (Observation)
   Penelitian dengan metode observation ini dilakukan dengan melakukan pengamatan langsung terhadap objek yang akan diteliti yang bertujuan untuk memperkuat data, mengetahui serta mendapatkan informasi secara langsung mengenai tentang prestasi siswa SMA di Kota jambi pada Dinas Provinsi Jambi.
b. Wawancara (Interview)
   Metode pengumpulan data yang dilakukan oleh penulis secara tatap muka antara penulis dengan narasumber, yaitu kepala Dinas Pendidikan untuk mendapatkan informasi yang dibutuhkan. Hal ini dilakukan agar penulis dapat memperoleh informasi langsung secara tepat dengan mengetahui permasalahan-permasalahan yang ada dan mempermudah dalam mengklasifikasi prestasi siswa pada beberapa SMA di kota Jambi.

3.4 Perhitungan dan Analisis dengan K-Nearest Neighbor

Pada tahap ini penulis melakukan perhitungan *decision tree* dengan menggunakan algoritma *K-Nearest Neighbor* dengan menghitung nilai entropy dan gain dari data-data yang telah dinormalisasikan dan digunakan untuk membuat hasil..

3.5 Evaluasi dan Hasil Analisis

Pada tahap ini hasil dari evaluasi dan analisis hasil akan dinilai keakuratan data yang sudah didapatkan dengan perhitungan Algoritma *K-Nearest Neighbor* dan mengelola data siswa mengenai prestasi siswa.

4. Hasil Penelitian dan Pembahasan

4.1 Identifikasi Masalah

Pada tahap ini penulis melakukan identifikasi masalah terhadap hal apa saja yang menjadi permasalahan data mining mengenai prestasi siswa SMA pada Dinas Pendidikan Provinsi Jambi. Identifikasi ini bertujuan untuk menentukan rencana kerja serta menentukan data apa saja yang akan dibutuhkan dalam penelitian ini.

4.1.1 Gambaran Umum Dinas Pendidikan Provinsi Jambi

Dinas Pendidikan Provinsi Jambi sebagai penanggung-jawab bidang pendidikan pada Tingkat Provinsi, salah satu aspek penting dalam hal tersebut adalah pengimplementasi Renstra SKPD dalam penyelenggaraan pendidikan. Namun demikian, koordinasi antar SKPD dan lembaga yang mengelola dan menyelenggarakan pendidikan, serta antara pemerintah pusat dengan pemerintah provinsi, pemerintah kota dan pemerintah kabupaten belum sepenuhnya tertata dengan baik. Demikian pula peran serta masyarakat dalam pengelolaan dan penyelenggaraan pendidikan belum dikelola dengan maksimal.

4.1.2 Analisis Permasalahan

Setiap tahunnya jumlah siswa yang lulus dan mendaftar di Universitas selalu bertambah, oleh karena itu nilai siswa yang lulus setiap tahunnya juga harus meningkat agar dapat mendaftar di universitas atau lembaha kedinasan yang diinginkan. Jika jumlah siswa yang lulus dengan nilai

rata-rata yang menurun maka dikhawatirkan akan terjadi ketidak seimbangan untuk melanjutkan pendidikan di tingkat yang lebih tinggi. Sejauh ini Dinas Pendidikan Provinsi jambi belum mempunyai suatu prediksi terhadap nilai-nilai yang telah dikumpulkan dalam database. Nilai yang telah dikumpulkan dapat dimanfaatkan menggunakan data minning agar nantinya dapat membantu Dinas Pendidikan Provinsi jambi dalam menentukan kebijakan terkait keberlanjutan studi siswa tersebut.

4.1.3 Solusi Permasalahan

Berdasarkan analisa permasalahan yang berjalan, maka penulis melakukan analisis klasifikasi prestasi siswa menggunakan metode klasifikasi KNN Pada tahun kelulusan 2020/2021 sebagai data training dan data angkatan 2021/2022 sebagai data testing yang dimana agar dapat mengetahui klasifikasi prestasi yang mendapatkan lulusan pada universitas, sekolah kedinasan, dan yang melanjutkan untuk bekerja. Selain itu penulis juga melakukan pengujian dengan menambahkan metode pembanding untuk mendapatkan hasil evaluasi model yang maksimal, sehingga diharapkan prediksi yang dilakukan peneliti menjadi lebih akurat. Hasil dari penelitian ini diharapkan dapat membantu Dinas Pendidikan Provinsi Jambi dalam membuat kebijakan untuk memberikan peningkatan kualitas pendidikan bagi siswa SMA di Provinsi Jambi.

4.2 Pembahasan

Pada Pembahasan yang dijelaskan adalah proses dari awal pengumpulan data, dimulai dari *Data Preparation, Data Cleaning, Data integrating, Data Transformating, Data Mining, Evaluation model.*

4.2.1 Data Preparation

Data preparation adalah tahapan pengumpulan data yang akan dijadikan ini diawali dengan melakukan pengambilan data training dari data siswa yaitu :

1. Data siswa yang telah lulus pada tahun 2020 dan 2021 sebanyak 1102 siswa.
2. Siswa dengan jurusan IPA sebanyak 659 siswa dan IPS sebanyak 443 siswa.

Sementara data testing yang akan digunakan adalah :

1. Data siswa pada Tahun 2022 sebanyak 645 siswa
2. Siswa dengan jurusan IPA 438 Siswa dan IPS sebanyak 71 siswa.

Sehingga total data yang di gunakan dalam pengujian adalah sebanyak 1747 siswa. Data ini didapatkan dari 2 Sekolah Menengah Atas Negeri yaitu SMAN 2 Kota Jambi dan SMAN Titian Teras H. Abdurrahman Sayoeti pada Dinas Pendidikan Provinsi Jambi.

4.2.2 Data Cleaning

Data cleaning merupakan, mengidentifikasi atau menghapus outlier, dan menyelesaikan inkonsistensi. Selain itu, data kotor dapat menyebabkan kebingungan pada prosedur *mining*, sehingga menghasilkan keluaran yang tidak dapat dipercaya. Meskipun sebagian besar rutinitas mining memiliki beberapa prosedur untuk menangani data yang tidak lengkap atau *noise*. Proses data cleaning yang dilakukan peneliti sepenuhnya menggunakan aplikasi *microsoft excel*.

4.2.3 Data Integrating

Data integrating adalah proses menggabungkan beberapa database menjadi 1 buah file sumber data yang besar. Penggabungan data pada siswa jurusan IPA dan IPS dilakukan untuk dapat menghasilkan sumber data yang lebih luas sesuai dengan keperluannya yaitu data mining. Pada data tabel 4.4 adalah data attribut nilai siswa pada attribut Fisika, Kimia, Biologi, Ekonomi, Sejarah, Geografi ketika digabungkan kita akan mendapatkan nilai attribut yang kosong. Sehingga data nilai penjurusan diperlukan penggabungan data menjadi data dengan nama attribut baru.

4.2.4 Data Transformation

Sebelum data digunakan dalam prediksi, data yang diterima akan diproses dulu sehingga dapat digunakan sebagai dataset untuk proses train. Data diolah secara manual dengan

Microsoft Excel, dengan menyesuaikan kondisi data, misalnya yang categorical diubah menjadi numerical, atau pengelompokan, rata-rata dan lainnya. Setelah data siap, maka selanjutnya disimpan dalam format csv. File Comma Separated Values (CSV) adalah file teks biasa yang berisi daftar data, file ini kadang bisa disebut Character Separated Values atau Comma Delimited files. File-file ini sering digunakan untuk bertukar data antara aplikasi yang berbeda. Pada umumnya file CSV menggunakan karakter koma untuk memisahkan (atau membatasi) antar data, tetapi terkadang menggunakan karakter lain, seperti titik koma.

4.2.5 Data Mining Menggunakan Python

Pembuatan Model

Dalam tahap implementasi ini, model dibuat berdasarkan dua hal, yaitu dataset yang akan menjadi train data, dan algoritma prediksi yang akan menjadi bentuk dari pemodelan, pada penelitian ini sudah ditetapkan akan menggunakan KNN atau K-Nearest Neighbors. Selain itu pada tahap implementasi ini akan menggunakan bahasa python, compiler yang digunakan adalah google colab yang dapat mendukung machine learning pandas dan sklearn dalam mengolah data dan pembentukan model. Berikut tahap pembuatan model:

1. *Import Library*

```
[190] import itertools
      import numpy as np
      import matplotlib.pyplot as plt
      from matplotlib.ticker import NullFormatter
      import pandas as pd
      import sklearn
      from sklearn import preprocessing
      %matplotlib inline
```

Gambar 2. *Import Library Pandas Dan Sklearn*

Penggunaan pustaka pada python sangat dianjurkan untuk mempermudah perhitungan data, adapun library yang digunakan untuk menghitung KNN seperti pada Gambar 2.

2. *Import Dataset*

Pada proses ini, data set yang miliki diproses menggunakan library pandas untuk dikonversi dari file semula csv menjadi data frame, sehingga dapat digunakan dalam pengolahan statistikal pembuatan model machine learning. Data set sebanyak 1102 siswa yang berupa file csv disimpan dengan format baru menggunakan library Pandas berupa data frame. Pembacaan data file dapat dilihat pada gambar 3.

```
df = pd.read_csv("/content/drive/MyDrive/Mbak Widya/datatraining_new.csv")
df.head
```

```
<bound method NDFrame.head of        No                          Nama  JK JK1  BI1  BI2  BI3  BI4  BI5  \
0       1          Andi Wildan Ar'rizky   1   L   84   90   88   94   85
1       2                Ardian Sinurat   1   L   83   85   85   94   88
2       3            B. Apri Ardiansyah   1   L   80   90   87   83   85
3       4       Christina Mutiara Ishak   0   P   81   90   88   95   83
4       5       Daniel Elvitto Jonathan   1   L   81   88   88   91   86
...   ...                           ...  ..  ..  ...  ...  ...  ...  ...
1097 1098            SYAKIRA YUMNA YUJO   0   P   85   88   86   87   86
1098 1099  SYALSYABILA MAHARANNI ZAFIRAH   0   P   88   89   88   93   93
1099 1100               Yayuk Ernawati   0   P   87   89   92   94   95
1100 1101                 Zalfa Zahira   0   P   79   79   82   84   83
1101 1102                Zikri alnazri   1   L   86   88   90   93   91
```

Gambar 3. *Import Dataset Sebagai menjadi Data Frame menggunakan Pandas*

Pada Gambar 3 terlihat bahwa terdapat pembacaan data Head frame yang label dari setiap attribut, dan jumlah data yang telah dikonversi menjadi data frame sejumlah 1102 data.
Data set yang telah diolah dan dijadikan data frame dikonversi menjadi sebuah variable data baru berupa array. Penggunaan data array dimaksudkan untuk menyimpan data setiap atribut agar bisa diolah pada proses minning menggunakan KNN.

```
[ ] x = df[[ 'BI1', 'BI2', 'BI3', 'BI4', 'BI5', 'M1',
           'M2', 'M3', 'M4', 'M5', 'BA1', 'BA2', 'BA3', 'BA4', 'BA5', 'KSA1',
           'KSA2', 'KSA3', 'KSA4', 'KSA5', 'KSAB1', 'KSB2', 'KSB3', 'KSB4', 'KSB5',
           'KSC1', 'KSC2', 'KSC3', 'KSC4', 'KSC5']] .values
    x[0:5]

    array([[88.2, 88.2, 85.2, 87.2, 85.8, 88.2],
           [87. , 83.6, 86. , 89.2, 85.6, 87.4],
           [85. , 80.6, 78.6, 86. , 78.8, 87.8],
           [87.4, 83. , 86. , 87.2, 82.6, 86.4],
           [86.8, 89.4, 87.6, 86. , 89. , 87.2]])
```

Gambar 4. *Konversi Data Frame menjadi Data Array*

Proses selanjutnya yang dilakukan adalah menormalisasikan data pada setiap value array, ini dimaksudkan agar menghilangkan data jumping dikarenakan perbedaan dari yang besar.

```
x = preprocessing.StandardScaler().fit(x).transform(x.astype(float))
x[0:5]

array([[ 0.02212613,  0.21730298, -0.05617575,  0.70411058,  0.22638026,
         0.04920128, -0.22027579,  0.03393114, -0.00641515,  0.49109994,
         0.54481441, -1.35304533,  1.06721136,  0.65948893,  0.09422437,
         0.87886216,  0.93176044,  1.1263736 , -0.63218456,  0.15165911,
        -0.13179381, -0.81090871,  0.13552586,  0.03802386,  0.24718671,
         0.71321092,  0.32992712,  1.39623193, -0.46325333,  0.09742162],
       [-0.28345369, -0.61917025, -0.32197024,  0.52643132,  0.22638026,
         0.90436646,  0.61973114,  0.72061676,  0.95774376,  0.32323902,
         1.28672025,  0.80448711,  1.39445108, -0.34587755, -0.70825317,
         1.28769107,  1.37384919,  1.60225364,  1.43715124, -0.93548444,
        -0.3044775 , -0.81090871, -0.66417059, -0.61509182, -1.4214052 ,
        -0.85963461, -0.67062337,  0.15799084, -0.71375328, -2.12500913],
       [-0.89461333, -0.61917025, -0.05617575,  0.17107279,  0.22638026,
        -1.14802996, -0.5002781 ,  0.54894535,  0.5720882 ,  1.16254359,
        -1.12447372, -0.6891892 ,  0.24911204, -0.74802414,  0.57571089,
        -0.96086795, -1.05763894, -1.01508658, -1.32196316, -0.15895333,
         0.04088987, -0.51765015, -0.98404917, -1.10492858, -1.58826439,
        -1.53371127, -0.42048575, -1.08025025, -0.96425322, -1.01379376],
       [ 0.32770595,  0.00818467, -0.05617575,  0.34875206,  0.06322044,
         0.04920128, -0.22027579,  0.54894535,  0.76491198, -0.01248281,
         0.1738615 , -0.35726113,  1.39445108, -0.54695084,  0.41521538,
         0.47003324,  0.48967169,  0.65049356,  0.05759404,  0.46227155,
        -0.13179381, -0.81090871,  0.13552586,  0.03802386,  0.4140459 ,
        -0.63494239, -1.17089861, -1.57554669, -0.46325333, -0.34706453],
       [-0.28345369, -1.03740687, -0.05617575,  0.17107279,  0.22638026,
        -0.12183175, -0.36027695, -1.1677687 , -0.58491049,  0.49109994,
         0.1738615 , -0.6891892 , -0.24174755,  1.06163552,  1.05719741,
        -0.7564535 , -0.83659456, -0.77714656, -0.86211076, -0.00364711,
        -0.64984487, -1.10416728, -0.98404917, -0.4518129 , -0.08653167,
        -1.08432683, -1.17089861, -0.83260203, -0.96425322,  1.43088007]])
```

Gambar 5. *Normalisasi Data Array*

Setelah hasil normalisasi data array didapat maka selanjutnya adalah menyimpang attribut yang dijadikan targer label pada data CSV menjadi sebuah data array, sehingga dapat dilakukan data minning. Target label yang dipilih adalah attribut dengan nama ”Lanjut”.

```
[195] y = df['Lanjut'].values
      y[0:5]

      array([0, 0, 0, 0, 0])
```

Gambar 6. *Menambah Target Label Kedalam Variabel Baru*

Setelah data target objek dan value objek yang digunakan sebagai data training telah diketahui, maka data training dipisah menjadi 80% data training dan 20% data test. Hal ini dimaksudkan agar kita tetap dapat menguji keakuratan data dari model yang telah kita buat dari data training.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.20, random_state=0)
print ('Train set:', x_train.shape, y_train.shape)
print ('Test set:', x_test.shape, y_test.shape)

Train set: (881, 30) (881,)
Test set: (221, 30) (221,)
```

Gambar 7. *Perbandingan Data Training dan Data Set*

3. *Membuat model dengan KNN berdasarkan dataset yang sudah disetup*
   Pada tahap ini, model akan dibuat dengan algoritma KNN. Dapat dilihat dari Gambar 8, hasil dari pembuatan model dengan dukungan sklearn dan memasukan data set x dan y. penentuan nilai K bisa diubah sesuai diinginkan pada penelitian ini K yang digunakan adalah 3 dan metric perhitungan jarak menggunakan Euclidean.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
k=3
knn = KNeighborsClassifier(n_neighbors = k,metric='euclidean').fit(x_train,y_train)
knn
yhat = knn.predict(x_test)
yhat
print ("Train set Accuracy:", metrics.accuracy_score(y_train, knn.predict(x_train)))
print ("Test set Accuracy:",metrics.accuracy_score(y_test,yhat))

Train set Accuracy: 0.7990919409761634
Test set Accuracy: 0.6108597285067874
```

Gambar 8**.** *Perhitungan KNN Dengan Perhitungan Jarak Eculidean Distance*

Selain menggunakan metric perhitungan jarak eculidegambaran kita dapat menggunakan perhitungan jarak default minkowski data hasil training akuraasi dan test akurasi juga mendapatkan hasil yang sama.

```
[15] from sklearn.neighbors import KNeighborsClassifier
     from sklearn import metrics
     k=3
     knn = KNeighborsClassifier(n_neighbors = k).fit(x_train,y_train)
     knn
     yhat = knn.predict(x_test)
     yhat

     array([1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1,
            1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1,
            1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0,
            1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1,
            1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0,
            0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0,
            1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1,
            1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1,
            1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
            0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0,
            1])

[16] print ("Train set Accuracy:", metrics.accuracy_score(y_train, knn.predict(x_train)))
     print ("Test set Accuracy:",metrics.accuracy_score(y_test,yhat))

     Train set Accuracy: 0.7990919409761634
     Test set Accuracy: 0.6108597285067874
```

Gambar 9. *Pengujian Dengan Metric Minkowski*

4. *Membuat model dengan naive bayes berdasarkan dataset*

Pada tahap ini, model akan dibuat dengan algoritma naive bayes. Dapat dilihat dari Gambar 10, hasil dari pembuatan model dengan dukungan sklearn dan memasukan data set x dan y. Penentuan algoritma perhitungan menggunakan Gaussian.

```
[114] from sklearn.compose import ColumnTransformer
      prepocessor = ColumnTransformer([('numeric',num_pipe(),['BI1', 'BI2', 'BI3', 'BI4
            'BA2', 'BA3', 'BA4', 'BA5', 'KSA1', 'KSA2', 'KSA3', 'KSA4', 'KSA5',
            'KSB1', 'KSB2', 'KSB3', 'KSB4', 'KSB5', 'KSC1', 'KSC2', 'KSC3', 'KSC4',
            'KSC5'])])

[115] from sklearn.naive_bayes import GaussianNB
      pipeline = Pipeline([
          ('prep',prepocessor),
          ('algo',GaussianNB())
          ])
```

Gambar 10. *Pembuatan Model Data Mining Klasifikasi Naive Bayes*

```
[127] pipeline.score(x_train,y_train)

      0.619750283768445

[128] pipeline.score(x_test,y_test)

      0.583710407239819
```

Gambar 11. *Hasil Pengujian Akurasi Menggunakan Naive Bayes*

Berdasarkan hasil pengujian menggunakan metode naive bayes maka didapat hasil pengujian score akurasi data training sebesar 61% dan data test sebesar 58% ini menunjukan bahwa perhitungan naive bayes kurang cocok digunakan untuk meprediksi prestasi siswa.

5. *Uji Prediksi*

Pada tahap ini, peneliti akan memasukan beberapa data yang akan digunakan sebagai test. Data test ini akan terdiri dari features saja tanpa label, sehingga dapat dilihat hasil dari prediksi tersebut. Berikut data yang dijadikan test dari siswa tahun SMA pada Dinas Pendidikan Provinsi Jambi jumlah yang digunakan adalah sebanyak 645 siswa.

Tabel 1. *Hasil Prediction menggunakan Python*

No	Nama	BI1	BI2	BI3	BI4	BI5	MM1	MM2	MM3	MM4	MM5	BA1	BA2	BA3	BA4	BA5	KSA1	KSA2	KSA3	KSA4	KSA5	KSB1	KSB2	KSB3	KSB4	KSB5	KSC1	KSC2	KSC3	KSC4	KSC5	Pre1
1	Adityarahmansyah Putra	81	83	96	93	90	80	83	83	94	94	85	85	90	92	90	83	93	86	88	90	86	80	87	83	95	85	88	86	88	88	1
2	Alya Nabila Abas	86	88	94	92	90	82	82	85	80	85	80	83	88	91	80	72	86	86	92	88	81	83	89	81	90	85	85	73	86	87	1
3	Ananda Zahwa Dasyifa	83	85	96	94	91	80	84	83	92	93	71	80	86	90	80	75	86	86	91	86	82	87	86	82	90	80	87	77	87	88	1

4	Anita Yuliana Tampubolon	87	89	84	94	05	59	84	95	84	85	90	91	90	93	94	90	93	93	80	88	86	80	96	90	86	91	92	93	1
5	Aufaditama Rytya Putra	80	82	70	71	09	90	82	95	86	81	81	75	81	80	77	82	85	88	89	78	75	76	77	90	85	81	71	83	84	1
6	Azmi Muhammad Padhil	84	86	63	30	58	85	95	96	82	83	88	90	90	76	89	85	90	94	84	81	88	80	90	85	89	91	90	91	1
7	Aznof Fefri Al'amin	85	87	63	30	03	32	93	77	80	85	90	80	77	84	87	91	92	83	77	88	80	90	85	82	85	88	89	0		
8	Billy Widarta Salim	83	85	41	13	25	58	91	91	92	94	94	90	82	86	90	93	93	89	82	90	81	85	85	88	90	92	92	1		
9	Dwi Putri Rahmadani	87	89	63	30	01	35	96	83	84	88	88	85	84	87	90	91	96	85	86	88	90	95	90	84	85	90	91	1		
10	Evlyn Shinta Marito Sitanggang	85	87	56	60	91	12	85	73	73	73	78	85	76	78	73	82	90	81	79	77	82	90	80	81	70	80	80	1		
644.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

4.2.6 Evaluasi

Evaluasi dilakukan untuk menghitung kembali akurasi dari model yang sudah dikembangkan. Klasifikasi adalah pendekatan pembelajaran terawasi di mana variabel target bersifat diskrit (atau kategoris). Mengevaluasi model pembelajaran mesin sama pentingnya dengan membangunnya. Kami membuat model untuk digunakan pada data baru yang sebelumnya tidak terlihat. Oleh karena itu, evaluasi menyeluruh dan serbaguna diperlukan untuk membuat model yang kuat. Evaluasi akan dilakukan dengan mempelajari :

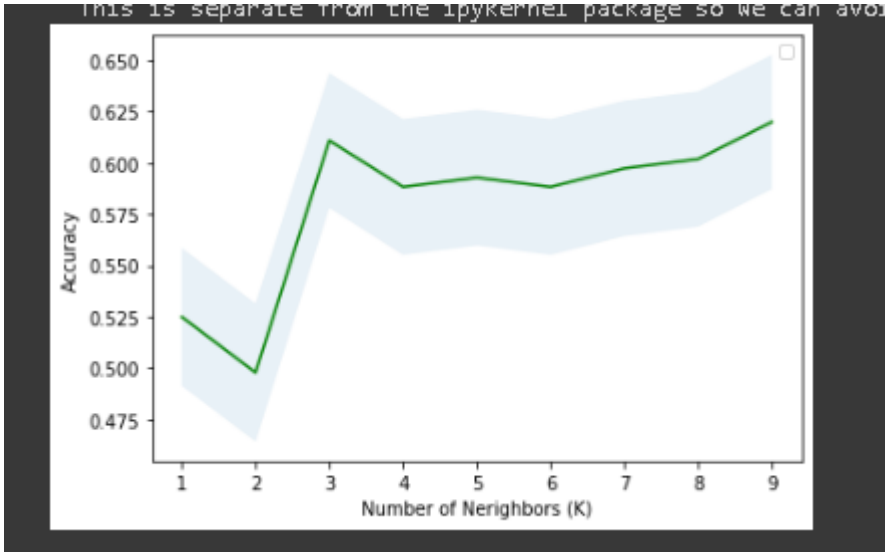
```
Ks=10
mean_acc =np.zeros((Ks-1))
std_acc = np.zeros((Ks-1))
ConfustionMx = []
for n in range (1,Ks):
  neigh = KNeighborsClassifier(n_neighbors= n).fit(x_train,y_train)
  yhat = neigh.predict(x_test)
  mean_acc[n-1] = metrics.accuracy_score(yhat,y_test)
  std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])

mean_acc
```

```
array([0.52488688, 0.49773756, 0.61085973, 0.58823529, 0.59276018,
       0.58823529, 0.59728507, 0.60180995, 0.6199095 ])
```

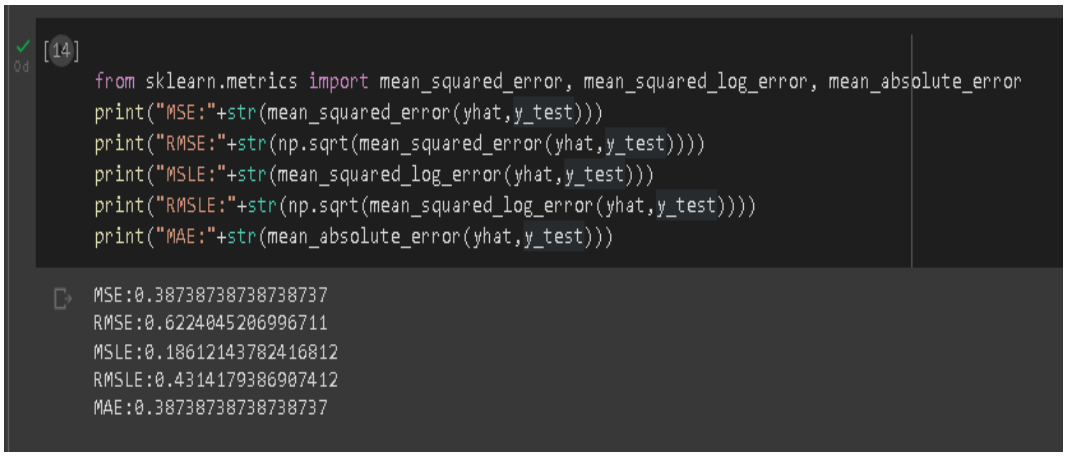
Gambar 12. *Membuat Perhitungan Akurasi Dengan Perulangan Ks*

Pada perhitungan diatas akan mendapatkan metrix skore akurasi nantinya akan ditampilkan untuk mendapatkan nilai maksimal pada perulang ke-n. Percobaan menggunakan 10 kali perulangan dengan hasil tertinggi terdapat pada perulangan ke 9 maka dari itu gambar grafik akan ditampilkan pada gambar 13.



Gambar 13. *Grafik Perhingan Akurasi F1 Score Penggunaan Python*

Selain mengevaluasi model menggunakan metrik F1 score penulis juga mengevaluasi model dengan tinggkat error dengan perhitungan mean squared error, root mean squred error, mean absoulte error. Hasil perithungan error dapat dilihat pada gambar 14.



```
[14]
from sklearn.metrics import mean_squared_error, mean_squared_log_error, mean_absolute_error
print("MSE:"+str(mean_squared_error(yhat,y_test)))
print("RMSE:"+str(np.sqrt(mean_squared_error(yhat,y_test))))
print("MSLE:"+str(mean_squared_log_error(yhat,y_test)))
print("RMSLE:"+str(np.sqrt(mean_squared_log_error(yhat,y_test))))
print("MAE:"+str(mean_absolute_error(yhat,y_test)))

MSE:0.38738738738738737
RMSE:0.6224045206996711
MSLE:0.18612143782416812
RMSLE:0.4314179386907412
MAE:0.38738738738738737
```

Gambar 14. *Hasil Perhitungan Error Evaluasi Model KNN*

5. Kesimpulan

5.1 Simpulan

Berdasarkan hasil penelitian yang telah penulis lakukan dalam Penerapan Data Mining Untuk Meprediksi prestasi Siswa SMA Pada Dinas Pendidikan Provinsi Jambi, maka dapat disimpulkan bebarapa hal, antara lain :

1. Penentuan prestasi siswa SMA pada Dinas Pendidikan Provinsi Jambi saat ini masih dilakukan secara manual, sehingga Dinas Pendidikan selaku instansi Pemerintah Provinsi Jambi yang menaungi urusan peningkatan mutu pendidikan membutuhkan waktu yang agak lama untuk menentukan kebijakan yang akan diambil. Hal ini dikarenakan belum adanya penerapan Educational Data Mining untuk memprediksi prestasi siswa yang membantu memberikan pertimbangan bagi Dinas Pendidikan Provinsi Jambi dalam menetukan Rencana strategis (RENSTRA) pada tahun yang akan datang.
2. Hasil klasifikasi menggunakan data training 80% dan testing 20%, dengan K = tmenggunakan KNN dengan k =3 pada Python dengan perhitungan jarak menggunakan metode euclidean didapatkan F1 *score accuracy* sebesar 61.99% dan untuk MSE 0.38 sementara itu untuk metode

klasifikasi menggunakan naive bayes mendapatkan F1 *score accuracy* sebesar 58% dan untuk MSE sebesar 0.4. Prediksi menggunakan KNN pada siswa tahun lulusan 2022 pada data test sebesar 56.89% siswa akan melanjutkan jenjang pendidikan yang lebih tinggi. Sementara untuk dengan menggunakan aplkasi Rapid Miner dengan perhitungan jarak menggunakan eculidean didapatkan F1 Score accuracy sebesar 51% bahwa diprediksi siswa tahun siswa tahun lulusan 2022 pada data test sebesar 57.51% siswa akan melanjutkan jenjang pendidikan yang lebih tinggi.

3. Dengan menggunakan hasil data mining yang telah diolah dapat membantu memberikan pertimbangan kepada Dinas Pendidikan Provinsi Jambi dalam menentukan kebijakan dan langkah-langkah yang tepat sebagai acuan rencana strategis berdasarkan hasil prediksi prestasi siswa SMA.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan maka dapat dikemukakan saran-saran sebagai berikut:

1. Diharapkan kedepannya penelitian ini menggunakan data siswa yang lebih banyak dari semua Sekolah SMA yang ada di Provinsi Jambi serta mencakup atribut lainnya agar memiliki presentasi akurasi lebih baik.
2. Penelitian ini diharapkan kedepannya akan ada penelitian yang melakukan perbandingan dengan metode algoritma data mining lainnya.

6. Daftar Rujukan

[1] Abdullah, T. Mohd, dkk. 2015. *A Survey of Anomaly Detection Using Data Mining Methods for Hypertext Transfer Protocol Web Services*, Journal of Computer Science. Volume 11. No,1, 2015.

[2] Larose, D.T & Larose, C.T. 2014. *Discovering Knowledge In Data An Introductionto Data Mining*. New Jersey: Willey.

[3] Budiman, dkk .2015. *Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi*. Jurnal Penelitian Ilmu dan Teknologi Komputer, Vol 7 No 1 (2015): Jupiter April 2015.

[4] F. Yang and F. W. B. Li. *Study on student performance estimation, student progress analysis, and student potential prediction based on data mining*, Comput. Educ., vol. 123, pp. 97–108, Aug. 2018.

[5] Gunawan, Harry, and Vega Purwayoga. "Data Mining Menggunakan Algoritma K-Means Clustering Untuk Mengetahui Potensi Penyebaran Virus Corona di Kota Cirebon." Jurnal Sisfokom (Sistem Informasi dan Komputer) 11.1 (2022): 1-8.

[6] Mardi, Yuli (2014) Analisa Data Rekam Medis untuk Menentukan Penyakit Terbanyak Berdasarkan International Classification Of Disease (ICD) Menggunakan Decision Tree C4.5 (Studi Kasus : RSU. CBMC Padang). UPI YPTK Padang

[7] Bramer, Max (2007), *Principles of Data Mining*, Springer Science.

[8] Laudon & Laudon. 2020. *Management Information Systems (Sixteenth)*. United Kingdom: Pearson Education.

[9] Kantardzic, Mehmed (2020), *Data Mining: Concepts, Models, Methods, And Algorithms, Edisi ke-3*, New Jersey: John Wiley & Sons, Inc.

[10] Pine, J. David. 2019. *Introduction to Python for Science and Engineering (Series in Computational Physics)*. New York: CRC Press.