PM-19

PREDIKSI PERFORMA AKADEMIK SISWA PADA PELAJARAN MATEMATIKA MENGGUNAKAN BAYESIAN NETWORKS DAN ALGORITMA KLASIFIKASI MACHINE LEARNING

Betha Nurina Sari
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
betha.nurina@staff.unsika.ac.id

Abstrak

Mengidentifikasi data yang relevan agar bisa memprediksi performa akademik siswa merupakan hal yang menarik untuk diteliti. Penelitian ini menggunakan data akademik dan personal siswa menengah pada nilai mata pelajaran matematika di Portugal. Tujuan penelitian ini adalah mempelajari bagaimana Bayesian Networks diterapkan agar dapat menentukan keterkaitan antar variabel dan mengetahui variabel apa saja yang berpengaruh, lalu digunakan sebagai teknik seleksi fitur. Struktur graf Bayesian Networks yang divisualisasikan menggunakan software CaMML versi 1.4.1 dan package J-API. Algoritma klasifikasi machine learning yang diterapkan di antaranya Random Forest, MLP, SVM, dan Naïve Bayes. Evaluasi prediksi dilakukan dengan membandingkan akurasi sebelum dan sesudah melalui pemodelan struktur graf Bayesian Networks. Terdapat sebelas struktur Graf Bayesian Networks yang terbentuk dari eksperimen tahap pertama dan digunakan untuk langkah menyeleksi variabel yang relevan. Eksperimen tahap kedua menggunakan 10 fold cross validation menunjukkan bahwa model dari struktur graf Bayesian Networks dapat meningkatkan performa algoritma klasifikasi machine learning dalam memprediksi performa akademik siswa.Hasil penelitian ini juga menunjukkan ada 25 variabel yang efektif untuk memprediksi perfoma akademik siswa pada mata pelajaran matematika

Kata Kunci: *bayesian networks, klasifikasi, performa akademik siswa, prediksi,*

1. PENDAHULUAN

Mengidentifikasi data yang relevan agar bisa memprediksi performa siswa merupakan hal yang menarik untuk diteliti. Dalam proses *discovering knowledge*, mengidentifikasi data yang relevan tersebut berguna untuk basis model klasifikasi atau prediksi (Transition, Osmanbegović, & Suljić, 2014). Penelitian ini menggunakan data akademik dan personal siswa menengah pada nilai mata pelajaran matematika di Portugal. Tujuan penelitian ini adalah mempelajari bagaimana *bayesian networks* diterapkan agar dapat menentukan keterkaitan antar variabel dan mengetahui variabel apa saja yang berpengaruh, lalu digunakan sebagai teknik seleksi fitur. Selanjutnya variabel yang relevan dipilih untuk digunakan prediksi performa akademik siswa dalam mata pelajaran matematika dengan algoritma klasifikasi *machine learning*.

Topik penelitian tentang prediksi performa akademik siswa pada mata pelajaran matematika sebelumnya dilakukan oleh Paulo Cortez dan Alice Silva dengan beberapa teknik klasifikasi *machine learning* (Cortez, Silva, Trees, & Forest, 2008). Pada penelitian tersebut dilakukan komparasi akurasi prediksi pada tiga pola eksperimen, peneliti memberikan saran untuk menerapkan

teknik seleksi fitur untuk menyeleksi variabel yang relevan sehingga bisa meningkatkan perfoma akurasi prediksi.

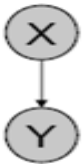
Pada penelitian sebelumnya, prediksi performa akademik siswa pada mata pelaajaran matematika telah dilakukan dengan penerapan teknik seleksi fitur *Information Gain* dan algoritma klasifikasi *machine learning*. Hasil penelitian menunjukkan bahwa dengan 10 variabel yang terseleksi melalui teknik seleksi *Information Gain* terbukti dapat meningkatkan tingkat akurasi prediksi algoritma klasifikasi *machine learning* (Sari, 2016). Ramaswami dan Rathinasabapthu menerapkan tiga teknik seleksi fitur dan metode *bayesian networks* untuk memprediksi performa akademik siswa (Ramaswami & Rathinasabapathy, 2012). Beberapa algoritma klasifikasi *machine learning* juga diterapkan Ramesh dkk untuk meneliti variabel apa saja yang relevan yang berhubungan dengan performa hasil ujian akhir siswa dengan terlebih dahulu menerapkan lima teknik seleksi fitur (Ramesh et al., 2013).

Bayesian Networks (BN) merupakan representasi grafis dari distribusi probabilitas, yaitu berupa *Directed Acyclic Graph* (DAG) yang terdiri dari satu set *node* untuk mewakili variabel dan satu set *link* diarahkan untuk menghubungkan pasang *node*. Setiap *node* memiliki distribusi probabilitas bersyarat yang mengkuantifikasi hubungan probabilistik antara *node* dan induknya (Aminian et al., 2014).

Bayesian networks menerapkan sifat Markov dimana dapat secara eksplisit menunjukkan independen bersyarat dalam distribusi probabilitas. Independen bersyarat seperti $A \perp C|B$ menunjukkan bahwa nilai B menghalangi informasi tentang C yang relevan dari A. Konsep ini tidak hanya berlaku pada sepasang *node*, tapi juga untuk sekumpulan *node*. Untuk dapat menentukan apakah suatu *node* independen terhadap *node* yang lain digunakan algoritma *d-separation*.

D-separation dapat mengidentifikasi hubungan independen bersyarat yang terdapat pada graf, yaitu di mana antara *node* X dan *node* Y yang dipisahkan di dalam graf G dan diberikan *node* Z, dan tidak ada jalur aktif antara setiap *node* X dan *node* Y di dalam graf G tersebut. *D-separation* dinotasikan sebagai berikut : $\mathbf{d\text{-}sep_G\ (X;Y|Z)}$ jika tidak ada jalur aktif antara *node* $X \in G$ dan *node* $Y \in G$ di dalam graf G (Korb & Nicholson, 2011).

Hubungan antar *node* dalam graf *bayesian networks* dibagi menjadi dua macam, yaitu hubungan langsung dan hubungan tidak langsung. Hubungan langsung adalah ketika *node* X dan *node* Y langsung berhubungan melalui anak panah, $X \rightarrow Y$. Pada gambar 1 ditunjukkan contoh hubungan langsung antara *node* X dan *node* Y.

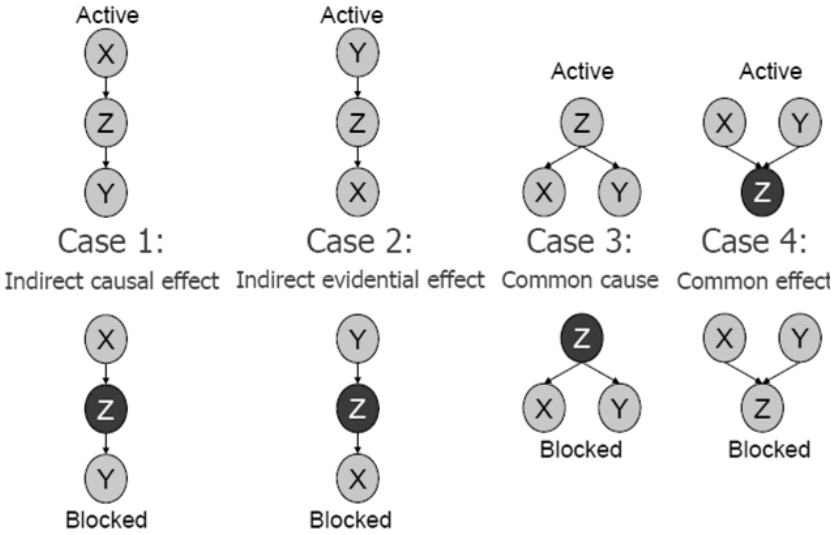


Gambar 1. Hubungan Langsung antara *node* X dan *node* Y

Sedangkan hubungan tidak langsung antar *node* adalah saat *node* X dan *node* Y tidak langsung terhubung, tetapi ada jalur yang menghubungkan antara kedua *node* tersebut dalam graf. Ada empat macam keadaan yang termasuk dalam hubungan tidak langsung (Koller & Friedman, 2009) :

a. *Indiret causal effect* atau efek sebab akibat tidak langsung, yaitu saat ada jalur aktif yang menghubungkan dari *node* X menuju *node* Y, tetapi ada *node* Z di antara kedua *node* tersebut. Kondisi dapat dikatakan ada jalur aktif ketika node Z belum diketahui nilai datanya, sebaliknya jika *node* Z sudah diketahui nilai datanya maka jalur antara *node* X dan *node* Y terhalangi.
b. *Indiret evidential effect* atau efek penting tidak langsung, yaitu saat ada jalur aktif yang menghubungkan dari *node* Y menuju *node* X, tetapi ada *node* Z di antara kedua *node* tersebut. Kondisi dapat dikatakan ada jalur aktif ketika *node* Z belum diketahui nilai datanya, sebaliknya jika *node* Z sudah diketahui nilai datanya maka jalur antara *node* X dan *node* Y terhalangi.
c. *Common cause* atau sebab umum, yaitu saat ada jalur aktif dari *node* Z yang menghubungkan ke *node* X dan *node* Y. Kondisi dapat dikatakan ada jalur aktif jika *node* Z belum diketahui nilai datanya, sebaliknya jika *node* Z sudah diketahui nilai datanya maka jalur antara *node* X dan *node* Y terhalangi.
d. *Common effect* atau efek umum, yaitu saat ada jalur dari *node* X dan *node* Y yang mengarah ke *node* Z, sebagai indikasi bahwa kedua *node* dapat mempengaruhi nilai dari *node* Z. Berbeda dengan 3 kondisi sebelumnya, kondisi pada jenis ini dapat dikatakan aktif jika nilai pada *node* Z sudah diketahui dan dikatakan tidak jalur aktif jika *node* Z belum diketahui nilai nilainya.

Berikut ini adalah gambar dari hubungan tidak langsung antara *node* X dan *node* Y.



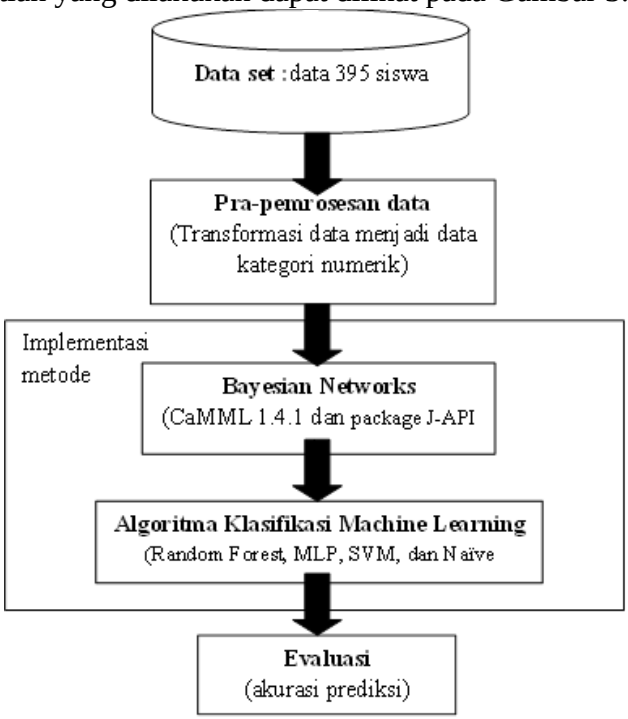
Gambar 2. Hubungan tidak langsung antara *node* X dan *node* Y

Karl W Kuschner menerapkan pendekatan *bayesian networks* untuk teknik seleksi fitur pada data *mass spectrometry*. Graf *bayesian networks* bisa mengorganisasi hubungan antar fitur atau atribut dan selanjutnya bisa digunakan untuk metode klasifikasi yang stabil. Graf *bayesian networks* menyediakan informasi tambahan tentang hubungan antar fitur yang sangat berguna untuk analisis (Kuschner et al., 2010) . Graf *bayesian networks* pada penelitian ini diterapkan untuk teknik seleksi fitur ditujukan untuk meningkatkan akurasi prediksi pada algoritma klasifikasi *machine learning*.

2. METODE PENELITIAN

Pada penelitian ini, data yang digunakan adalah data akademik dan personal siswa menengah pada nilai mata pelajaran matematika di Portugal. *Dataset* ini dikumpulkan oleh Paulo Cortez dan Alice Silva selama 2005-2006 dari dua sekolah umum melalui berkas laporan sekolah dan pengisian kuisioner. Dataset terdiri dari 395 data sampel dengan 33 variabel untuk evaluasi performa akademik siswa pada mata pelajaran matematika. Penelitian ini sebagai tahap awal pengujian terkait teknik *bayesian networks* bisa digunakan untuk studi kasus prediksi perfoma akademik siswa pada mata pelajaran matematika.

Prosedur atau langkah penelitian yang dilakukan pada penelitian ini terdiri dari 4 tahap, yaitu tahap pra pemrosesan data, implementasi metode *bayesian networks*, tahap klasifikasi dengan menggunakan algoritma klasifikasi *machine learning*, lalu tahap evaluasi. Adapun prosedur atau langkah penelitian yang dilakukan dapat dilihat pada Gambar 3.



Gambar 3. Prosedur/langkah penelitian

Tahap pra pemrosesan data dilakukan dengan melakukan transformasi data menjadi data kategori numerik. Dataset yang digunakan dalam penelitian ini berisi beberapa variabel dengan data kategori dalam bentuk string, dan beberapa variabel dengan data numerik kontinyu. Agar bisa dilanjutkan menuju fase pembentukan struktur graf *bayesian networks* maka dilakukan fase transformasi data. Variabel dengan data kontinyu dihitung nilai rata-ratanya, lalu dijadikan batas untuk mengkategorikan. Apabila data numerik kontinyu di bawah rata-rata maka diubah menjadi kategori 1, sedangkan data numerik kontinyu di atas rata-rata maka diubah menjadi kategori 2. Untuk variabel target hasil performa akademik pada mata pelajaran matematika dikategorikan menjadi dua (*binary classification*) yaitu lulus jika G3 ≥ 10 dan gagal jika G3 < 10.

Tahap implementasi metode dalam penelitian ini terdiri dalam dua tahap metode, yaitu implementasi metode *bayesian networks* dan proses klasifikasi dengan algoritma klasifikasi *machine learning*. Eksperimen dengan metode *bayesian networks* menggunakan *Causal discovery via MML* (CaMML) untuk membentuk struktur graf *bayesian networks*. Struktur graf *bayesian networks* yang divisualisasikan menggunakan software CaMML versi 1.4.1 dan *package* J-API.

Struktur graf *bayesian networks* dalam tahap ini digunakan sebagai tahap *feature selection*. Setelah itu, proses klasifikasi dengan algoritma klasifikasi *machine learning* dilakukan untuk memprediksi performa akademik siswa pada pelajaran matematika. Algoritma klasifikasi *machine learning* yang diterapkan di antaranya Random Forest, MLP, SVM, dan Naïve Bayes. Eksperimen ini dilakukan menggunakan *data mining library* pada Java dalam lingkungan Weka (Witten,IH. & Frank, E.,2000). Eksperimen menggunakan skenario 10 *fold cross validation.*

Tahap terakhir adalah tahap evaluasi, yaitu mengevaluasi tingkat akurasi sebelum dan sesudah melalui pemodelan struktur graf *bayesian networks*. Hasil evaluasi prediksi performa akademik siswa akan bisa dilihat menggunakan *confusion matrix*. Tingkat akurasi serta tingkat kesalahan prediksi dapat menunjukkan kinerja algoritma klasifikasi. Confusion matrix terdiri dari 4 bagian, yaitu *true positive, true negative, false positive dan false negative* (Gorunescu,2011). *Confusion Matrix* dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

		Prediksi	
		Positive	*Negative*
Aktual	*Positive*	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

Akurasi adalah jumlah prediksi benar dari semua data yang diprediksi, yaitu dapat dihitung dengan rumus :

$$akurasi\ (\%) = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

3. HASIL PENELITIAN DAN PEMBAHASAN

Eksperimen tahap pertama yaitu implementasi metode *bayesian networks* membentuk struktur graf *bayesian networks*. Hasil yang didapatkan melalui program

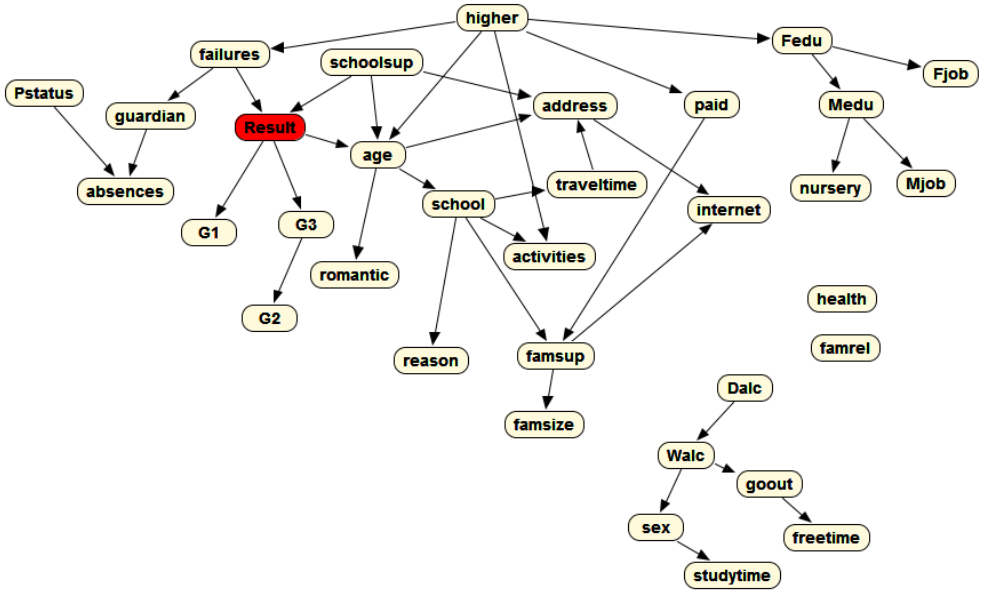
CaMML ada sebelas jenis struktur Graf *Bayesian Networks* yang digunakan untuk langkah menyeleksi variabel yang relevan. Hasil seleksi fitur berdasarkan 11 skenario dapat dilihat pada Tabel 2.

Tabel 2. Hasil Struktur Graf *Bayesian Networks*

Skenario	Variabel yang independen = variabel dieliminasi	jumlah *feature*
skenario 1	*famrel, health*	32
skenario 2	*famrel, health, reason*	31
skenario 3	*famrel,health, famsize, romantic*	30
skenario 4	*famrel,heath, Pstatus, reason, famsize*	29
skenario 5	*famrel,health, nursery*	31
skenario 6	*sex, study time, Dalc, Walc, go out, freetime, famrel, health*	25
skenario 7	*famrel, health, Pstatus, famsize*	30
skenario 8	*famrel,health, Pstatus*	31
skenario 9	*famrel,health, reason, nursery*	30
skenario 10	*famrel,health, Pstatus, activities*	30
skenario 11	*famrel, health, activities, famsize, Pstatus*	29

Setiap skenario menghasilkan struktur graf *bayesian networks* yang berbeda, masing-masing graf menunjukkan adanya keterkaitan antar variabel baik hubungan langsung atau hubungan tidak langsung. Hasil pada tahap ini, struktur graf *bayesian networks* digunakan untuk menentukan variabel apa saja yang relevan terhadap variabel performa akademik siswa sebagai variabel target. Untuk dapat menentukan apakah suatu *node* variabel itu independen terhadap *node* yang lain digunakan algoritma *d-separation*. Pada tabel 2 dapat dilihat daftar variabel yang independen yang akhirnya akan digunakan untuk mengeliminasi variabel atau bisa disebut dengan *feature selection*. Didapatkan hasil yang paling banyak mengeliminasi variabel adalah skenario eksperimen ke-6, dimana menyisakan 25 variabel dari 34 keseluruhan variabel yang digunakan.

Graf struktur *bayesian networks* pada skenario 6 terdapat pada Gambar 4.



Gambar 4. Struktur Graf *Bayesian Networks*

Pada Gambar 4 dapat dilihat bahwa ada 8 variabel yang *independent* atau tidak mempunyai jalur aktif untuk terhubung dengan variabel hasil. Kedelepan variabel itu adalah *sex, study time, Dalc, Walc, go out, freetime, famrel, health,* yang akan dieliminasi dan selanjutnya 25 variabel yang lain digunakan untuk prediksi hasil performa akademik siswa. Hasil akurasi dengan algoritma klasifikasi *Machine Learning* dapat dilihat pada Tabel 3.

Tabel 3. Hasil akurasi dengan algoritma klasifikasi *Machine Learning*

Skenario	Naïve Bayes	MLP	Random Forest	SMO
Sebelum (34 variabel)	88.66	93.62	99.22	92.25
skenario 1	88.99	92.73	99.42	92.3
skenario 2	88.71	93.65	99.52	92.78
skenario 3	88.78	93.97	99.72	92.51
skenario 4	88.71	94.58	99.47	92.25
skenario 5	88.9	94.08	99.47	92.61
skenario 6	89.19	94.46	99.87	92.61
skenario 7	88.84	93.95	99.65	92.63
skenario 8	88.91	93.7	99.72	92.2
skenario 9	88.68	95.04	99.67	92.2
skenario 10	88.91	94.81	99.59	92.38
skenario 11	88.86	94.71	99.8	92.76

Pada eksperimen tahap dua dilakukan uji model hasil struktur graf *bayesian networks*, yaitu mengeliminasi variabel independen terlebih dahulu

lalu mengimplementasikan empat algoritma klasifikasi *machine learning*. Eksperimen tahap kedua ini menggunakan 10 *fold cross validation* untuk masing-masing uji akurasi pada algoritma klasifikasi *machine learning*. Pada tabel 2 dapat dilihat bahwa adanya peningkatan performa akurasi prediksi pada empat algoritma klasifikasi *machine learning*. Pada skenario 6 yang menggunakan 25 variabel menunjukkan performa yang terbaik, dimana peningkatan akurasi maksimal pada algoritma *Naïve Bayes* dan *Random Forest*, yaitu 89.19% dan 99,87%. Algoritma *Neural Networks* (*Multiple Layer Perceptron*) menujukkan performa terbaiknya pada skenario ke-9. Sedangkan algoritma SVM (SMO) mencapai performa akurasi terbaik pada skenario 2, yaitu dengan tingkat akurasi 92.78%.

Hasil eksperimen menunjukkan ada 25 variabel yang relevan untuk memprediksi perfoma akademik siswa pada mata pelajaran matematika adalah umur, sekolah, alamat, status ortu, pendidikan ibu, pekerjaan ibu, pendidikan ayah, pekerjaan ayah, pendamping belajar anak (ibu/ayah/orang lain), jumlah anggota keluarga, alasan memilih sekolah, waktu perjalanan dari rumah ke sekolah, jumlah kegagalan dalam kelas, keikutsertaan kegiatan ekstrakulikuler, dukungan keluarga, ikut les tambahan, tambahan kelas di sekolah, ada akses internet di rumah, kehadiran ortu ke sekolah, motivasi studi lanjut, hubungan dengan lawan jenis (*romatic relationship*), jumlah ketidakhadiran (absen) kelas, nilai ujian matematika periode pertama, kedua dan ujian akhir matematika. Hasil penelitian ini bisa dijadikan rekomendasi pemilihan variabel untuk melakukan penelitian yang sejenis dengan menyebarkan kuisioner untuk siswa di Indonesia pada mata pelajaran matematika.

4. SIMPULAN

Kesimpulan yang didapatkan dari penelitian ini adalah struktur graf *bayesian networks* diterapkan untuk menentukan keterkaitan antar variabel dan mengetahui variabel apa saja yang berpengaruh terhadap variabel target, yaitu variabel performa akademik siswa pada pelajaran matematika. Untuk dapat menentukan apakah suatu *node* variabel itu independen terhadap *node* yang lain digunakan algoritma *d-separation*. Eksperimen untuk prediksi menggunakan algoritma klasifikasi *machine learning* menggunakan 10 *fold cross validation*. Hasil eksperimen menunjukkan adanya peningkatan performa akurasi prediksi pada empat algoritma klasifikasi *machine learning*. Hasil penelitian menunjukkan ada 25 variabel yang efektif untuk memprediksi perfoma akademik siswa pada mata pelajaran matematika.

5. DAFTAR PUSTAKA

Aminian, M., Couvin, D., Shabbeer, A., Hadley, K., Vandenberg, S., Rastogi, N., & Bennett, K. P. (2014). Predicting Mycobacterium tuberculosis Complex Clades Using Knowledge-Based Bayesian Networks, *2014*.

Cortez, P., Silva, A., Trees, D., & Forest, R. (2008). Using Data Mining to Predict Secondary School Student Performance, *2003*(2000).

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian Artificial Intelligence* (second edi). CRC Press.

Kuschner, K. W., Malyarenko, D. I., Cooke, W. E., Cazares, L. H., Semmes, O. J., & Tracy, E. R. (2010). A Bayesian network approach to feature selection in mass spectrometry data.

Sari, B. N. (2016). Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning untuk Prediksi Perfomasi Akademik Siswa, 6–7.

Transition, T., Osmanbegović, E., & Suljić, M. (2014). Determining Dominant Factor for Students Performance Prediction by Using Data Mining, *XVII*, 147–158.