

BAB I

PENDAHULUAN

1.1 Latar Belakang

Sekolah merupakan lembaga atau sarana dalam melaksanakan pelayanan belajar atau proses pendidikan. Sebagai organisasi pendidikan formal, sekolah memiliki tanggung jawab dalam meningkatkan mutu pendidikan. Penyelenggaraan satuan pendidik secara baik, tertata dan sistematis hingga proses yang terjadi didalamnya dapat menjadi suatu sumbangan besar bagi kehidupan sosial masyarakat.

Sekolah sebagai suatu institusi yang melaksanakan proses pendidikan dalam tatanan mikro menempati posisi penting, karena di lembaga inilah setiap anggota masyarakat dalam ini siswa dapat mengikuti proses pendidikan dengan tujuan membekali mereka dengan ilmu dan pengetahuan sehingga menjadi manusia yang berkualitas.

Saat ini, masalah nilai akhir siswa dan faktor penyebabnya menjadi topik yang menarik untuk diteliti. Sekolah perlu mendeteksi perilaku siswa yang memiliki nilai “rendah” tersebut sehingga dapat diketahui faktor – faktor penyebab nilainya. Beberapa penyebab nilai siswa diantaranya kurangnya kemampuan akademik dalam menilai, bagaimana pendidikan orang tuanya, dan faktor lainnya.

Database siswa menyimpan data akademik, administrasi dan biodata siswa. Data tersebut apabila digali dengan tepat maka dapat diketahui pola atau pengetahuan untuk mengklasifikasikan siswa. Serangkaian proses pengetahuan atau pola dari kumpulan data disebut dengan *Data Mining* atau *Machine learning*. *Machine learning* memecahkan masalah dengan menganalisis data yang telah ada dalam *database*.

Data Mining dalam dunia pendidikan dikenal dengan *Educational Data Mining*. *EDM* mengembangkan metode untuk menggali data pendidikan dan

menggunakan metode tersebut untuk lebih memahami siswa. *EDM* dapat membantu pendidik untuk menganalisis cara belajar, mendeteksi siswa yang memerlukan dukungan dan memprediksi performa siswa.

Dengan berlimpahnya data siswa, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengolahan terhadap data siswa sehingga berguna bagi sekolah. Pengolahan data siswa perlu dilakukan untuk mengetahui informasi penting berupa pengetahuan baru (*Discovering knowledge*) dan dengan begitu berdasarkan data yang diperoleh dari *UCI Repository Machine learning* yaitu data perolehan siswa di pendidikan menengah. Atribut data meliputi kelas siswa, demografi, Sosial dan sekolah. Data ini didapat dari laporan sekolah dan kuesioner, 2 *dataset* didapatkan terkait performa siswa. Terdiri dari 2 subyek yaitu matematika dan bahasa Portugis, dan di sini penulis mengambil subyek matematika.

Pengkategorian atribut yang terpengaruh pada klasifikasi nilai akhir siswa masih cukup rumit dan cenderung subyektif. Masih sangat sulit untuk mendefinisikan permasalahan tersebut. Tetapi ada beberapa ciri khusus yang dapat digunakan untuk pengkategorian nilai akhir berdasar atribut pada siswa.

Penelitian mengenai prediksi performa akademik siswa telah dilakukan untuk mengidentifikasi faktor apa saja yang mempengaruhi hasil akhir akademik siswa. Paulo Cortez dan Alice Silva memprediksi performa akademik siswa dan mengidentifikasi faktor yang relevan dengan memperhatikan hasil model klasifikasi yang tertinggi tingkat akurasinya.

Prediksi performa akademik siswa tingkat menengah juga dilakukan Ramaswami dan Rathinasabapathy dengan menerapkan metode *Bayesian Networks* dan tiga teknik seleksi fitur (*Information Gain, Chi Square, dan Consistency Subset Evaluation*). Ramesh dkk juga mengidentifikasi faktor yang mempengaruhi performa akademik siswa di ujian akhir dan menerapkan beberapa algoritma klasifikasi (*Naïve Bayes, Multi Layer Perception, SMO, J48, REPTree*) dengan 10 atribut yang

didapatkan dari rata-rata ranking terbaik dari lima teknik seleksi fitur yang diterapkan (*Chi Square, Information Gain, OneR, Symmetrical Uncertainty (SU), dan ReliefF*). Osmanbegovic,dkk menerapkan teknik seleksi fitur *Information Gain* dan *Gain Ratio* dengan batas *Threshold* 0.01 untuk mereduksi dimensi data siswa lalu diterapkan pada teknik klasifikasi algoritma *rule based, tree-based, function based, dan Bayes-based*.

Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. “Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages.” (2009), mengatakan bahwa “Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya”. Namun Naive Bayes memiliki kekurangan yaitu Asumsi independence antar atribut membuat akurasi berkurang (karena biasanya ada keterkaitan).

Feature Selection adalah salah satu cara untuk menentukan atribut yang paling berpengaruh di dalam *dataset*. *Feature Selection* berperan memilih *subset* yang tepat dari set fitur asli, karena tidak semua fitur / atribut relevan dengan masalah. Bahkan beberapa dari fitur atau atribut tersebut mengganggu dan dapat mengurangi akurasi. *Noisy features* atau fitur tidak terpakai tersebut harus dihapus untuk meningkatkan akurasi. Selain itu degan fitur atau atribut yang banyak memperlambat proses komputasi. Dinakaran dkk menerapkan teknik seleksi fitur *Information Gain* dengan metode perangkingan fitur terbaik dan algoritma klasifikasi *Decision Tree J48*. Ramaswami dan Bhaskaran melakukan studi komparasi lima teknik seleksi fitur dan menerapkan empat algoritma klasifikasi pada *Data Mining* pendidikan. Hasil eksperimen menunjukkan bahwa teknik seleksi fitur *Information Gain* menunjukkan hasil yang terbaik.

Dari kesimpulan di atas maka diajukan Judul ***FITUR SELEKSI INFORMATION GAIN UNTUK MENENTUKAN ATRIBUT YANG BERPENGARUH PADA KLASIFIKASI NILAI AKHIR SISWA MENGGUNAKAN ALGORITMA NAÏVE BAYES.***

1.2 Rumusan Masalah

Berdasarkan latar belakang sebagaimana diuraikan di atas, maka dapat diambil rumusan masalah sebagai berikut :

- a. Bagaimana penerapan algoritma *Naïve Bayes* disertai fitur seleksi *Information Gain* pada klasifikasi nilai akhir siswa ?
- b. Bagaimana cara mengukur tingkat *confusion matriks* algoritma *Naïve Bayes* dengan algoritma *Naïve Bayes* disertai seleksi fitur *Information Gain* pada klasifikasi nilai akhir siswa ?

1.3 Batasan Masalah

Agar tidak terjadi penyimpangan dalam proses penelitian pada fitur seleksi *Information Gain* pada klasifikasi nilai akhir siswa, maka diterapkan beberapa batasan permasalahan, adapun batas masalahnya :

- a. Penelitian ini sebatas Analisis klasifikasi menggunakan *Naïve Bayes* dan seleksi fitur / atribut *Information Gain*
- b. Penelitian ini hanya menampilkan fitur/atribut yang sangat berpengaruh dalam klasifikasi performa siswa
- c. Dataset diambil dari *UCI Machine Learning* yaitu *Dataset Student Performance* pada mata pelajaran Matematika sekolah menengah di dua sekolah Portugis
- d. *Tools Data Mining* yang digunakan yaitu menggunakan *Tools WEKA*

1.4 Tujuan dan Manfaat

Adapun tujuan dari penelitian ini adalah sebagai berikut :

- a. Mengidentifikasi atribut yang sangat berpengaruh dalam nilai siswa dan mengurangi *noisy feature* (atribut tidak terpakai)
- b. Pengetahuan yang didapat dapat berguna bagi akademisi untuk klasifikasi nilai akhir siswa

Adapun Manfaat dari penelitian ini adalah sebagai berikut :

- a. Hasil penelitian ini dapat menjadi referensi untuk meningkatkan akurasi suatu *Data Mining* dengan menambahkan *Feature Selection*
- b. Hasil penelitian ini diharapkan dapat memberikan kontribusi bagi akademisi bagaimana menilai siswa