

**PERBANDINGAN ANALISIS KLASIFIKASI ANTARA DECISION TREE
DAN SUPPORT VECTOR MACHINE MULTICLASS
UNTUK PENENTUAN JURUSAN PADA SISWA SMA
(Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)**



SKRIPSI

Oleh:

Rizky Ade Putranto

NIM : 24010210141018

**JURUSAN STATISTIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
SEMARANG**

2015

**PERBANDINGAN ANALISIS KLASIFIKASI ANTARA DECISION TREE
DAN SUPPORT VECTOR MACHINE MULTICLASS
UNTUK PENENTUAN JURUSAN PADA SISWA SMA
(Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)**



SKRIPSI

Oleh:

Rizky Ade Putranto

NIM : 24010210141018

**JURUSAN STATISTIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
SEMARANG**

2015

HALAMAN PENGESAHAN I

Judul : Perbandingan Analisis Klasifikasi Antara *Decision Tree* dan *Support Vector Machine Multiclass* Untuk Penentuan Jurusan Pada Siswa SMA
(Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)

Nama : Rizky Ade Putranto

NIM : 24010210141018

Telah diujikan pada sidang Tugas Akhir tanggal 24 Agustus 2015 dan dinyatakan lulus pada tanggal 29 September 2015.

Semarang, 29 September 2015

Mengetahui,

Ketua Program Studi Statistika

Jurusan Statistika



Dra. Dwi Ispriyanti, M.Si

NIP. 195709141986032001

Panitia Penguji Ujian Tugas Akhir

Ketua,

Yuciana Wilandari, S.Si, M.Si

NIP. 197005191998022001

HALAMAN PENGESAHAN II

Judul : Perbandingan Analisis Klasifikasi Antara Decision Tree dan Support Vector Machine Multiclass Untuk Penentuan Jurusan Pada Siswa SMA (Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)

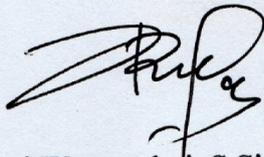
Nama : Rizky Ade Putranto

NIM : 24010210141018

Telah diujikan pada Sidang Tugas Akhir tanggal 24 Agustus 2015.

Semarang, 29 September 2015

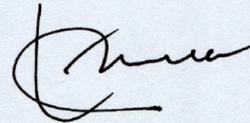
Pembimbing I



Triastuti Wuryandari, S.Si, M.Si.

NIP. 197109061998032001

Pembimbing II



Drs. Sudarno, M.Si.

NIP. 196407091992011001

KATA PENGANTAR

Puji syukur penulis panjatkan atas hadirat Allah SWT yang telah memberikan rahmat dan karunia-Nya, sehingga Laporan Tugas Akhir ini terselesaikan. Laporan Tugas akhir yang berjudul **“Perbandingan Analisis Klasifikasi Antara *Decision Tree* dan *Support Vector Machine Multiclass* Untuk Penentuan Jurusan Pada Siswa SMA (Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)”** dapat terselesaikan. Pada kesempatan ini penulis ingin mengucapkan terima kasih kepada :

1. Ibu Dra. Dwi Ispriyanti, M.Si selaku Ketua Jurusan Statistika FSM Universitas Diponegoro Semarang.
2. Ibu Triastuti Wuryandari, S.Si, M.Si selaku Dosen Pembimbing I dan Bapak Drs. Sudarno, M.Si selaku Dosen Pembimbing II yang telah meluangkan waktu memberikan masukan, motivasi, bimbingan dan pengarahan kepada penulis.
3. Bapak/Ibu Dosen dan teman-teman mahasiswa Statistika Undip yang telah memberikan motivasi dan dukungan kepada penulis.
4. Semua pihak yang tidak dapat disebutkan satu per satu.

Penulis menyadari bahwa Laporan Tugas Akhir ini belumlah sempurna. Oleh karena itu, kritik dan saran yang bersifat membangun sangat penulis harapkan. Penulis berharap semoga Laporan Tugas Akhir ini bisa membawa manfaat bagi penulis sendiri khususnya maupun bagi para pembaca pada umumnya.

Semarang, 29 September 2015

Penulis

ABSTRAK

Data mining adalah proses yang mempekerjakan satu atau lebih teknik *Machine Learning* untuk menganalisis dan mengekstraksi pengetahuan secara otomatis. Analisis klasifikasi data mining adalah menentukan sebuah *record* data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya, disebut juga dengan *Supervised Learning*. Klasifikasi *Decision Tree* merupakan salah satu teknik terkenal dalam data mining dan merupakan salah satu metode yang populer dalam menentukan keputusan suatu kasus dimana dalam proses metodenya diperoleh kriteria *entropy*, *information gain* dan *gain ratio*. Klasifikasi *Support Vector Machine Multiclass* (SVMM) dikenal sebagai teknik *Machine Learning* paling mutakhir menangani kasus multi kelas dimana output dari himpunan data memiliki lebih dari dua kelas atau kategori. Penulisan Tugas Akhir ini bertujuan untuk membandingkan tingkat ketepatan dan laju *error* klasifikasi *Decision Tree* dan SVMM untuk prediksi penjurusan siswa SMA di SMA Negeri 1 Jepara. Total akurasi sebesar 88,57% dan laju *error* 11,43% untuk klasifikasi *decision tree* dan total akurasi sebesar 87,14% dan laju *error* 12,86% untuk klasifikasi SVMM.

Kata kunci : Data Mining, *Machine Learning*, *Supervised Learning*, *Decision Tree*, *Support Vector Machine Multiclass*.

ABSTRACT

Data mining is a process that employs one or more of Machine Learning techniques to analyze and extract knowledge automatically. Analysis of data mining is to determine the classification of a new data record into one of several categories that have been defined previously, also known as Supervised Learning. Classification Decision Tree is one of the well-known technique in data mining and is one of the popular methods in the decision making process of a case in which the method is obtained entropy criteria, information gain and gain ratio. Classification Support Vector Machine Multiclass (SVMM) is known as the most advanced machine learning techniques to handle multi-class case where the output of the data set has more than two classes or categories. This final project aims to compare the level of accuracy and error rate of Decision Tree classification and prediction majors SVMM for high school students at SMAN 1 Jepara. The total accuracy of 88,57% and 11,43% error rate for the classification decision tree and the total accuracy of 87,14% and the error rate for the classification SVMM 12,86%.

Keywords : Data Mining, Machine Learning, Supervised Learning, Decision Tree, Support Vector Machine Multiclass

DAFTAR ISI

	Halaman
HALAMAN JUDUL	i
HALAMAN PENGESAHAN I	ii
HALAMAN PENGESAHAN II	iii
KATA PENGANTAR	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR TABEL	ix
DAFTAR GAMBAR	xi
DAFTAR LAMPIRAN.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Batasan Masalah	6
1.4 Tujuan	6
BAB II TINJAUAN PUSTAKA	7
2.1 Konsep Dasar Data Mining	7
2.2 Operasi Data Mining	10
2.3 Permasalahan dalam Data Mining	11
2.4. Teknik Data Mining	12
2.5 Klasifikasi (<i>Classification</i>).....	13
2.6 Konsep <i>Decision Tree</i>	17
2.6.1 Algoritma C 4.5.....	20
2.6.2 Memilih Atribut <i>Decision Tree</i>	21
2.6.3 Kriteria Pemilihan Atribut	22
2.7 <i>Support Vector Machine</i>	23
2.7.1 <i>Hyperplane</i>	24
2.7.2 Klasifikasi <i>Linear Separable</i>	24
2.7.3 Klasifikasi <i>Linear Non-Separable</i>	28
2.7.4 Klasifikasi <i>Non Linear</i>	30

2.8 Konsep SVM (Support Vector Machine Multiclass)	32
2.8.1 Metode Satu Lawan Semua	33
2.8.2 Metode Satu Lawan Satu	34
2.8.3 Max Voting	35
2.9 Pengukuran Uji Ketepatan Klasifikasi	36
BAB III METODOLOGI	38
3.1 Sumber Data	38
3.2 Metode Analisis	38
3.2.1 Tahap Pengumpulan Data Sekunder.....	38
3.2.2 Melakukan Klasifikasi Menggunakan <i>Decision Tree</i>	39
3.2.3 Melakukan Klasifikasi Menggunakan SVM.....	39
3.2.4 Melakukan Analisis Perbandingan.....	40
3.3 <i>Flowchart</i>	41
BAB IV HASIL DAN PEMBAHASAN	42
4.1 Analisis Klasifikasi <i>Decision Tree</i>	42
4.1.1 Menyiapkan Data <i>Training</i>	42
4.1.2 Menentukan Akar Pohon Keputusan	42
4.1.3 Klasifikasi Penjurusan Menggunakan <i>Decision Tree</i>	58
4.2 Analisis Klasifikasi Metode SVM	60
4.2.1 Pembentukan Persamaan Menggunakan Fungsi	
<i>Kernel Linier</i> untuk Pembuatan GUI	61
4.2.2 Fungsi <i>Kernel Polynomial</i> untuk Perhitungan Akurasi	
dan Laju <i>Error</i>	61
4.3 Membuat Aplikasi Klasifikasi SVM Berbasis <i>text</i> dan GUI.....	64
4.4 Analisis Hasil Perbandingan	66
BAB V KESIMPULAN	72
DAFTAR PUSTAKA	73
LAMPIRAN	75

DAFTAR TABEL

	Halaman
Tabel 1. Matriks Konfusi untuk Klasifikasi Tiga Kelas	36
Tabel 2. Posisi v Untuk Pemecahan Fitur Nilai IPA	43
Tabel 3. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Training</i> Simpul Akar “Nilai IPA”	45
Tabel 4. Posisi v Untuk Pemecahan Fitur Nilai IPS.....	46
Tabel 5. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Training</i> Simpul Akar “Nilai IPS”	48
Tabel 6. Posisi v Untuk Pemecahan Fitur Nilai Bahasa.....	48
Tabel 7. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Training</i> Simpul Akar “Nilai Bahasa”	50
Tabel 8. Posisi v Untuk Pemecahan Fitur Nilai IPA	52
Tabel 9. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Testing</i> Simpul Akar “Nilai IPA”.....	54
Tabel 10. Posisi v Untuk Pemecahan Fitur Nilai IPS	54
Tabel 11. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Testing</i> Simpul Akar “Nilai IPS”	55
Tabel 12. Posisi v Untuk Pemecahan Fitur Nilai Bahasa	56
Tabel 13. Hasil Perhitungan <i>Entropy</i> dan <i>Gain</i> Data <i>Testing</i> Simpul Akar “Nilai Bahasa”	57
Tabel 14. Hasil Prediksi Klasifikasi Data <i>Training Decision Tree</i>	66
Tabel 15. Hasil Prediksi Klasifikasi Data <i>Testing Decision Tree</i>	67
Tabel 16. Hasil Prediksi Klasifikasi Data <i>Training SVM</i>	67

Tabel 17. Hasil Prediksi Klasifikasi Data <i>Testing</i> SVM.....	68
Tabel 18. Perbandingan Data <i>Training</i>	69
Tabel 19. Perbandingan Data <i>Testing</i>	70

DAFTAR GAMBAR

	Halaman
Gambar 1. Data Mining dan Teknologi Database Lainnya.....	8
Gambar 2. Proses KDD (<i>Knowledge Discovery in Database</i>).....	10
Gambar 3. Proses Pekerjaan Klasifikasi	13
Gambar 4. Klasifikasi Dua Kelas.....	16
Gambar 5. Klasifikasi Multi Kelas	16
Gambar 6. <i>Decision Tree</i> untuk Klasifikasi Data Nilai Siswa	18
Gambar 7. Penelusuran <i>Decision Tree</i> untuk Menentukan Kelas Dari Objek Nilai Siswa.....	19
Gambar 8. <i>Hyperplane</i> pada SVM	24
Gambar 9. Klasifikasi SVM untuk Klasifikasi 3 Kelas	33
Gambar 10. <i>Flowchart</i> Analisis Klasifikasi.....	41

DAFTAR LAMPIRAN

	Halaman
Lampiran 1. Data Penjurusan SMA	75
Lampiran 2. Data <i>Training</i> Klasifikasi Nilai Penjurusan SMA.....	85
Lampiran 3. Data <i>Testing</i> Klasifikasi Nilai Penjurusan SMA.....	89
Lampiran 4. Pembentukan <i>Decision Tree</i> dengan Matlab	90
Lampiran 5. Output <i>Decision Tree</i> Data <i>Training</i>	91
Lampiran 6. Output <i>Decision Tree</i> Data <i>Testing</i>	93
Lampiran 7. Aplikasi GUI Berbasis <i>Decision Tree</i>	95
Lampiran 8. Data dan Group Data <i>Training</i> SVM	97
Lampiran 9. Data dan Group Data <i>Testing</i> SVM	99
Lampiran 10. Algoritma SVM Data <i>Training</i> Menggunakan <i>Software</i> Matlab	101
Lampiran 11. Algoritma SVM Data <i>Testing</i> Menggunakan <i>Software</i> Matlab	104
Lampiran 12. Sintaks Klasifikasi M-File	107
Lampiran 13. Rancangan GUI Berbasis SVM	108
Lampiran 14. <i>Sintaks Callback</i>	109

BAB I

PENDAHULUAN

1.1 Latar Belakang

Setiap lembaga mempunyai sistem operasional yang setiap transaksi kegiatan operasinya selalu dicatat dan didokumentasikan. Pendokumentasian setiap transaksi sangat berguna bagi lembaga tersebut untuk segala keperluan. Data-data tersebut tersimpan dalam sebuah basis data berkapasitas besar. Bagi suatu perusahaan, data-data yang tersimpan di basis data dapat dimanfaatkan untuk membuat laporan penjualan, kontrol inventaris, dan sebagainya, yang pada akhirnya dapat digunakan untuk mengetahui kondisi keuangan perusahaan. Dalam dunia pendidikan, proses penentuan bidang minat siswa juga menjadi hal yang penting untuk dapat diputuskan dengan benar. Faktor-faktor penentu masuknya siswa dalam sebuah bidang minat menjadi data awal yang dapat digunakan untuk membantu penentuan keputusan. Dalam dunia kesehatan, diagnosis penyakit pasien menjadi hal yang sangat sulit dilakukan. Namun demikian, catatan rekam medis telah menyimpan gejala-gejala penyakit pasien dan diagnosis penyakitnya. Hal seperti itu tentu sangat berguna bagi para dokter muda. Mereka dapat menggunakan catatan rekam medis yang sudah ada sebagai bantuan untuk mengambil keputusan tentang diagnosis penyakit pasien. Beberapa lembaga yang mempunyai sistem operasional data yang tersimpan dalam sebuah basis data kapasitas besar, memicu munculnya suatu ilmu yang dapat digunakan memecahkan banyak kasus, yaitu data mining (Prasetyo, 2012).

Munculnya data mining didasarkan pada kenyataan bahwa jumlah data yang tersimpan dalam basis data semakin besar. Data mining sendiri berisi pencarian trend atau pola tertentu yang diinginkan dalam basis data yang besar untuk membantu pengambilan keputusan di waktu yang akan datang. Harapannya, perangkat data mining mampu mengenali pola-pola ini dalam data dengan masukan yang minimal. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa data yang berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan lainnya. Data mining berhubungan dengan sub-area statistik yang disebut *Exploratory Data Analysis* (Analisis Data Eksplorasi) yang mempunyai tujuan sama dan bersandar pada ukuran statistik. Data mining berpotensi tinggi jika data yang tepat dikumpulkan dan disimpan dalam sebuah gudang data (*data warehouse*). Sebuah gudang data merupakan suatu sistem manajemen basis data relasional yang didesain khusus untuk memenuhi kebutuhan sistem pengolahan transaksi. Basis data cenderung menjadi besar dan dinamis dalam hal isinya yang selalu berubah saat informasi ditambahkan, dimodifikasi atau dihapus. Permasalahan dalam hal ini dari sudut pandang data mining adalah bagaimana menjamin bahwa aturan-aturan tersebut *up-to-date* dan konsisten dengan informasi paling terkini. Juga sistem penemuan dipengaruhi oleh ketepatan waktu dari data tersebut. Beberapa aplikasi yang menggunakan data mining bermaksud menyelesaikan permasalahan dengan membangun model berdasarkan data yang sudah digali untuk diterapkan terhadap data yang lain. Secara umum terdapat dua jenis tipologi aplikasi data mining (Hermawati, 2013):

1. Metode prediksi, yang bermaksud memprediksi nilai yang akan datang berdasarkan data-data yang telah ada variabelnya seperti *Classification*, *Regression*, *Deviation Detection* dan lain-lain.
2. Metode deskriptif, yang bertujuan membantu user agar mudah melihat pola-pola yang berasal dari data yang ada seperti *Clustering*, *Association Rule Discovery*, *Sequential Pattern Discovery* dan lain-lain.

Analisis klasifikasi data mining adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya, disebut juga dengan *supervised learning*. Di dalam klasifikasi keluaran dari setiap data adalah bilangan bulat atau diskrit. Klasifikasi pertama kali diterapkan pada bidang tanaman yang mengklasifikasikan suatu spesies tertentu, seperti yang dilakukan oleh Carolus von Linne (atau dikenal dengan nama Carolus Linnaeus) yang pertama kali mengklasifikasikan spesies berdasarkan karakter fisik dan dikenal dengan bapak klasifikasi. Metode-metode yang telah dikembangkan oleh periset untuk menyelesaikan kasus klasifikasi, antara lain: Pohon keputusan (*Decision Tree*), *Naïve Bayes*, Jaringan Syaraf Tiruan, Analisis Statistik, Algoritma Genetik, *Rough Sets*, *k-Nearest Neighbour*, Metode Berbasis Aturan, *Memory Based Reasoning*, *Support Vector Machine* (Sumathi, 2006).

Berdasarkan beberapa metode klasifikasi tersebut, ingin dilakukan perbandingan analisis klasifikasi antara metode *Decision Tree* dan *Support Vector Machine Multiclass* untuk penentuan jurusan pada siswa SMA. Tingkat penguasaan ilmu pengetahuan dan teknologi erat kaitannya dengan kesejahteraan dan perekonomian suatu negara. Masyarakat yang berpendidikan, berwawasan, berbudi, dan terampil dapat membawa bangsanya menjadi negara yang maju dan

disegani bangsa lain. Oleh karena itu, pendidikan disebut-sebut sebagai salah satu landasan utama meraih impian tersebut. Generasi berpendidikan turut mempengaruhi angkatan kerja di sebuah negara, tanpa terkecuali Indonesia.

Menurut Badan Pusat Statistik (BPS), jumlah angkatan kerja di Indonesia pada Agustus 2014 tercatat 118,2 juta orang, berkurang 3 juta orang dari jumlah angkatan kerja pada Februari 2014. Pada Agustus 2014, penduduk bekerja pada jenjang pendidikan Sekolah Dasar ke bawah sebanyak 52 juta orang atau 44 persen, pada jenjang pendidikan SMP sebanyak 21,1 juta orang atau 17,85 persen, pada jenjang pendidikan SMA sebanyak 34,6 juta orang atau 29,27 persen, pada jenjang diploma 2,9 juta orang atau 2,45 persen, dan pendidikan universitas 7,6 juta orang atau 6,43 persen.

Dari data Badan Pusat Statistik tersebut dapat disimpulkan bahwa penduduk Indonesia yang bekerja pada jenjang pendidikan diploma dan pendidikan universitas masih sangat rendah. Hal ini dikarenakan banyak lulusan-lulusan dari pendidikan universitas yang bekerja di berbagai macam sektor pekerjaan kebanyakan tidak sesuai dengan bidang atau jurusan yang digeluti pada masa pendidikan di tingkat universitas. Hal ini menjadikan output lulusan dari berbagai universitas tidak sesuai harapan. Sumber dari masalah ini adalah berakar dari ketidaksesuaian minat dan bakat dari seorang individu terhadap penjurusan suatu pendidikan. Sebelum masuk ke jenjang pendidikan diploma dan universitas, seorang individu menempuh jenjang pendidikan di SMA. Di SMA inilah seorang individu sudah diarahkan kepada bidang minat dan bakat seseorang terhadap potensi yang dimiliki, diantaranya yang umum dikenal di masyarakat adalah penjurusan minat bidang pendidikan IPA (Sains), Pendidikan IPS (Sosial), dan

pendidikan Bahasa. Tujuan penjurusan ini adalah agar minat dan bakat seseorang dapat terarah dengan spesifik pada bidang ilmu tersebut yang selanjutnya dapat melanjutkan ke jenjang perguruan tinggi dan setelah itu memperoleh suatu pekerjaan yang sesuai minat dan bakat seseorang. Akhirnya tercapai suatu tujuan peningkatan penguasaan ilmu pengetahuan dan teknologi yang erat kaitannya dengan kesejahteraan dan perekonomian suatu negara. Fakta di lapangan masih banyak terdapat kesalahan dalam proses penjurusan pendidikan tersebut. Atas dasar itulah maka penulis ingin mengkaji aplikasi data mining untuk penentuan bidang minat pada siswa SMA dengan menyusun tugas akhir yang berjudul **“Perbandingan Analisis Klasifikasi Antara *Decision Tree* dan *Support Vector Machine Multiclass* Untuk Penentuan Jurusan Pada Siswa SMA (Studi Kasus Nilai Mata Pelajaran Pokok di SMA Negeri 1 Jepara)”**.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut dapat dirumuskan masalah sebagai berikut :

1. Bagaimana klasifikasi *decision tree* dihasilkan dari kriteria *entropy*, *information gain*, dan *gain ratio* terhadap penentuan jurusan siswa SMA menggunakan algoritma C4.5
2. Bagaimana klasifikasi SVM dihasilkan dari kriteria pendekatan satu lawan satu (*One against one*) terhadap penentuan jurusan siswa SMA
3. Bagaimana hasil model terbaik dari perbandingan klasifikasi *decision tree* dan SVM yang dapat digunakan untuk prediksi penjurusan terbaik

1.3 Batasan Masalah

Penulisan tugas akhir ini dibatasi pada pengujian data rata-rata nilai yang digunakan dalam penentuan penjurusan pendidikan IPA (Sains), IPS (Sosial), dan Bahasa pada tingkat pendidikan di SMA Negeri 1 Jepara. Metode pengujian data ini adalah berurutan dari *Decision Tree* dilanjutkan dengan *Support Vector Machine Multiclass* dan akhirnya dapat dilakukan perbandingan dari kedua metode data mining klasifikasi tersebut.

1.4 Tujuan

Adapun tujuan dari penulisan tugas akhir ini adalah sebagai berikut :

1. Mendapatkan kriteria *entropy*, *information gain*, dan *gain ratio* analisis klasifikasi *decision tree* menggunakan algoritma C 4.5
2. Mendapatkan kriteria SVM dengan pendekatan satu lawan satu (*One against one*)
3. Membandingkan tingkat ketepatan akurasi dan laju error klasifikasi *decision tree* dan SVM untuk mendapatkan model terbaik yang dapat digunakan untuk prediksi penjurusan terbaik