Komparasi Algoritma Data Mining Decision Tree, KNN, dan Random Forest Classifier untuk Menentukan Prediksi Pengguna Asuransi Perjalanan

Ahmad Izza Zain Firdaus

Sistem Informasi, Universitas Negeri Surabaya
ahmad.19063@mhs.unesa.ac.id

***Abstrak*— Keperluan untuk memperkirakan target pasar sangat penting, termasuk dalam bidang asuransi perjalanan. Sehingga untuk menentukan target pelanggan dapat menggunakan algoritma data mining seperti decision tree, knn dan random forest. Penelitian dilakukan menggunakan metodologi SEMMA dimana menjadi metodologi yang cocok untuk melakukan segmentasi pasar. Dataset yang digunakan berisi 1277 sampel dengan 9 kriteria untuk menentukan prediksi model**

***Kata Kunci*— Data Mining, SEMMA, Decission Tree, KNN, Random Forest,**

I. PENDAHULUAN

Bagi beberapa orang melakukan perjalanan (travelling) menjadi hal yang sangat penting, karena dengan melakukan perjalanan memberikan kita istirahat dari rutinitas yang menjemukan. Dalam sebuah perjalanan tentu saja memungkinkan terjadi resiko terjadi hal yang tidak diinginkan misalkan saja kecelakaan dalam perjalanan. Untuk itu ada sebuah asuransi yang memberikan kita jaminan apabila terjadi kecelakaan dalam perjalanan. Setelah pandemi COVID-19 yang memberikan penurunan drastis pada bidang perjalanan menjadikan diperlukan melakukan analisa prediksi pengguna jasa asuransi perjalanan, sehingga pihak penyedia asuransi perjalanan dapat menentukan target pasar seperti apa yang layak dituju untuk melakukan iklan asuransi perjalanan. Untuk itu dalam melakukan prediksi menggunakan 3 algoritma data mining yakni decision tree, KNN, dan Random Forest

II. TINJAUAN PUSTAKA

A. *Data Mining*

Data mining menurut Bellazi dan Zupan adalah bagaimana kita menemukan pola atau hubungan dari suatu data dengan cara mengeksplorasi dan memodelkan data ketika melakukan analisa data [1]

Dalam ilmu kita belajar mengenai bagaimana kita memanfaatkan sebuah data, karena data sendiri apabila tidak diproses belum bisa menjadi informasi yang bermanfaat. Dengan adanya data mining diharapkan dapat membantu organisasi untuk melakukan berbagai hal, seperti memprediksi kebutuhan produksi, mengkategorikan target pasar dan masih banyak lagi

B. *SEMMA*

Dalam melakukan penelitian analisa data mining diperlukan beberapa model tahapan, salah satunya adalah SEMMA. SEMMA menjadi salah satu métodо penelitian dalam data mining yang mana memiliki tahapan yang cukup singkat karena hanya tergabung dalam 5 proses.

Tetapi meskipun hanya memiliki 5 proses, métode tersebut memiliki proses yang lengkap karena mencangkup berbagai hal.

III. METODOLOGI

Dalam Penelitian menggunakan metode SEMMA dimana tahapan pencarian keputusan terbaik melalui beberapa tahapan meliputi:.

A. *Sample*

Pada tahapan ini dilakukan identifikasi dari data yang tersedia, menganalisa data dan menentukan bagaimana data akan diolah, variabel mana yang menjadi variabel independen maupun independen dan disiapkan untuk kategori berikutnya..

B. *Explore*

Dalam tahapan ini dilakukan analisa untuk mempelajari hubungan dari antar-variabel dan memeriksa apakah semua variabel saling berhubungan atau tidak.

C. *Modify*

Pada tahapan ini dilakukan pemrosesan data mentah agar dapat dilakukan permodelan. Data mentah sebelumnya diurai dan dibersihkan, data juga dilakukan perbaikan jika diperlukan. Pada tahapan ini sering dikenal sebagai tahapan preprocessing.

D. *Model*

Dengan data yang sudah diberishkan maka dapat dilakukan permodelan data berdasarkan alogoritma data mining untuk mendapatkan model terbaik sehingga didapatkan prediksi terbaik

E. *Asses*

Pada tahapan ini model dievaluasi untuk mengecek seberapa sesuai dengan topik penelitian yang dilakukan..

IV. PEMBAHASAN

A. *Sample*

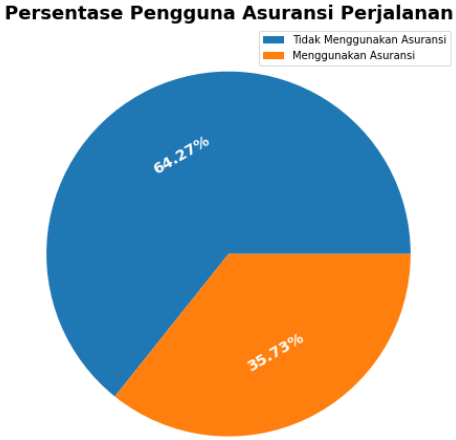
Dataset yang digunakan adalah sebuah dataset mengenai kriteria yang mungkin digunakan untuk mengelompokkan calon pengguna asuransi perjalanan data berisi 10 kolom dengan masing-masing kolom berisi 1987 data. Kolom yang ada adalah sebagai berikut:

TABEL I
DAFTAR NAMA KOLOM DARI DATASET

Nama Kolom	Tipe Data	Keterangan
Id	numerik	0-1986
Age	Numerik	25-35
Employment type	Kategorik	Government Sector, Private Sector/Self Employed
Graduation	Kategorik	Yes, No
annualincome	Numerik	300000-1800000
familymembers	Numerik	2-9
chronicdiseases	Kategorik	0, 1
Frequentflyer	Kategorik	Yes, No
evenrtravelledabroad	Kategorik	Yes, No
travelinsurance	Kategorik	0, 1

B. *Explore*

Pembagian data yang digunakan permodelan dibagi menjadi dua yakni mereka yang menggunakan asuransi perjalanan dan mereka yang tidak menggunakan asuransi perjalanan. Semua variable selain variable id akan digunakan karena variable id hanya digunakan dalam pengumpulan data dan membedakan antar satu data dengan data lain. Dari data pengguna dan data yang tidak menggunakan asuransi memiliki jumlah yang berbeda.



Gbr1. Diagram persebaran data pengguna asuransi

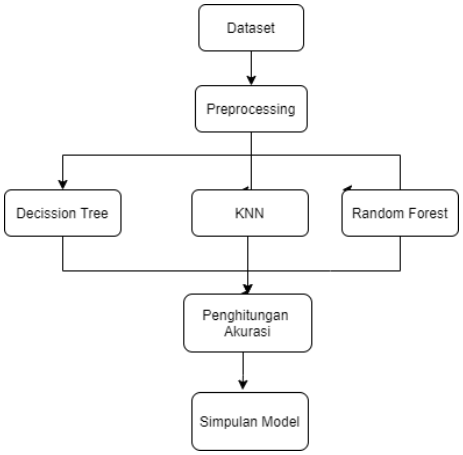
Dari diagram tersebut persentase yang tidak menggunakan asuransi sebesar 64,27% dan yang menggunakan asuransi sebesar 35,73% dan dalam jumlah yang tidak menggunakan asuransi sebanyak 1277 sampel dan yang menggunakan asuransi sebanyak 710 sampel.

C. *Modify*

Pada tahapan modify dilakukan beberapa pengubahan dari mentah yang sebelumnya. Dimulai dari melakukan encoding data yang masih berupa objet menjadi integer. Dalam kasus kolom yang masih berupa object adalah employment type, graduateornot, frequentflyer dan evertravelledabroad. Lalu karena isi kolom dari beberapa data masih dalam skala yang tidak konsisten antar kolom, maka dilakukan scaling agar data lebih mudah dimodelkan. Setelah semua data tersebut siap tetapi persebaran kelas masih belum seimbang maka dilakukanlah imbalancing data dari sebelumnya kategori 0 berjumlah 1277 sampel dan kategori 1 sejumlah 710 dengan menggunakan *oversampling* SMOTE data seimbang sejumlah 1277 data

D. *Model*

Dalam tahapan model dilakukan komparasi beberapa algoritma yakni menggunakan decission tree, knn, dan random forest sehingga dapat digambarkan menjadi sebagai berikut:



Gbr2. Proses Pencarian Algoritma Terbaik

Dari berbagai tahapan tersebut menghasilkan beberapa hasil yang berbeda. Hasil perhitungan perbandingan 3 algoritma dapat dilihat di tabel berikut:

TABEL IIIII
HASIL PERBANDINGAN ALGORITMA

Algoritma	Hasil
Decission Tree	0,81
KNN	0,75
Random Forest Classifier	0,81

Dari komparasi tiga algoritma tersebut masih perlu dilakukan pengecekan akurasi dan didapati hasil sebagai berikut:

TABEL IVVVI
NILAI AKURASI DARI MASING-MASING ALGORITMA

Algoritma	Hasil

Decission Tree	0,80
KNN	0,74
Random Forest Classifier	0,80

Nilai akurasi tersebut didapatkan menggunakan metode Accuracy, F-Score, dan AUC.

E. *Assess*

Dari hasil perbandingan pada tahapan sebelumnya, dapat dinyatakan bahwa algoritma Decission tree dan Random Forest Classifier menjadi yang terbaik untuk mengelola dataset tersebut, karena nilai berada pada angka 0,81 dan akurasi pada nilai 0,80.

V. KESIMPULAN

Dari penelitian yang dilakukan dengan membandingkan ketiga algoritma didapati algoritma decision tree dan random forest classifier memiliki hasil dan akurasi yang paling tinggi dalam model ini, sehingga kedua model tersebut bisa digunakan untuk melakukan prediksi dalam kasus pelanggan asuransi perjalanan.

REFERENSI

[1] R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: Current issues and guidelines,” Int. J. Med. Inform., vol. 77, no. 2, pp. 81–97, 2008