KLASIFIKASI KETEPATAN LAMA STUDI MAHASISWA MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST*

(Studi Kasus : Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Program Studi Statistika

Syauqi Amri Yahya

14611140

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan Metode *Support Vector Machine* Dan *Random Forest* (Studi Kasus: Studi Kasus : Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017)
Nama Mahasiswa	:	Syauqi Amri Yahya
Nomor Mahasiswa	:	14611140

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, Mei 2018

Pembimbing,

Ayundyah Kesumawati, S.Si, M.Si

HALAMAN PENGESAHAN

TUGAS AKHIR

KLASIFIKASI KETEPATAN LAMA STUDI MAHASISWA MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST*

(Studi Kasus : Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017)

Nama Mahasiswa : Syauqi Amri Yahya

Nomor Mahasiswa : 14611140

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL 14 MEI 2018

Nama Penguji	Tanda Tangan
1. Ir. Ali Parkhan, M.T.	.....
2. Tuti Purwaningsih, S.Stat., M.Si.	.....
3. Ayundyah Kesumawati, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

UNIVERSITAS ISLAM INDONESIA YOGYAKARTA FAKULTAS MIPA

Drs. Allwar, M.Sc. Ph.D.

KATA PENGANTAR

Assalamu'alaikum Warahmatullaahi Wabarakaatuh

Alhamdulillahirabbil'alamiin, puji syukur kehadirat Allah SWT yang telah memberikan hidayah, kesempatan, dan kemudahan kepada kita semua dalam menjalankan amanah yang menjadi tanggung jawab kita. Shalawat serta salam tak henti-hentinya kita panjatkan kepada junjungan kita Nabi Besar Muhammad SAW beserta seluruh keluarga dan sahabatnya, karena dengan *syafa'at*nya kita dapat hijrah dari zaman *jahiliyah* menuju zaman *islamiyah.*

Tugas akhir yang berjudul "**Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan Metode *Support Vector Machine D*an *Random Forest* (Studi Kasus : Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017)**" sebagai salah satu syarat untuk memperoleh gelar sarjana Jurusan Statistika di Universitas Islam Indonesia. Dalam penyusunan skripsi ini penulis banyak mengalami hambatan, namun berkat bantuan, bimbingan, dan kerjasama yang ikhlas dari berbagai pihak, akhirnya skripsi ini dapat terselesaikan dengan baik.

Pada kesempatan ini penulis mengucapkan terima kasih dengan tulus kepada :

1. Bapak Nandang Sutrisno, SH., M.Hum., LLM., Ph.D. selaku Rektor Universitas Islam Indonesia.
2. Bapak Drs. Allwar, M.Sc., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta beserta seluruh jajarannya.
3. Bapak Dr. RB. Fajriya Hakim, S.Si., M.Si. selaku Ketua Jurusan Statistika beserta seluruh jajarannya.

4. Ibu Ayundyah Kesumawati, S.Si., M.Si. selaku dosen pembimbing. Terimakasih atas waktu, tenaga, motivasi, ilmu, nasehat serta bimbingannya sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik.
5. Dosen-dosen Statistika Universitas Islam Indonesia yang telah mendidik dan menginspirasi.
6. Kedua orang tua yang telah memberikan kepercayaan kepada peneliti untuk kuliah dan telah memberikan do'a tulusnya sepanjang hari hingga bisa lulus jenjang sarjana ini.
7. Kedua kakak yang selalu memberikan dukungan, bantuan, dan nasihat selama ini untuk kuliah.
8. Sahabat-sahabatku semua yang selalu ada untuk penulis selama masa kuliah, atas kebersamaan dari awal kuliah hingga sekarang terimakasih atas segala hal yang pernah dilalui bersama.
9. Teman – teman seperjuanganku di Jurusan Statistika 2014, terimakasih atas kebersamaannya selama ini, menuntut ilmu bersama kalian adalah pengalaman yang tak akan pernah terlupakan.
10. Semua pihak yang telah membantu dalam penyusunan skripsi ini yang tidak dapat disebukan satu persatu

Semoga segala bantuan, bimbingan dan pengajaran yang telah diberikan kepada penyusun mendapatkan imbalan dari Allah SWT. Tidak lupa penulis memohon maaf apabila selama dalam proses penyusunan tugas akhir ini terdapat kekhilafan dan kesalahan. Penulis menyadari sepenuhnya akan keterbatasan kemampuan yang dimiliki. Oleh karena itu, penulis mengharapkan adanya kritik dan saran yang membangun demi kesempurnaan penyusunan dan penulisan tugas akhir ini. Semoga tugas akhir ini dapat bermanfaat bagi semua yang membaca dan membutuhkannya, Amin amin ya robbal 'alamiin.

Wassalamu'alaikum Warahmatullaahi Wabarakaatuh

Yogyakarta, Mei 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan disuatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain kecuali yang secara tertulis diacu dalam naskah ini dan diterbitkan dalam daftar pustaka.

Yogyakarta, Mei 2018

Penulis

Syauqi Amri Yahya

KLASIFIKASI KETEPATAN LAMA STUDI MAHASISWA MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST*

(Studi Kasus : Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017)

Syauqi Amri Yahya

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Setiap perguruan tinggi berusaha untuk terus memperbaiki manajemennya, supaya meningkatkan mutu pendidikan dan meningkatkan akreditasi. Salah satu elemen penilaian akreditasi perguruan tinggi adalah lulus tepat waktu. Selain itu, ketepatan lama studi mahasiswa merupakan isu yang penting karena ketepatan tersebut menjadi dasar efektifnya suatu perguruan tinggi. Universitas Islam Indonesia (UII) adalah salah satu perguruan tinggi swasta terkemuka di Indonesia. Sebagai universitas yang sudah cukup tua di Indonesia, UII sudah berakreditas-A dan sudah menghasilkan banyak alumni dari berbagai daerah dan latar belakang. Untuk terus meningkatkan universitas, UII tentunya harus mempertimbangkan juga aspek ketepatan waktu mahasiswanya untuk menempuh lama studi, karena itu merupakan salah satu aspek penilaian akreditasi dari BAN-PT. Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik tertentu. Analisis klasifikasi bisa digunakan untuk memprediksi bahwa seorang mahasiswa dikatakan lulus tepat waktu atau tidak. Metode Support Vector Machine (SVM) dan Random Forest adalah bagian dari metode klasifikasi. Analisis klasifikasi SVM dan Random Forest dilakukan dengan menggunakan data historis dari alumni UII tahun kelulusan 2000-2017. Tingkat akurasi SVM kernel RBF dengan nilai optimum C=1 dan gamma = 1 adalah 77%, akurasi SVM kernel sigmoid dengan nilai optimum C=10, dan gamma = 1 adalah 68%, dan akurasi Random Forest dengan nilai optimum m = 2 dan k = 500 adalah 80%. Oleh karena itu, metode terbaik untuk menentukan ketepatan lama studi mahasiswa UII adalah Random Forest.

Kata Kunci : *Perguruan Tinggi, UII, Klasifikasi, SVM, Random Forest.*

CLASSIFICATION OF GRADUATION RATE USING SUPPORT VECTOR MACHINE AND RANDOM FOREST METHOD

(Case Study: Data of Alumni Universitas Islam Indonesia Graduation Year 2000-2017)

Syauqi Amri Yahya

Department Statistics, Faculty of Mathematics and Natural Science

Islamic University of Indonesia

ABSTRACT

Each university strives to continue to improve its management, in order to improve the quality of education and improve accreditation. One of the elements of a college accreditation assessment is to pass on time. In addition, the timeliness of student studies is an important issue because it is the basis for the effectiveness of a university. Universitas Islam Indonesia (UII) is one of the leading private universities in Indonesia. As an old university in Indonesia, UII is already A-accredited and has produced many alumni from various regions and backgrounds. To continue to improve the university, UII must also consider the timeliness aspect of its students to take a long study, because it is one aspect of accreditation assessment from BAN-PT. Classification is a method to predict an event or an upcoming decision to be at a certain point. Classification analysis can be used to predict that a student is said to pass on time or not. Support Vector Machine (SVM) and Random Forest methods are part of the classification method. SVM and Random Forest classification analysis is done by using historical data from UII alumni of graduation year 2000-2017. SVM accuracy level of RBF kernel with optimum value C = 1 and gamma = 1 is 77%, SVM svm accuracy with optimum value C = 10, and gamma = 1 is 68%, and Random Forest accuracy with optimum value m = 2 and k = 500 is 80%. Therefore, the best method for determining the accuracy of the study duration of UII students is Random Forest.

Keywords: *College, UII, Classification, SVM, Random Forest.*

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan tinggi merupakan pendidikan jenjang untuk menuntut ilmu setelah pendidikan menengah (SMA/Sederajat) yang terdiri dari berbagai jenjang, yaitu jenjang diploma, sarjana, magister, doktor, dan program profesi. Badan yang menjalankan pendidikan tinggi disebut perguruan tinggi dan dikenal dengan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Bentuk perguruan tinggi ada bermacam-macam, seperti universitas, institut, sekolah tinggi, politeknik, spesialis, dan akademi. (Undang-Undang Republik Indonesia Nomor 12 pasal 1, 2012). Banyaknya perguruan tinggi pada era ini menggambarkan bahwa semakin pedulinya pemerintah dan masyarakat dalam menangani pendidikan. Kualitas dari perguruan tinggi diharapkan terus berkembang agar peranan pendidikan di Indonesia semakin sejahtera.

Salat satu tolak ukur kualitas pendidikan tinggi adalah bisa dilihat dari akreditasi yang diperoleh. Lembaga yang berwenang untuk menilai akreditasi perguruan tinggi adalah Badan Akreditasi Nasional-Perguruan Tinggi (BAN-PT). Setiap perguruan tinggi berusaha untuk terus memperbaiki mutu universitas dengan meningkatkan pendidikan dan akreditasinya. Salah satu elemen penilaian akreditasi perguruan tinggi adalah lulus tepat waktu. Selain itu, ketepatan lama studi mahasiswa merupakan isu yang penting karena ketepatan tersebut menjadi dasar efektifnya suatu perguruan tinggi (Untari, 2014). Ketepatan lama studi mahasiswa sudah lama diatur oleh pemerintah dalam peraturan menteri. Mahasiswa dikatakan tepat waktu jika menempuh masa studi sarjana/S1 maksimal selama 4 tahun atau kurang (PERMENDIKNAS No. 232, 2000).

Jumlah Perguruan Tinggi Negeri di Indonesia sebanyak 418 dan Jumlah Perguruan Tinggi Swasta sebanyak 4.222 (http://forlap.ristekdikti.go.id). Universitas Islam Indonesia (UII) adalah salah satu perguruan tinggi swasta terkemuka di Indonesia. Terinspirasi oleh semangat nasionalisme dan berpedoman

pada nilai-nilai perennial, UII didirikan satu bulan sebelum proklamasi kemerdekaan Indonesia pada tahun 1945 (https://www.uii.ac.id). Sebagai universitas yang sudah cukup tua di Indonesia, UII sudah berakreditas-A dan sudah menghasilkan banyak alumni dari berbagai daerah dan latar belakang. Untuk terus meningkatkan universitas, UII tentunya harus mempertimbangkan juga aspek ketepatan waktu mahasiswanya untuk menempuh lama studi, karena itu merupakan salah satu aspek penilaian akreditasi dari BAN-PT. Lama studi mahasiswa tentu bisa saja terjadi karena banyak hal atau faktor, misalnya saja dalam penelitian ini ada faktor pekerjaan orangtua dan pendidikan orang tua.

Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik. Klasifikasi merupakan suatu pekerjaan yang melakukan penilaian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012).

Berdasarkan latar belakang yang telah dijabarkan, peneliti merasa perlu melakukan suatu analisis klasifikasi untuk memprediksi bahwa seorang mahasiswa dikatakan lulus tepat waktu atau tidak berdasarkan data yang dimiliki oleh mahasiswa itu sendiri. Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik. Klasifikasi merupakan suatu pekerjaan yang melakukan penilaian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012). Metode *Support Vector Machine* (SVM) dan *Random Forest* adalah bagian dari metode klasifikasi. Metode analisis klasifikasi yang digunakan adalah SVM karena memiliki tingkat akurasi yang cukup tinggi dalam hal klasifikasi teks dan cocok untuk data yang memiliki dimensi besar (Naradhipa dan Purwarianti, 2011). Sedangkan *Random Forest* digunakan karena tingkat akurasi dari metode ini lebih tinggi dari pada metode *regresi* logistik dan *decision tree*, yang mana *Random Forest* akan membuat pohon klasifikasi dengan jumlah banyak dan kemudian digabungkan untuk mencari tingkat akurasi yang tinggi (Sartono dan Syafitri, 2010). *Random Forest* dianggap sebagai "obat mujarab" untuk klasifikasi data, karena saat menemukan data dan tidak tau harus digunakan metode apa, maka cobalah *random forest* (*https://www.analyticsvidhya.com*).

Berdasarkan penjelasan tersebut, peneliti menentukan untuk melakukan analisis klasifikasi menggunakan metode SVM dan *Random Forest*. Analisis SVM dan *Random Forest* memerlukan nilai parameter optimum yang ditentukan oleh peneliti. Pada metode SVM parameter yang digunakan adalah *cost* dan *gamma*, adapun nilai yang diuji coba untuk *cost* = 0.1, 1, 5, 10, dan 50, sedangkan untuk nilai *gamma* = 1, 2, 3, dan 4. Pada metode *random forest*, parameter yang digunakan adalah m (banyaknya variabel acak) dan k (banyaknya pohon percobaan), adapun nilai yang diuji cobakan untuk m = 2, 3, dan 6, sedangkan untuk nilai k = 25, 50, 100, 500, dan 1000. Analisis dilakukan dengan menggunakan data historis dari alumni UII tahun kelulusan 2000-2017.

1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan di atas, maka rumusan masalah yang dapat diidentifikasi dalam tugas akhir ini adalah:

1. Bagaimana gambaran secara umum tentang ketepatan lama studi dari alumni UII tahun kelulusan 2000-2017?
2. Berapa besar ketepatan metode klasifikasi menggunakan metode SVM dan *Random Forest* dalam memprediksi lama studi mahasiswa?
3. Manakah metode terbaik antara SVM dan *Random Forest* dalam melakukan prediksi ketepatan lama studi mahasiswa?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Penelitian dilakukan dengan menggunakan data historis alumni UII jenjang sarjana tahun kelulusan 2000-2017.
2. Metode yang digunakan untuk memprediksi ketepatan lama studi adalah metode SVM dan *Random Forest* menggunakan *software* RStudio.
3. Metode SVM menggunakan parameter *cost* (C) = 0,1 ; 1 ; 5 ; 10 ; dan 50, serta nilai parameter *gamma* = 1 ; 2 ; 3 ; 4.
4. Metode *Random Forest* menggunakan nilai *mtry* (m) = 2 ; 3 ; dan 6, serta nilai *ntree* (k) = 10 ; 50 ; 100 ; 500 ; dan 1.000.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengetahui gambaran secara umum mengenai ketepatan lama studi dari alumni UII tahun kelulusan 2000-2017.
2. Mengetahui seberapa tepat metode SVM dan *Random Forest* dalam memprediksi ketepatan lama studi dengan melihat akurasinya.
3. Melihat metode mana yang memiliki tingkat akurasi terbaik untuk memprediksi ketepatan lama studi.

1.5 Manfaat Penelitian

Adapun manfaat dilakukan penelitian ini adalah:

1. Mengetahui penerapan metode klasifikasi yang tepat untuk memprediksi ketepatan lama studi.
2. Dengan metode terbaik, dapat diterapkan untuk melakukan klasifikasi data mahasiswa agar bisa melihat/memprediksi seseorang lulus dengan tepat watu atau tidak tepat waktu.
3. Dengan mengetahui kemungkinan tidak tepat waktu, maka bisa dilakukan pencegahan dengan misalnya mengingatkan untuk giat belajar, selalu aktif dikelas saat kuliah, atau lebih fokus dalam kuliah.

.

BAB II
TINJAUAN PUSTAKA

Penelitian ini terbentuk karena telah membaca dan memahami tentang penelitian-penelitian sebelumnya yang memiliki keterkaitan dengan penelitian ini baik dari segi metode maupun objek penelitian. Penelitian terdahulu merupakan hal penting dalam membuat suatu penelitian. Penelitian terdahulu digunakan sebagai suatu kajian bagi penulis untuk mengetahui hubungan antara penelitian sebelumnya dengan penelitian yang dilakukan saat ini. Hal ini dilakukan untuk menghindari adanya unsur duplikasi. Tidak hanya itu, dengan menggunakan penelitian terdahulu, penelitian yang saat ini dilakukan memiliki arti penting sehingga dapat memberikan kontribusi pada perkembangan ilmu pengetahuan.

Penerapan *Support Vector Machine* dalam dunia medis oleh Rachman dan Purnami (2012) yang digunakan untuk mengetahui tingkat keganasan dari penyakit kanker payudara (*breast cencer*). Penelitian tersebut dilakukan karena di Indonesia adalah Negara berkembang yang penderita penyakit tersebut sangat banyak. Hal itu dibuktikan dengan data Sistem Informasi Rumah Sakit (SIRS) 2007 yang menunjukan kanker payudara menempati urutan pertama pada pasien rawat inap di seluruh rumah sakit Indonesia. Penelitian tersebut menggunakan pendekatan 2 metode, yaitu regresi logistik ordinal dan *support vector machine*. Hasil tingkat akurasi untuk menentukan keganasan kanker payudara untuk metode SVM memiliki akurasi yang lebih tinggi dari pada metode regresi logistik ordinal. Tingkat akurasi menggunakan SVM adalah sebesar 98,11% dan akurasi untuk regresi logistik ordinal hanya sebesar 56,60%. Tingkat akurasi tersebut didapatkan dengan melakukan uji coba untuk pembagian data *training* dan *testing* sebanyak 3, yaitu 50% *training* dan 50% *testing*, 75% *training* dan 25% *testing*, dan 80% *training* dan 20% *testing*. Selain itu, uji coba untuk mencari parameter C = 10, 100, dan 1000, serta nilai *gamma* = 1, 2, dan 3. Setelah uji coba, didaptlah bahwa pembagian data optimum di 75% *training* dan 25% *testing* menggunakan nilai parameter C=1000 dan nilai *gamma* =1.

Penelitian selanjutnya adalah penelitian oleh Octaviani, Wilandari, dan Ispriyanti (2014) dengan menggunakan metode SVM untuk menentukan klasifikasi akreditasi pada SD di Kota Malang. Data akreditasi yang digunakan adalah data akreditasi SD di Kota Malang tahun 2011 sampai 2013 yang didapat dari *website* resmi BAN-S/M dengan 3 status, yaitu Akreditasi A, Akreditasi B, dan Akreditas C. Sedangkan untuk variabel *independen* terdiri dari 8 komponen standar, yaitu Standar Isi (X1), Standar Proses (X2), Standar Kompetensi Lulusan (X3), Standar Pendidik dan Tenaga Kependidikan (X4), Standar Sarana dan Prasarana (X5), Standar Pengelolaan (X6), dan Standar penilaian pendidikan (X8). Penelitian tersebut menggunakan metode SVM dengen kernel *Radian Basis Function* (RBF) dan kernel *Polynomial*. Tingkat akurasi yang didapat untuk menentukan akreditasi SD di Kota Malang menggunakan SVM kernel RBF sebesar 93,92% dan untuk SVM kernel *polynomial* adalah sebesar 92,67%.

Penelitian lainnya dilakukan oleh Amalia (2018) dengan membandingkan metode SVM dan *Neural Network* (NN) untuk klasifikasi penyakit ginjal kronis. Dalam Penelitian dia menyebutkan bahwa penyakit ginjal kronis adalah satu satu penyakit yang mematikan, karena penyakit ini bisa menimpulkan penyakit-penyakit mematikan lainnya. Diketahui dari hasil penelitian yaitu dengan metode *Neural Network* diperoleh nilai akurasi 93,36% dan metode *Support Vector Machine* (SVM) diperoleh nilai akurasi 95.16%. Hasil yang diperoleh termasuk dalam jenis klasifikasi sangat baik. Sehingga dapat disimpulkan bahwa SVM dan NN memiliki performa kinerja yang baik untuk pengolahan *dataset* penyakit ginjal kronis. Dan dari hasil penelitian diketahui untuk *dataset* ginjal kronis bahwa metode SVM menghasilkan nilai akurasi yang lebih tinggi dari metode *Neural Network*.

Sartono dan Syafitri (2010) dalam penelitian mengatakan bahwa *Random Forest* adalah solusi pilihan untuk mengatasi kelemahan pada *regression tree* dan *classification tree*. Metode *regression tree* dan *classification tree* adalah metode yang sangat umum digunakan karena memiliki tingkat *error* yang kecil dalam menduga sesuatu dengan membuat pohon tunggal yang dianggap terbaik. Namun kelinieran hubungan antara peubah *predictor* dan peubah respon sering menjadi

kendala. Masalah tersebut biasanya dilakukan dengan tranformai data, tapi tidak semua bisa dilakukan tranformasi. Seletah melalui perkembangan yang pesat, para peneliti membuat pohon gabungan yang dimana data yang akan dianalisis diambil secara acak untuk dibuat beberapa pohon dan kemudian digabungkan menjadi 1 pohon, hal tersbut disebut dengan *Random Forest*. Dengan kata lain, ada dua tahapan utama dalam analisis ini, yaitu *bootstrap* yang tidak lain adalah pengambilan contoh dari data contoh yang dimiliki (*resampling*) dan *aggregating* yaitu menggabungkan banyak nilai dugaan menjadi satu nilai dugaan untuk mendapatkan dugaan terbaik. Berbagai literatur mencatat bahwa metode pohon gabungan mampu bekerja dengan baik dan memberikan dugaan yang lebih tinggi akurasinya dibandingkan pohon tunggal. Menimbang bahwa *software-software open-source* juga banyak yang menyediakan fungsi untuk menghasilkan analisis pohon gabungan ini, maka tidak berlebihan kiranya jika penulis manyebutkan bahwa metode ini layak untuk menjadi alternatif bagi pengguna pohon tunggal.

Penelitian selanjutnya oleh Daqiqil, Astread, dan Mahdiyah (2017) tentang penerapan metode *Random Forest* dan *boosted* C5.0 untuk diagnosa kanker payudara. Penelitian tersebut adalah mencar tingkat akurasi dari metode *Random Forest* dan *Boosted* C5.0 dan kemudian diterapkan untuk deteksi diagnosa penyakit kanker payudara. Penelititan ini menjelaskan bahwa *Boosted* C5.0 adalah pengembangan dari algoritma C4.5 dan ID3 (*Iterative Dichotomiser* 3). Algoritma C5.0 memiliki fitur yang lebih lengkap (*winnowing* dan *boosting*), lebih cepat, lebih efisien, dan menghasilkan *tree* yang lebih sederhana dari C.45. Sedangkan *Random Forest* bekerja dengan cara membangun membangun prediktor dengan sekumpulan *decision tree* yang berkembang secara acak pada subruang data *training*. Hasil yang diberikan oleh *Random Forest* untuk klasifikasi adalah modus dari *decision tree*nya. Hasil akurasi algoritma *Random Forest* menunjukkan performa yang lebih baik dengan akurasi 95.7% dan C50 dengan akurasi 93.7%.

Penelitian lainnya oleh Perdana, Soelaiman, dan Farichah (2017) tentang pengelompokan berkas musik berdasarkan kemiripan karakteristik suara menggunakan metode *Random Forest*, *Multi layer perceptron*, dan J48. Pengelompokan musik berdasarakan karakteristik suara sangat penting bagi

penikmat musik. Banyak penikmat musik yang menikmati musik berdasarakan *genre* favorit masing-masing. Karena itu dibutuhkan metode ekstraksi fitur yang tepat untuk dapat merepresentasikan berkas musik berdasarkan *genre* dengan baik. Studi ini melakukan ekstraksi fitur berkas musik dengan mengekstraksi fitur *spectral centroid, spectral flux, spectral rolloff*, dan *short time energy* pada tiap berkas musik yang diolah dan kemudian dihitung nilai *mean, median, skewness*, dan kurtosisnya, dan selanjutnya dikelompokkan menggunakan metode klasifikasi. Tingkat akurasi dari metode klasifikasi tertinggi adalah metode *Random Forest* dengan akursi 80%, kemudian metode *multi layer perceptron* dengan akurasi 77,62%, dan yang terakhir adalah metode J48 dengan akurasi 62,86%. Dari tingkat akurasi, berarti dapat dikatakan bahwa untuk mengelompokan suara musik berdasarkan *genre* terbaik adalah menggunakan metode *Random Forest*.

BAB III
LANDASAN TEORI

3.1 Pendidikan Tinggi

Pendidikan tinggi merupakan pendidikan jenjang untuk menuntut ilmu setelah pendidikan menengah (SMA/Sederajat). Pendidikan tinggi terdiri dari berbagai jenjang, yaitu jenjang diploma, sarjana, magister, doktor, dan program profesi. Satuan badan yang menjalankan pendidikan tinggi disebut perguruan tinggi dan dikenal dengan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Bentuk perguruan tinggi ada bermacam-macam, seperti universitas, institute, sekolah tinggi, politeknik, spesialis, dan akademi. (Undang-Undang Republik Indonesia Nomor 12 pasal 1, 2012).

Fungsi pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 pasal 4 tahun 2012 adalah sebagai berikut ;

1. Mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa.
2. Mengembangkan sivitas akademika yang inovatif, responsif, kreatif, terampil, berdaya saing, dan kooperatif melalui pelaksanaan tridharma, dan
3. Mengembangkan ilmu pengetahuan dan Teknologi dengan memperhatikan dan menerapkan nilai humaniora.

Adapun tujuan pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 pasal 5 tahun 2012 adalah ;

1. Berkembangnya potensi Mahasiswa agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa dan berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, terampil, kompeten, dan berbudaya untuk kepentingan bangsa.
2. Dihasilkannya lulusan yang menguasai cabang Ilmu Pengetahuan dan/atau Teknologi untuk memenuhi kepentingan nasional dan peningkatan daya saing bangsa.

3. Dihasilkannya Ilmu Pengetahuan dan Teknologi melalui Penelitian yang memperhatikan dan menerapkan nilai Humaniora agar bermanfaat bagi kemajuan bangsa, serta kemajuan peradaban dan kesejahteraan umat manusia.
4. Terwujudnya Pengabdian kepada Masyarakat berbasis penalaran dan karya Penelitian yang bermanfaat dalam memajukan kesejahteraan umum dan mencerdaskan kehidupan bangsa.

3.2 *Data Mining*

Data mining adalah proses mengolah atau merangkum data yang berjumlah besar melalui proses analisis agar bisa mengambil kesimpulan data yang berharga. Selain itu, bisa diartikan dengan gabungan antara metode statistik dan *artificial intelligence*/kecerdasan buatan yang terus berkembang (Gorunescu, 2011).

3.2.1 Model *Supervised Learning*

Model ini digunakan untuk memprediksi hasil masa depan berdasarkan data historis untuk dipelajari menggunakan metode tertentu agar bisa memprediksi dengan akurat. Contoh penerapan metode *Supervised Learning* adalah untuk memprediksi kemungkinan terjadinya bahaya yang akan terjadi dengan melihat beberapa faktor sesuai dengan data historis yang telah dipelajari, seperti bencana gempa bumi, banjir, dan lain-lain (Jain, 2015).

Supervised learning adalah sebuah pendekatan dengan cara melatih data yang sudah ada dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang sudah ada. Metode analisis untuk *supervised learning* adalah *Decision tree, Random Forest, Nearest - Neighbor Classifier, Naive Bayes Classifier, Artificial Neural Network, Support Vector Machine, Fuzzy K-Nearest Neighbor* (Chandra, 2017).

3.2.2 Model *Unsupervised Learning*

Unsupervised learning tidak memiliki data latih, sehingga dari data yang ada, kita mengelompokan data tersebut menjadi 2 bagian atau 3 bagian dan seterusnya. Contoh metode ini adalah seseorang belum pernah membeli buku sama sekali, namun dalam suatu hari, orang tersebut membeli banyak buku dan ingin membaginya kedalam beberapa kategori agar nantinya mudah dicari namun belum

diketahui kategori dari buku tersebut. Maka pembagian buku dengan cara mengidentifikasi buku mana yang mirip berdasarkan isinya. Metode analisis *unsupervised learning* adalah *K-Means, Hierarchical Clustering, DBSCAN, Fuzzy C-Means, Self-Organizing Map* (Chandra, 2017).

3.3 Klasifikasi

Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik. Klasifikasi merupakan suatu pekerjaan yang melakukan penilaian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012). Tahapan klasifikasi ada beberapa, yaitu ;

3.3.1 *Data Collection*

Data collection adalah proses mengumpulkan dan memastikan informasi pada *variable of interest* (subjek yang akan dilakukan uji coba), dengan cara yang sistematis yang memungkinkan seseorang dapat menjawab pertanyaan dari permasalahan yang ada dan mengevaluasi hasil. Komponen pengumpulan data dari penelitian ini bersifat umum, bisa dilakukan untuk semua bidang studi termasuk ilmu fisik dan sosial, humaniora, bisnis, dan lainnya (Redaksi, 2014).

3.3.2 *Data Cleaning*

Data yang baik dan berkualitas adalah kunci dasar untuk menghasilkan data yang berkualitas, dengan cara menghapus data *error,* mengahapus data *incomplete* atau data yang nilai atributnya hilang atau datanya kosong, dan menyamakan satuan atau nilai dari data (Subhan, 2017). *Data cleaning* adalah proses analisa kualitas dari suatu data dengan cara mengubah, mengoreksi, atau menghapus data-data yang salah, tidak lengkap, tidak akurat, atau memiliki format yang salah dalam basis data guna menghasilkan data berkualitas tinggi (Tawakal, 2015). Contoh ;

Tabel 3.1 *Data Sebelum Cleaning*

No	Jenis Klmn	Prodi	Fakultas	Tempat Lahir	Kabupaten	Propinsi
1	P	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	WERU	KAB.INDRAMAYU	JAWA BARAT
2	L	Ilmu Komunikasi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Bumi Agung	KAB.SLEMAN	D.I. YOGYAKARTA
3	L	Manajemen	EKONOMI	Tuban	KODYA YOGYAKARTA	D.I. YOGYAKARTA
4			TEKNIK SIPIL DAN PERENCANAAN	Lamongan		
5				Semarang		
6	L		TEKNIK SIPIL DAN PERENCANAAN	BLORA		
7	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Jakarta	KODYA JAKARTA UTARA	DKI JAKARTA
8	L	Teknik Sipil	TEKNIK SIPIL DAN PERENCANAAN	Bantul	KAB.BANTUL	D.I. YOGYAKARTA
9	L	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	Klaten	KAB.KLATEN	JAWA TENGAH
10	L	Teknik Industri	TEKNOLOGI INDUSTRI	Kediri	KAB.KEDIRI	JAWA TIMUR
11	L	Informatika	TEKNOLOGI INDUSTRI	Tuban	KAB.TUBAN	JAWA TIMUR
12	L	Teknik Kimia	TEKNOLOGI INDUSTRI	Kendal	KAB.KENDAL	JAWA TENGAH
13	P	Hukum	HUKUM	Madiun		
14	P	Teknik Kimia	TEKNOLOGI INDUSTRI	Klangenan Cirebon	KAB.CIREBON	JAWA BARAT
15	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Serang	X kab.pandegelang	JAWA BARAT
16	L	Pendidikan Agama Islam	ILMU AGAMA ISLAM	Sleman	KAB.SLEMAN	D.I. YOGYAKARTA
17	L	Manajemen	EKONOMI	Serang		
18	L	Manajemen	EKONOMI	Yogyakarta	KAB.KENDAL	JAWA TENGAH
19	L	Hukum	HUKUM	Sanggau	KAB.SANGGAU	KALIMANTAN BARAT
20	L	Hukum	HUKUM	Samarinda	KODYA YOGYAKARTA	D.I. YOGYAKARTA

Tabel 3.2 *Data Setelah Cleaning*

No	Jenis Klmn	Prodi	Fakultas	Tempat Lahir	Kabupaten	Propinsi
1	P	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	WERU	KAB.INDRAMAYU	JAWA BARAT
2	L	Ilmu Komunikasi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Bumi Agung	KAB.SLEMAN	D.I. YOGYAKARTA
3	L	Manajemen	EKONOMI	Tuban	KODYA YOGYAKARTA	D.I. YOGYAKARTA
7	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Jakarta	KODYA JAKARTA UTARA	DKI JAKARTA
8	L	Teknik Sipil	TEKNIK SIPIL DAN PERENCANAAN	Bantul	KAB.BANTUL	D.I. YOGYAKARTA
9	L	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	Klaten	KAB.KLATEN	JAWA TENGAH
10	L	Teknik Industri	TEKNOLOGI INDUSTRI	Kediri	KAB.KEDIRI	JAWA TIMUR
11	L	Informatika	TEKNOLOGI INDUSTRI	Tuban	KAB.TUBAN	JAWA TIMUR
12	L	Teknik Kimia	TEKNOLOGI INDUSTRI	Kendal	KAB.KENDAL	JAWA TENGAH
14	P	Teknik Kimia	TEKNOLOGI INDUSTRI	Klangenan Cirebon	KAB.CIREBON	JAWA BARAT
15	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Serang	X kab.pandegelang	JAWA BARAT
16	L	Pendidikan Agama Islam	ILMU AGAMA ISLAM	Sleman	KAB.SLEMAN	D.I. YOGYAKARTA
18	L	Manajemen	EKONOMI	Yogyakarta	KAB.KENDAL	JAWA TENGAH
19	L	Hukum	HUKUM	Sanggau	KAB.SANGGAU	KALIMANTAN BARAT
20	L	Hukum	HUKUM	Samarinda	KODYA YOGYAKARTA	D.I. YOGYAKARTA

3.3.3 Data *Reduction*

Pada tahap data *reduction* adalah tahap dimana data yang telah terkumpul, kemudian di bersihkan, proses selanjutnya adalah memilih variabel atau atribut yang akan digunakan dalam penelitian. Tahap ini dilakukan untuk mengurangi atribut yang tidak digunakan akan tetapi tetap bersifat informatif (Subhan, 2017).

Tabel 3.3 *Data Sebelum Dan Sesudah Reduce Variabel*

Variabel Asli		Variabel Setelah Reduce
No Mhs		Jenis Klmn
Nama Mhs		Prodi
Jenis Klmn		Propinsi
Prodi	→	Pekerjaan Ayah
Propinsi		Pekerjaan Ibu
Pekerjaan Ayah		Pendidikan Ayah
Pekerjaan Ibu		Pendidikan Ibu
Pendidikan Ayah		
Pendidikan Ibu		

3.3.4 Data *Training* dan Data *Testing*

Data *training* digunakan oleh algoritma untuk membentuk sebuah model klasifikasi. Model ini merupakan representasi pengetahuan yang akan digunakan untuk mengukur sejauh mana klasifikasi berhasil melakukan prediski dengan benar. Karena itu, data yang ada yang ada pada data *testing* seharusnya tidak boleh ada pada data *training* sehingga dapat diketahui apakah model klasifikasi dapat melakukan klasifikasi dengan baik. Proporsi antara data *training* dan data *testing* tidak mengikat tetapi agar variasi dalam model tidak terlalu besar maka disarankan data *training* lebih besar dibandingkan data *testing.* Biasanya 3/4 dari total data dijadikan data *training* sedangkan sisanya dijadikan data *testing.* Selain itu, ada pula penelitian yang menghasilkan keakuratan model klasifikasi optimum dengan proporsi 75 : 25 untuk data *training* dan data *testing* (Rachman dan Purnami, 2012).

3.4 *Balancing Data*

Balancing data adalah merubah data yang tidak seimbang (i*mbalance* data) menjadi data yang seimbang (*balance*). *Imbalance* data adalah kondisi ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*majority class*) sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (*minority class*) (Sastrawan, Baizal, dan Bijaksana, 2010). Metode ini secara luas dikenal sebagai 'Metode Sampling'. Umumnya, metode ini bertujuan untuk memodifikasi data yang tidak seimbang ke dalam distribusi yang seimbang

menggunakan beberapa mekanisme. Modifikasi terjadi dengan mengubah ukuran kumpulan data asli dan menyediakan proporsi keseimbangan yang sama.

Menurut *Analytics Vidhya Content Team* (2016), menjelaskan bahwa *balancing* data ada beberapa metode, yaitu diantaranya adalah *under sampling*. Metode *under sampling* ini bekerja dengan kelas mayoritas (*majority class*), dengan mengurangi jumlah observasi dari kelas mayoritas untuk membuat kumpulan data dengan jumlah yang seimbang. Metode ini paling baik digunakan ketika kumpulan data sangat besar dan mengurangi jumlah sampel pelatihan yang membantu meningkatkan waktu operasi dan masalah penyimpanan. Metode *under sampling* secara acak memilih observasi dari kelas mayoritas yang dieliminasi sampai set data menjadi seimbang antar masing-masing kelas.

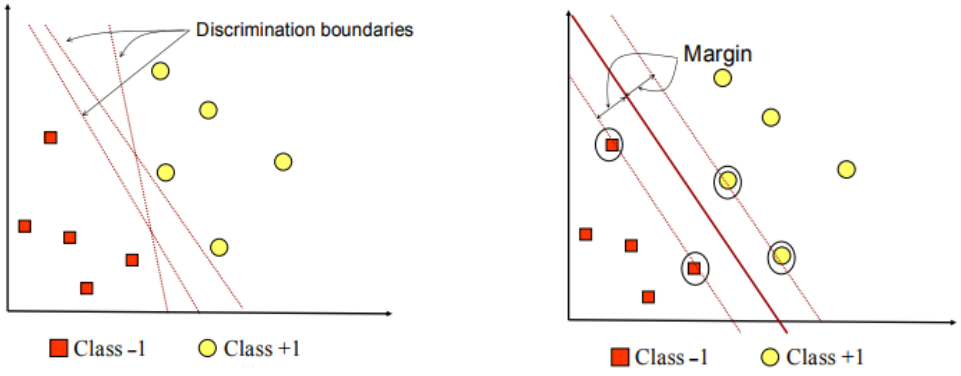
3.5 *Support Vector Machine* (SVM)

Vapnik memperkenalkan SVM untuk pertama kali pada tahun 1992 sebagai rangkaian konsep unggulan pada bidang *pattern recognition*. Usia SVM sebagai salah satu metode *pattern recognition* masih terbilang relatif muda. Dewasa ini SVM meruapakn salah satu metode yang berkembang pesat. SVM merupan salah satu metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperplane* terbaik untuk memisahkan dua kelas data. *SVM* bekerja dengan memaksimalkan *margin* yang merupakan jarak pemisah antara kedua kelas data tersebut (Pratiwi 2017). Pada dasarnya SVM memiliki prinsip linear, akan tetapi kini SVM telah berkembang sehingga dapat bekerja pada masalah *non*-linear. Cara kerja SVM pada masalah *non*-linear adalah dengan memasukkan konsep kernel pada ruang berdimensi tinggi. Pada ruang yang berdimensi ini, nantinya akan dicari pemisah atau yang sering disebut *hyperplane*.

Hyperplane dapat memaksimalkan jarak atau *margin* antara kelas data. *Hyperplane* terbaik antara kedua kelas dapat ditemukan dengan mengukur *margin* dan kemudian mencari titik maksimalnya. Usaha dalam mencari *hyperplane* yang terbaik sebagai pemisah kelas-kelas adalah inti dari proses pada metode SVM (Assaffat, 2015).

3.5.1 Linear *Separable* Data

Metode SVM dengan *hyperplane* yang berbentuk garis lurus disebut dengan *linear saparable*. **Gambar 3.1** merupakan ilustrasi dari *hyperplane linear separable* data ;



a. b.

Gambar 3.1 *Garis Linear Pemisah Dua Kelas (Sumber : Nugroho, 2003)*

Dapat dilihat ilustrasi pada **Gambar 3.1** adalah beberapa *pattern* yang merupakan anggota dari dua buah kelas yaitu kelas +1 dan kelas –1. Simbol untuk *pattern* pada kelas -1 adalah kotak yang berwarna merah, sedangkan simbol untuk *pattern* pada kelas +1 adalah lingkaran dengan warna kuning. Dalam SVM yang telah disebutkan diatas menemukan garis (*hyperplane*) yang dapat memisahkan antara kedua kelompok tersebut. Berbagai macam garis pemisah *(discrimination boundaries) alternative* yang ditunjukkan pada **gambar 3.1** bagian a. Dalam menemukan *hyperplane* yaitu dengan cara mengukur *Margin hyperplane* tersebut dan kemudian mencari titik maksimalnya. Jarak antara *hyperplane* dengan *pattern* pada masing-masing kelas biasa disebut dengan *margin*. Untuk *pattern* paling dekat disebut dengan *support vector*. Pada **gambar 3.1** bagaian b garis yang berada di tengah menunjukkan *hyperplane* yang terbaik, karena terletak tepat pada tengah-tengah antar kelas, sedangkan *support vector* adalah titik merah dan kuning yang berada dalam lingkaran hitam. Usaha dalam mencari lokasi *hyperplane* ini merupakan proses inti dari SVM.

Pada penelitian ini data yang tersedia dapat dinotasikan sebagai $\mathcal{x} \in R^d$, sedangkan label untuk masing-masing kelas dinotasikan $y_i \in \{-1, +1\}$ untuk i = 1,2,3, … n.

Pada kedua kelas dapat diasumsikan terpisah secara sempurna oleh *hyperplane* berdimensi *d*, yang didefinisikan sebagai berikut:

$$\vec{w} . \vec{x}+b=0 \quad \ldots(3.1)$$

Untuk *pattern* x_i yang termasuk dalam kelas -1 dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan:

$$\vec{w} . \vec{x}+b \leq -1 \quad \ldots (3.2)$$

Sedangkan untuk *pattern* x_i yang *term*asuk kelas +1 dapat dirumuskan sebagai berikut:

$$\vec{w} . \vec{x}+b \geq +1 \quad \ldots (3.3)$$

dimana

R^d = ruang vektor

d = dimensi

n = banyak data

$\vec{w}$ = vektor bobot

$\vec{x}$ = vektor data (*input*)

b = bias

Margin terbesar dapat diperoleh dengan cara memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Hal ini dapat dirumuskan sebagai masalah *Quadratic Programming* (QP), yaitu mencari titik minimal persamaan 3.4, dengan memperlihatkan *constraint* persamaan 3.5. (Pratiwi, 2017)

$$\min_w = \tau(w)=\frac{1}{2}\|w\|^2 \quad \ldots (3.4)$$

$$y_i(x_i . w+b) - 1 \geq 0 \quad \ldots (3.5)$$

dimana $\|w\|$ adalah vektor normal.

Salah satu teknis komputasi yaitu *lagrange multiplier* yang dapat memecahkan masalah ini dapat dinyatakan pada persamaan 3.6.

$$L(w,b,\alpha)=\frac{1}{2\|w\|^2}2-\sum \alpha_i(y_i((\vec{x}.\vec{w}+b)-1)) \; ; \; i = 1,2, \ldots ,n \quad \ldots (3.6)$$

Dimana α merupakan *lagrange multiplier*, yang bernilai $\alpha_i \geq 0$. Nilai optimal pada persamaan 3.6 dapat dihitung dengan meminimalkan nilai L terhadap w dan b, dan memaksimalkan nilai L terhadap α_i, dengan memperhatikan sifat bahwa pada titik optimal *gradient* $L = 0$. Pada persamaan 3.6 dapat dimodifikasi sebagai maksimalisasi *problem* yang hanya mengandung α_i, seperti yang terlihat pada persamaan 3.7 dan persamaan 3.8 dibawah ini.

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i . \vec{x}_i \quad \text{... (3.7)}$$

$$\text{dimana } \alpha_i \geq 0 \ (i=1, 2, \ldots, l) \ \sum_{i=1}^{l} \alpha_i y_i = 0 \quad \text{... (3.8)}$$

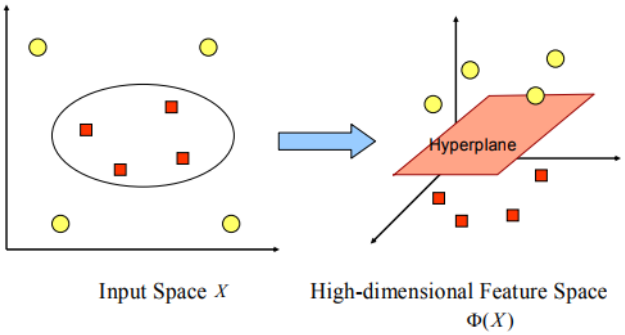
Dengan demikian, maka akan diperoleh α_i yang kebanyakan bernilai positif yang disebut sebagai *support vector* dan juga memperoleh persamaan 3.9 dan persamaan 3.10 sebagai solusi pemisah. (Pratiwi, 2017)

$$w = \sum \alpha_i y_i x_i \quad \text{... (3.9)}$$

$$b = y_k - w^T x_k \quad \text{... (3.10)}$$

3.5.2 *Non*-Linear *Separable* Data

Dalam dunia nyata *(real world problem)* pada umumnya masalah data yang diperoleh jarang yang bersifat *linear*, banyak yang bersifat *non* linear. Pada SVM terdapat sebuah fungsi kernel, yaitu fungsi yang digunakan untuk menyelesaikan *problem non linear*. Kernel berfungsi memungkinkan untuk mengimplementasikan suatu model pada ruang dimensi lebih tinggi (ruang fitur).



Gambar 3.2 *Hyperplane* (Sumber : Nugroho, 2003)

a) Kernel *Radian Basis Function* (RBF)

$$K(\vec{X}_i, \vec{X}_j) = exp\left(\frac{\left\|\vec{X}_i - \vec{X}_j\right\|^2}{2\sigma^2}\right) \quad \text{... (3.11)}$$

b) Kernel Sigmoid

$$K(\vec{X}_i, \vec{X}_j) = \tan(\sigma\, x_i^t\, x_j) \quad \text{...(3.12)}$$

3.6 *Random Forest*

Random Forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan *Random Forest* antara lain dapat menghasilkan *error* yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data *training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing* data (Breiman,2001).

Secara sederhana, algoritma pembentukan RF dapat disebutkan sebagai berikut. Andaikan gugus data *training* yang kita miliki berukuran *n* dan terdiri atas *d* peubah penjelas (*predictor*). Tahapan penyusunan dan pendugaan menggunakan RF adalah:

a. (tahapan *bootstrap*) tarik contoh acak dengan permulihan berukuran n dari gugus data *training*.
b. (tahapan *random sub-setting*) susun pohon berdasarkan data tersebut, namun pada setiap proses pemisahan pilih secara acak jumlah *variable* prediktor (*m) < d* peubah penjelas, dan lakukan pemisahan terbaik.
c. Ulangi langkah a-b sebanyak *k* kali sehingga diperoleh *k* buah pohon acak
d. Lakukan pendugaan gabungan berdasarkan *k* buah pohon tersebut (misal menggunakan *majority vote* untuk kasus klasifikasi, atau rata-rata untuk kasus regresi)

Perhatikan bahwa pada setiap kali pembentukan pohon, kandidat peubah penjelas yang digunakan untuk melakukan pemisahan bukanlah seluruh peubah yang terlibat namun hanya sebagian saja hasil pemilihan secara acak. Bisa dibayangkan bahwa proses ini menghasilkan kumpulan pohon tunggal dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah kumpulan pohon tunggal memiliki korelasi yang kecil antar pohonnya. Korelasi kecil ini

mengakibatkan ragam dugaan hasil RF menjadi kecil (Hastie *et al*, 2008) dan lebih kecil dibandingkan ragam dugaan hasil bagging (Zhu, 2008).

Jika melihat secara seksama algoritma pembentukan RF, salah satu yang bisa diubah adalah nilai *m*, yaitu banyaknya peubah penjelas yang digunakan sebagai kandidat pemisah dalam pembentukan pohon. Nilai *m* yang semakin besar akan menyebabkan korelasi (ρ) semakin besar. Contoh ekstrim adalah jika kita gunakan $m = d$ yang menyebabkan setiap kali pengulangan akan menghasilkan pohon yang sama sehingga nilai korelasi akan menjadi maksimum yaitu sebesar 1. Namun, jika nilai *m* kita buat sekecil mungkin yaitu hanya 1 peubah penjelas saja yang dijadikan kandidat pemisah, maka pohon yang diperoleh akan menjadi pohon dengan akurasi yang sangat rendah. Dengan demikian jelas bahwa pemilihan *m* memegang peranan dalam menentukan kebaikan RF yang dihasilkan.

Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain*. Untuk menghitung nilai *entropy* digunakan rumus seperti pada persamaan 3.12, sedangkan nilai *information gain* menggunakan persamaan 3.13 (Nugroho, 2017).

$$Entropy(Y) = -\sum p(c/Y) log^2\ p(c|Y) \quad \ldots(3.13)$$

Dimana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

Information gain (Y,a)

$$= Entropy(Y) - \sum v \in Values(a) \frac{/Yv/}{/Ya/} Entropy\ (Yv) \quad \ldots(3.14)$$

Dimana *Values(a)* merupakan semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, dan Ya adalah semua nilai yang sesuai dengan a.

3.7 *Confusionan Matrix*

Berikut ini merupakan hasil dari *confusionan matrix* (Sasongko, 2016)

Tabel 3.4 *Tabel Confusionan Matrix*

Prediksi	Aktual	
	True	*False*
True	TP	FN
False	FP	TN

Keterangan:

TP = Jumlah prediksi yang tepat bersifat positif *(True Positive).*

TN = jumlah prediksi yang tepat bersifat negatif *(True Negative).*

FP = jumlah prediksi yang salah bersifat positif *(False Positive).*

FN = jumlah prediksi yang salah bersifat negatif *(False Negative).*

Accuracy merupakan proporsi jumlah prediksi benar. Rumus akurasi adalah:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad \ldots (3.15)$$

BAB IV
METODE PENELITIAN

4.1 Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder yang diambil dari *database* Badan Sistem Informasi (BSI) Universitas Islam Indonesia. Data tersebut berupa data seluruh alumni Universitas Islam Indonesia jenjang sarjana dari tahun kelulusan 2000-2017 dengan berbagai *background* yang berbeda.

4.2 Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini adalah sebanyak 10 variabel. Berikut ini adalah tabel berisi variabel penelitian dan penjelasan definisi operasional masing-masing variabelnya ;

Tabel 4.1 *Variabel Penelitian dan Penjelasan Operasional*

Variabel	Definisi Penjelasan Operasional
Jenis kelamin	Jenis kelamin dari alumni, yaitu : laki-laki dan perempuan.
Prodi	Program studi alumni semasa menempuh jenjang kuliah. yaitu ; Hukum, Psikologi, Manajemen, Farmasi, Akuntansi, Teknik Sipil, Teknik Industri, Kedokteran, Arsitektur, Informatika, Teknik Kimia, Teknik Lingkungan, Ekonomi, Statistika, Kimia, Pendidikan Agama Islam, Teknik Elektro, Ekonomi Islam, Akhwal Al-Syakhshiyah, Ilmu Komunikasi, dan Teknik Mesin.
Provinsi	Provinsi asal alumni, yaitu ; Berisi 33 provinsi di Indonesia
Pekerjaan ayah	Pekerjaan ayah dari alumni, yaitu ; PNS, Wiraswasta, Peg. Swasta, Pensiun, Peteani, dan Lain-lain.
Pekerjaan ibu	Pekerjaan ibu dari alumni, yaitu ;

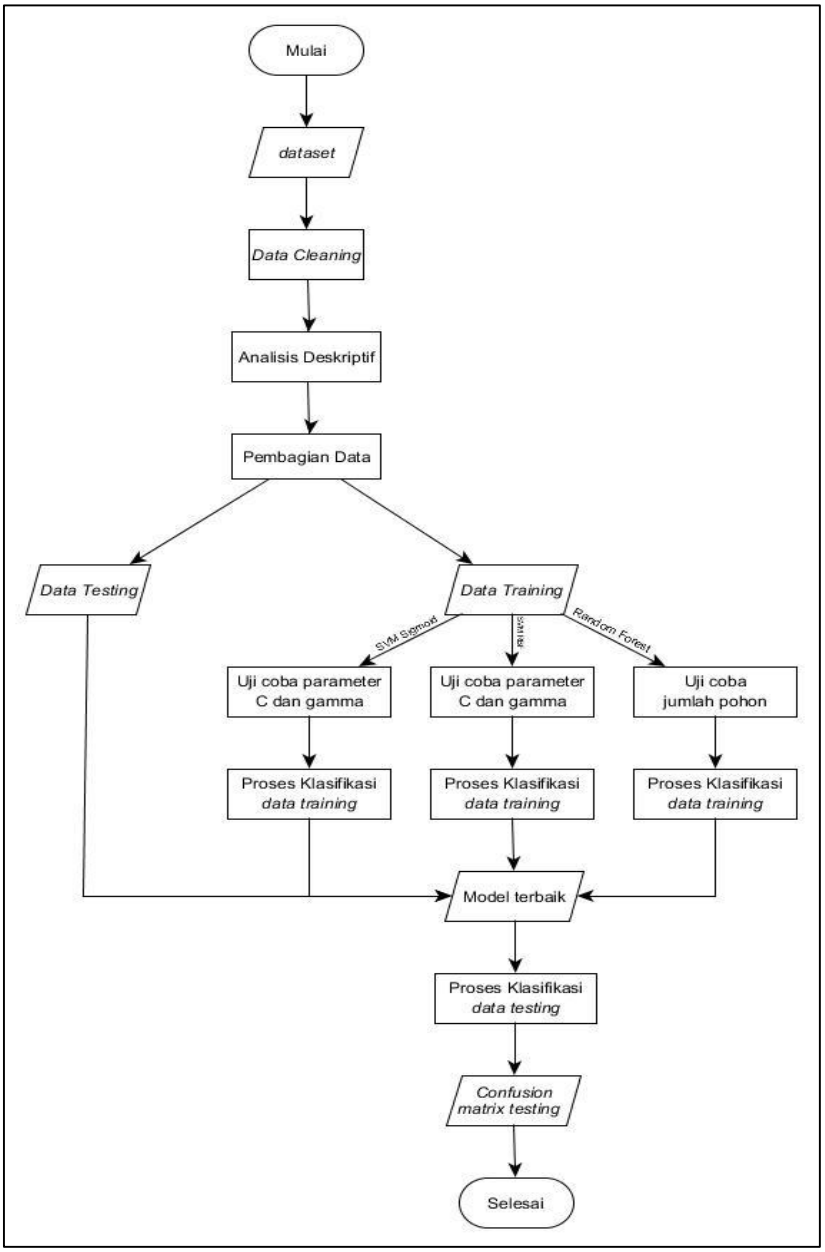
	Ibu Rumah Tangga (IRT), PNS, Wiraswasta, Peg. Swasta, Petani, dan Lain-lain.
Pendidikan ayah	Pendidikan terakhir yang ditempuh ayah dari alumni, yaitu; Tidak tamat SD, tamat SD, tamat SMP, SMA/Sederajat, Diploma, Sarjana muda, Sarjana, Pasca Sarjana, dan Doktor.
Pendidikan ibu	Pendidikan terakhir yang ditempuh ibu dari alumni, yaitu; Tidak tamat SD, tamat SD, tamat SMP, SMA/Sederajat, Diploma, Sarjana muda, Sarjana, Pasca Sarjana, dan Doktor.
Jurusan SMA	Jurusan alumni saat masih di bangku SMA/Sederajat, yaitu; IPA, IPS, dan Lain-lain.
IPK	Nilai akhir keseluruhan semasa kuliah, yaitu : 2,00 – 4,00
Keterangan	Berupa keputusan yang menjelaskan bahwa alumni berhasil lulus tepat waktu atau tidak, ditentukan dari lamanya studi alumni (tgl. registrasi sampai tgl. yudisium akhir), yaitu : Tepat waktu → studi ≤ 4 tahun, dan Tidak Tepat waktu → studi >4 tahun.

4.3 Metode Analisis Data

Penelitian ini menggunakan analisis deskriptif untuk mengetahui gambaran umum data dan perbandingan hasil analisis klasifikasi *Support Vector Machine* (SVM) dan *Random Forest*. Untuk metode SVM, peneliti menggunakan kernel RBF dan Sigmoid dengan parameter C =0.1, 1, 5, 10, 50 dan *gamma* = 1, 2, 3, 4. Untuk metode *Random Forest*, peneliti menggunakan pohon sebanyak 25,50,100,500, dan 1000.

4.4 Tahapan Analisis Data

Tahapan yang dilakukan dalam penelitian ini dapat digambarkan dengan gambar berikut:



Gambar 4.1 *Diagram Alur Penelitian*

1. Tahapan pertama adalah pengumpulan data, pengumpulan data ini kemudian terbentuk *dataset* yang akan digunakan dalam penelitian ini.
2. Data *cleaning* merupakan tahapan berikutnya setelah *dataset* terkumpul, data *cleaning* ini dilakukan untuk menghilangkan data *mising* dan variabel yang tidak digunakan.
3. Analisis deskriptif ini dilakukan pada masing-masing variabel dengan menggunakan tabel atau grafik.

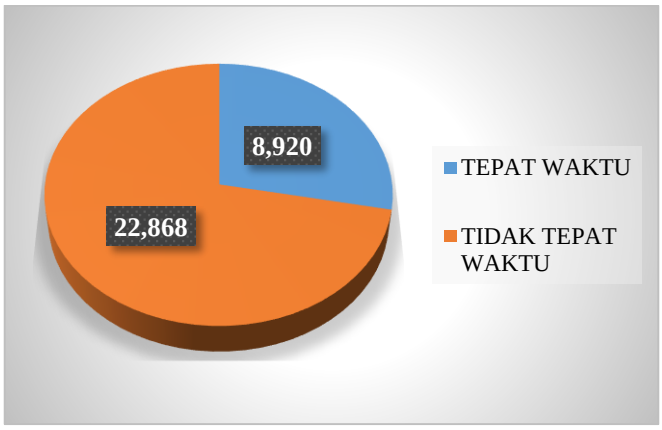
4. Pada klasifikasi jenis data pada umumnya dibagi menjadi dua yaitu data *training* dan data *testing*.
5. Data *training* yaitu data latih yang digunakan, data *training* pada penelitian ini yang digunakan adalah 75% dari total data
6. Data *testing* yaitu data yng digunakan untuk menguji tingkat akurasi metode klasifikasi. Data *testing* yang digunakan dalam penelitian adalah 25%.
7. Data *training* digunakan untuk melakukan analisis klasifikasi SVM untuk kernel RBF maupun sigmoid. Pada SVM kernel RBF dan sigmoid perlu adanya uji coba/inisiasi parameter C dan *gamma* yang ditentukan oleh peneliti sendiri. Setelah nilai paramater telah ditentukan, selanjutnya adalah melakukan optimasi untuk nilai parameter dengan proses *tuning* untuk menentukan nilai parameter optimum. Langkah selanjutnya pada proses SVM kernel RBF dan sigmoid adalah melakukan klasifikasi SVM terhadap data *training* menggunakan nilai parameter C dan *gamma* yang optimum agar menghasilkan model yang analisis terbaik.
8. Kalisifikasi *Random Forest* juga menggunakan data *training* dan menentukan banyaknya pohon yang terbentuk (k atau *ntree*) dan menentukan banyaknya jumlah *variable* prediktor (m) yang besarnya ditentukan oleh peneliti berdasarkan penelitian sebelumnya. Selanjutnya adalah melakukan optimasi untuk nilai parameter dengan proses *tuning* untuk menentukan nilai parameter optimum. Langkah selanjutnya pada proses *Random Forest* adalah melakukan klasifikasi terhadap data *training* menggunakan nilai k dan m yang optimum agar menghasilkan model yang analisis terbaik.
9. Tahapan selanjutnya setelah diperoleh model terbaik untuk klasifikasi SVM dan *Random Forest* pada data *training*, maka akan dilakukan proses klasifikasi dengan data *testing* untuk menguji tingkat akurasi dari masing-masing metode.
10. Selanjutnya setelah proses untuk data *testing* dilakukan, maka akan menghasilkan tabel *confusionan matrix* untuk data *testing*, yang mana dengan itu bisa terlihat tingkat akurasi dari masing-masing metode klasifikasi.

BAB V

ANALISIS DAN PEMBAHASAN

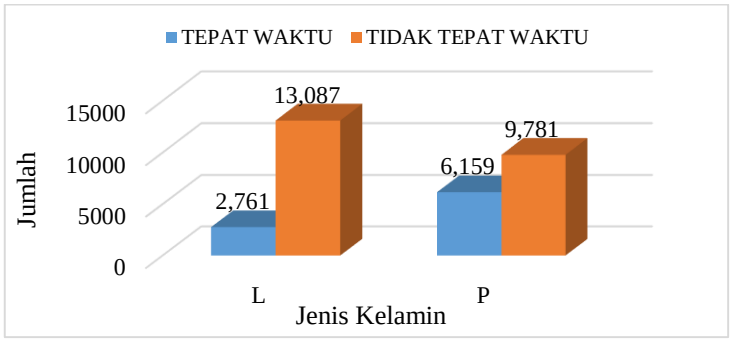
5.1 Analisis Deskriptif

Analisis deskriptif pada penelitian ini digunakan untuk mengetahui gambaran secara umum mengenai data alumni Universitas Islam Indonesia dari tahun kelulusan 2000-2017 berdasarkan ketepatan alumni lulus. Ketepatan lama studi alumni dibagi menjadi 2, yaitu tepat waktu dan tidak tepat waktu. Alumni dikatakan tepat waktu jika lulus dari universitas dengan menempuh studi maksimal 4 tahun lamanya, sedangkan dikatakan tidak tepat waktu jika alumni menyelesaikan masa studinya lebih dari 4 tahun. Penentuan lama studi dilihat dari tanggal alumni tersebut terdaftar jadi mahasiswa Universitas Islam Indonesia sampai tanggal alumni yudisium akhir, perhatikan gambar berikut ini :



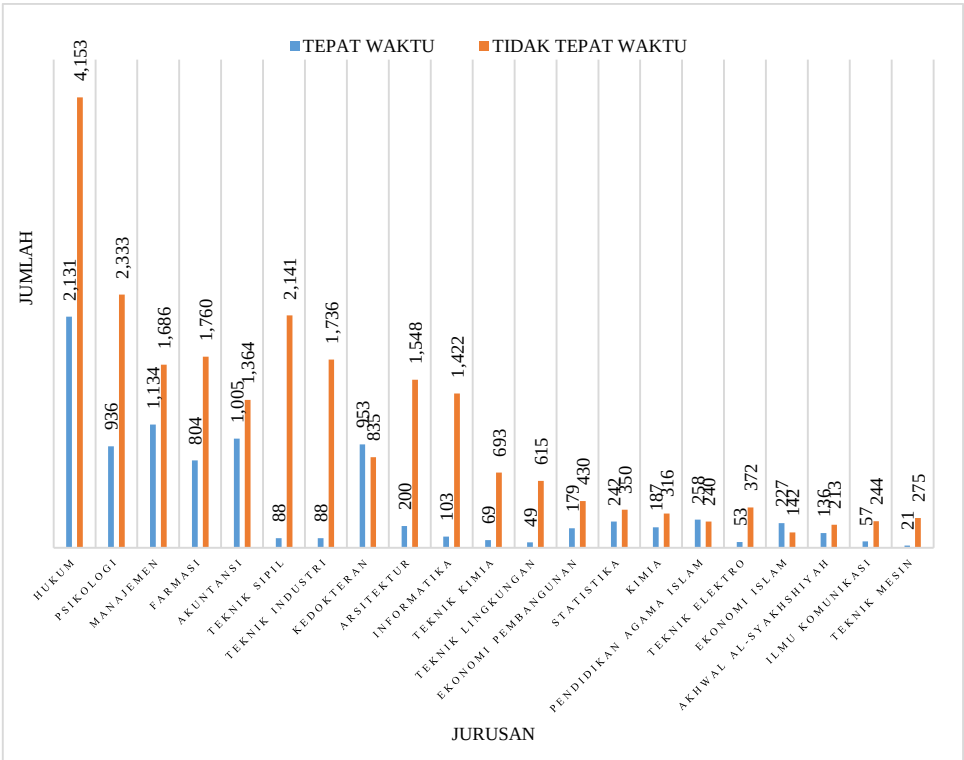
Gambar 5.1 *Jumlah Ketepatan Alumni Lulus*

Dapat dilihat dari gambar di atas bahwa jumlah jumlah alumni yang menempuh jenjang pendidikan tepat waktu hanya sebanyak 8.920 dan alumni yang tidak tepat waktu sangat banyak, yaitu mencapai 22.868. Persentase alumni yang tepat waktu adalah sebesar 28%, sedangkan persentase alumni yang tidak tepat waktu adalah sebesar 72%. Perbedaan jumlah alumni yang tepat waktu dan tidak tepat waktu cukup tinggi, yaitu sebanyak 13.948 orang. Untuk mengetahui berapa banyak alumni lulus berdasarkan jenis kelamin, perhatikan gambar di bawah ini ;



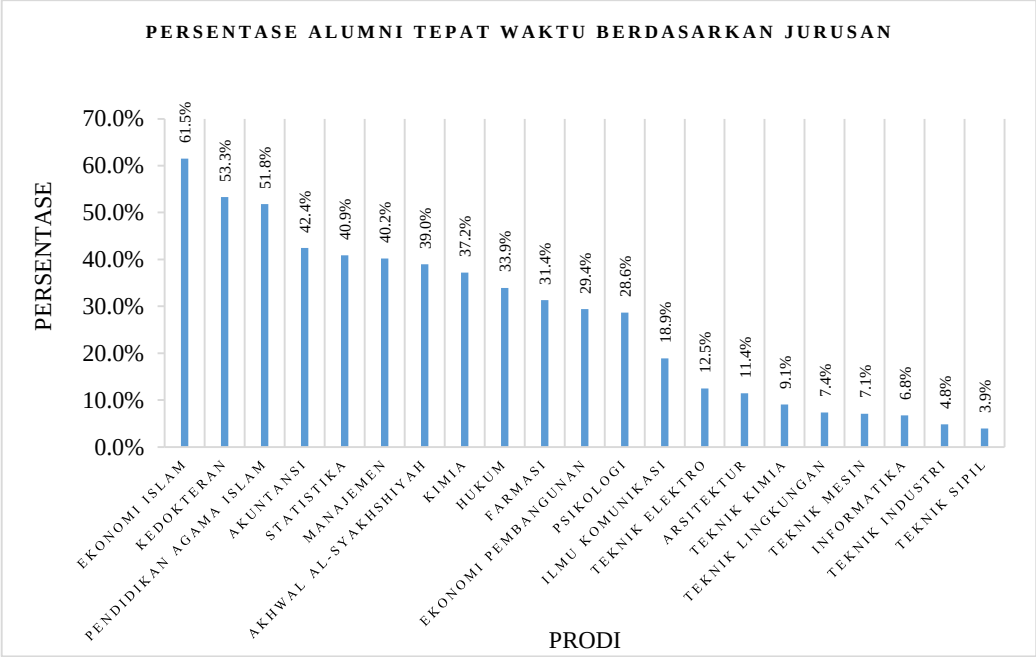
Gambar 5.2 *Ketepatan Lama Studi Alumni Berdasarkan Jenis Kelamin*

Jumlah alumni Universitas Islam Indonesia berdasarkan jenis kelamin seperti pada gambar di atas menunjukan jumlah perempuan sebanyak 15.940 dan alumni laki-laki sebanyak 15.848. Selisih antara jenis kelamin perempuan dan laki-laki tidak terlalu jauh berbeda, yaitu selisihnya sebesar 92 orang. Banyaknya alumni perempuan yang tepat waktu adalah 6.159 dan alumni perempuan yang tidak tepat waktu adalah 9.781, sedangkan alumni laki laki untuk yang tepat waktu adalah 2.761 dan yang tidak tepat waktu adalah 13.087. Berikut ini adalah jumlah alumni berdasarkan dari jurusan yang diambil ;



Gambar 5.3 *Jumlah Alumni Berdasarkan Jurusan*

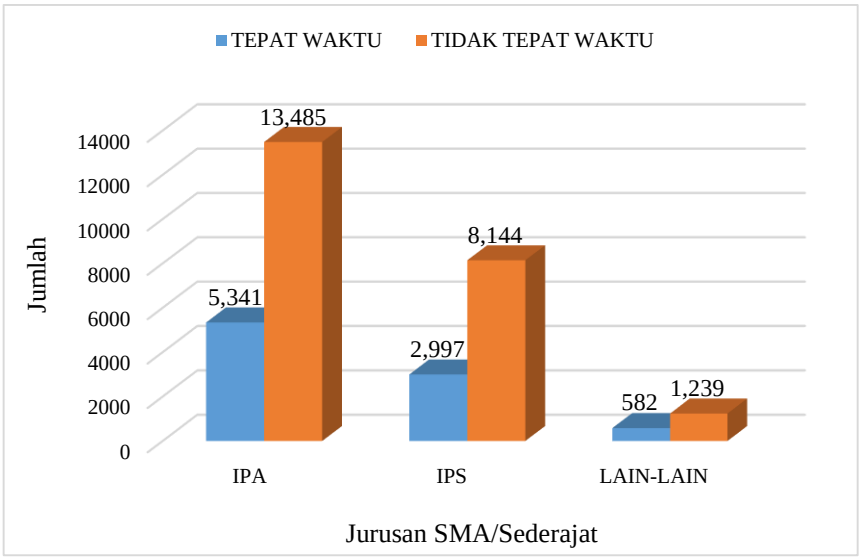
Gambar 5.3 di atas menunjukan bahwa jurusan dengan jumlah alumni terbanyak adalah jurusan hukum dengan 6.284 alumni dengan banyaknya alumni yang tepat waktu adalah 2.131 dan yang tidak tepat waktu adalah 4.153. Kemudian diikuti oleh jurusan psikologi dengan jumlah 3.269 alumni dengan banyaknya alumni yang tepat waktu adalah 936 dan yang tidak tepat waktu adalah 2.333. Sedangkan jurusan yang paling sedikit alumninya adalah jurusan teknik mesin dengan jumlah hanya 296 alumni dengan banyaknya alumni yang tepat waktu adalah 21 dan yang tidak tepat waktu adalah 275, lalu jurusan ilmu komunikasi dengan jumlah 301 alumni dengan banyaknya alumni yang tepat waktu adalah 57 dan yang tidak tepat waktu adalah 244. Namun jika dilihat dari proporsi alumni yang tepat waktu dari kelulusan 2000-2017 untuk semua jurusan di jenjang sarjana adalah sebagai berikut ;



Gambar 5.4 *Proporsi Mahasiswa Tepat Waktu Berdasarkan Jurusan*

Gambar di atas menjelaskan tentang proporsi ketepatan waktu studi berdasarkan jurusan yang ada di UII, didapat bahwa jurusan dengan jumlah proporsi alumni tepat waktu terbanyak adalah Ekonomi Islam dengan proporsi 61,5% mahasiswanya lulus tepat waktu dan kemudian diikuti oleh Kedokteran. Sedangkan untuk proporsi jurusan dengan jumlah proporsi alumni tepat waktu

terkecil adalah Teknik Sipil dengan proporsi 3,9% mahasiswanya lulus tepat waktu dan kemudian diikuti oleh Teknik Industri.

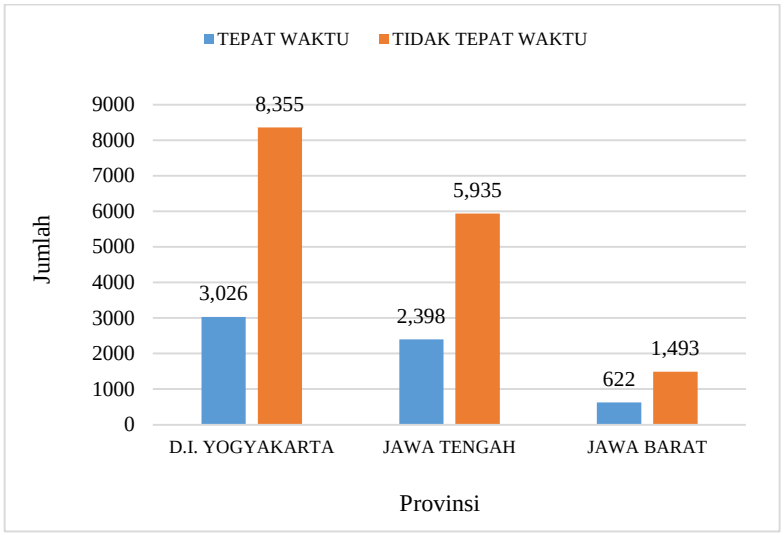


Gambar 5.5 *Jumlah Alumni Berdasarkan Jurusan SMA/Sederajat Sebelumnya*

Gambar di atas adalah menjelaskan tentang jumlah asal jurusan di bangku SMA/Sederajat sebelumnya. Dapat dilihat bahwa alumni yang berasal dari SMA/Sederajat yang paling banyak adalah dari jurusan IPA sebanyak 18.826, lalu jurusan IPS sebanyak 11.141, dan jurusan lain-lain sebanyak 1.821 alumni. Alumni yang berasal dari jurusan IPA pada SMA/Sederajat sebelumnya yang tepat waktu lulus sebanyak 5.341 dan tidak tepat waktu lulus sebanyak 13.485. Alumni yang berasal dari jurusan IPS pada SMA/Sederajat sebelumnya yang tepat waktu lulus sebanyak 2.997 dan tidak tepat waktu lulus sebanyak 8.144. Sedangkan alumni yang berasal dari jurusan lain-lain pada SMA/Sederajat sebelumnya yang tepat waktu lulus sebanyak 582 dan tidak tepat waktu lulus sebanyak 1.239. Alumni tersebut tentunya berasal dari berbagai daerah yang ada di Indonesia, berikut ini adalah tabel jumlah ketepatan alumni lulus berdasarkan provinsi asal ;

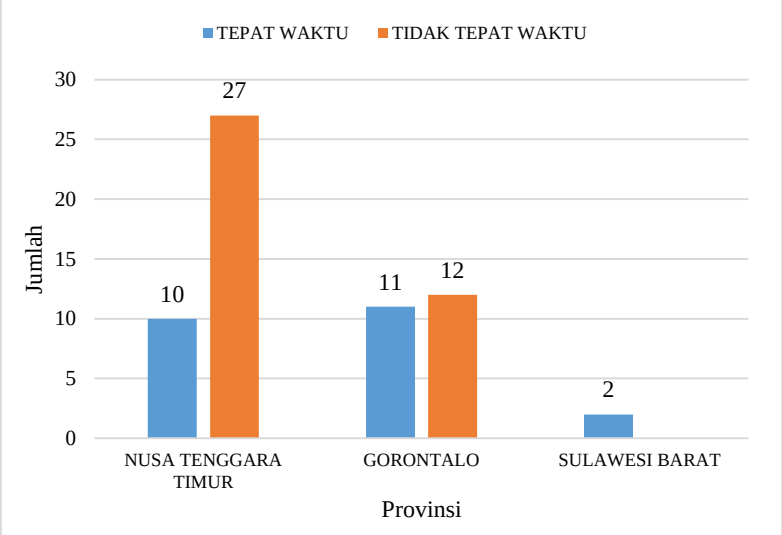
Table 5.1 *Ketepatan Alumni Berdasarkan Provinsi Asal*

Provinsi	Tepat Waktu	Tidak Tepat Waktu	Total
D.I. YOGYAKARTA	3026	8355	11.381
JAWA TENGAH	2398	5935	8.333
JAWA BARAT	622	1493	2.115
JAWA TIMUR	406	950	1.356
RIAU	365	920	1.285
KALIMANTAN TIMUR	280	759	1.039
LAMPUNG	209	487	696
SUMATERA SELATAN	193	479	672
NUSA TENGGARA BARAT	165	371	536
DKI JAKARTA	126	339	465
KALIMANTAN BARAT	138	315	453
JAMBI	122	324	446
KALIMANTAN SELATAN	119	312	431
BENGKULU	115	304	419
BANTEN	132	204	336
SUMATERA UTARA	63	209	272
KALIMANTAN TENGAH	63	169	232
KEPULAUAN RIAU	58	101	159
NANGGROE ACEH DARUSSALAM	37	113	150
SUMATERA BARAT	37	108	145
SULAWESI SELATAN	44	93	137
SULAWESI TENGGARA	35	85	120
BALI	30	81	111
KEPULAUAN BANGKA BELITUNG	35	66	101
PAPUA	22	61	83
SULAWESI TENGAH	25	58	83
MALUKU	6	38	44
MALUKU UTARA	5	38	43
SULAWESI UTARA	11	31	42
PAPUA BARAT	10	31	41
NUSA TENGGARA TIMUR	10	27	37
GORONTALO	11	12	23
SULAWESI BARAT	2		2



Gambar 5.6 *3 Provinsi Asal Alumni dengan Jumlah Terbanyak*

Provinsi asal alumni terbanyak pertama adalah dari D.I. Yogyakarta dengan jumlah alumni tepat waktu sebanyak 3.026 dan tidak tepat waktu sebanyak 8.355, terbanyak kedua adalah Jawa Tengah dengan jumlah alumni tepat waktu sebanyak 2.398 dan tidak tepat waktu sebanyak 5.935, dan terbanyak ketiga adalah Jawa Barat dengan jumlah alumni tepat waktu sebanyak 622 dan tidak tepat waktu sebanyak 1.493. Adapun 3 provinsi asal alumni dengan jumlah paling sedikit adalah sebagai berikut ini ;

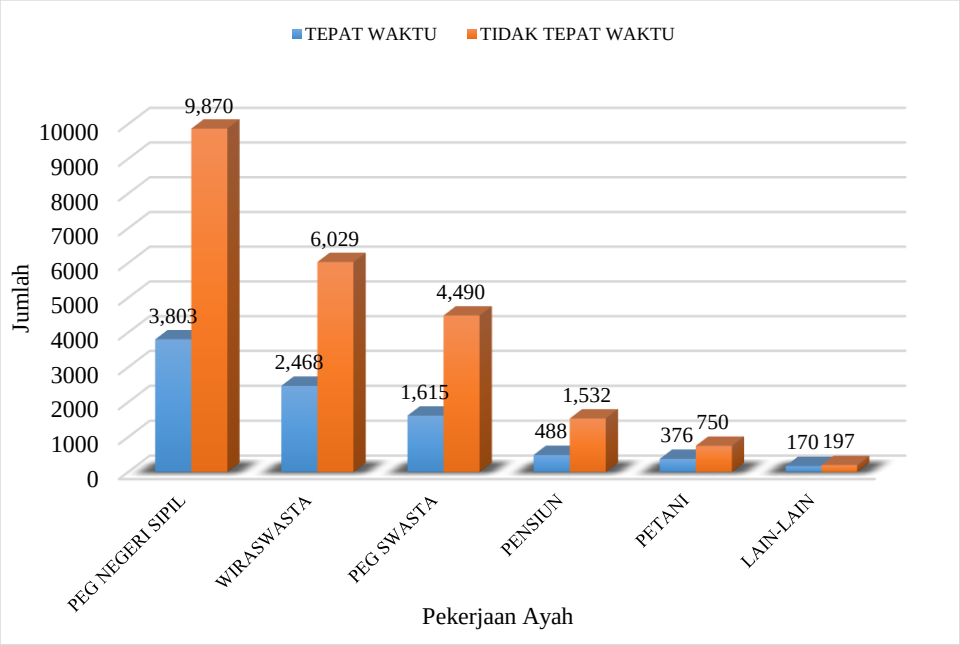


Gambar 5.7 *3 Provinsi Asal Alumni dengan Jumlah Paling Sedikit*

Berdasarkan gambar di atas, dapat dilihat bahwa provinsi asal alumni dengan jumlah paling sedikit adalah Sulawesi Barat dengan jumlah alumni hanya sebanyak

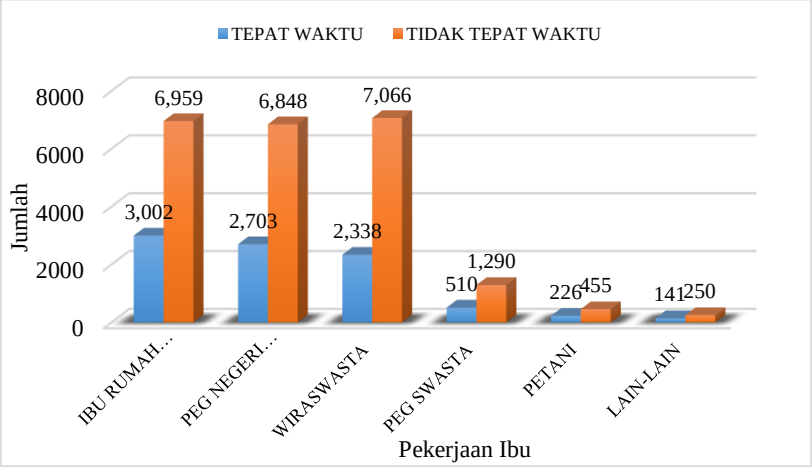
2 alumni dan semuanya tepat waktu, kemudia Gorontalo dengan jumlah alumni tepat waktu sebanyak 11 dan tidak tepat waktu sebanyak 12, dan selanjutnya Nusa Tenggara Timur dengan jumlah alumni tepat waktu sebanyak 10 dan tidak tepat waktu sebanyak 27.

Analisis deskriptif di atas sebelumnya adalah analisis deskriptif dari segi latar belakang individu alumni itu sendiri, kemudian analisis deskriptif selanjutnya adalah analisis berdasarkan latar belakang dari orang tua alumni. Latar belakang orang tua alumni dibagi menjadi 4, yaitu pekerjaan ayah, pekerjaan ibu, pendidikan terakhir ayah, dan pendidikan terakhir ibu. Jumlah alumni berdasarkan pekerjaan ayah adalah ;



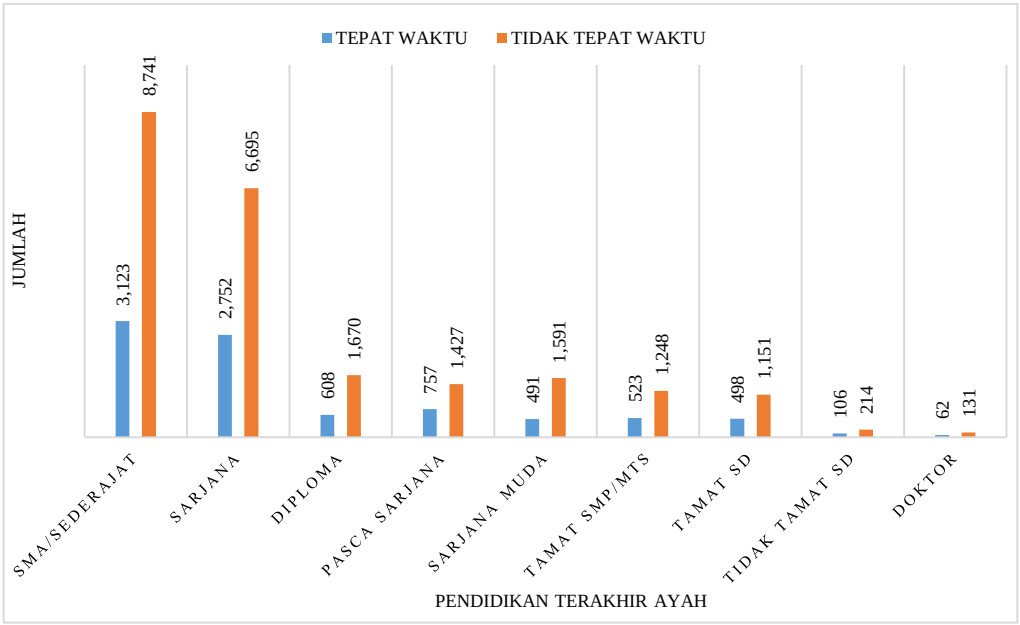
Gambar 5.8 *Jumlah Alumni Berdasarkan Pekerjaan Ayah*

Pada gambar 5.7 menerangkan tentang jumlah alumni berdasarkan pekerjaan ayahnya. Pekerjaan ayah terbanyak adalah pegawai negeri sipil dengan jumlah 13.673 alumni, kemudian pekerjaan wiraswasta dengan jumlah 8.497 alumni, kemudian pekerjaan pegawai swasta dengan jumlah 6.105 alumni, kemudian pekerjaan pension dengan jumlah 2.020, kemudian pekerjaan petani dengan jumlah 1.126 alumni, dan pekerjaan lain-lain dengan jumlah 367 alumni. Adapun jumlah alumni berdasarkan pekerjaan ibu adalah sebagai berikut ini ;



Gambar 5.9 *Jumlah Alumni Berdasarkan Pekerjaan Ibu*

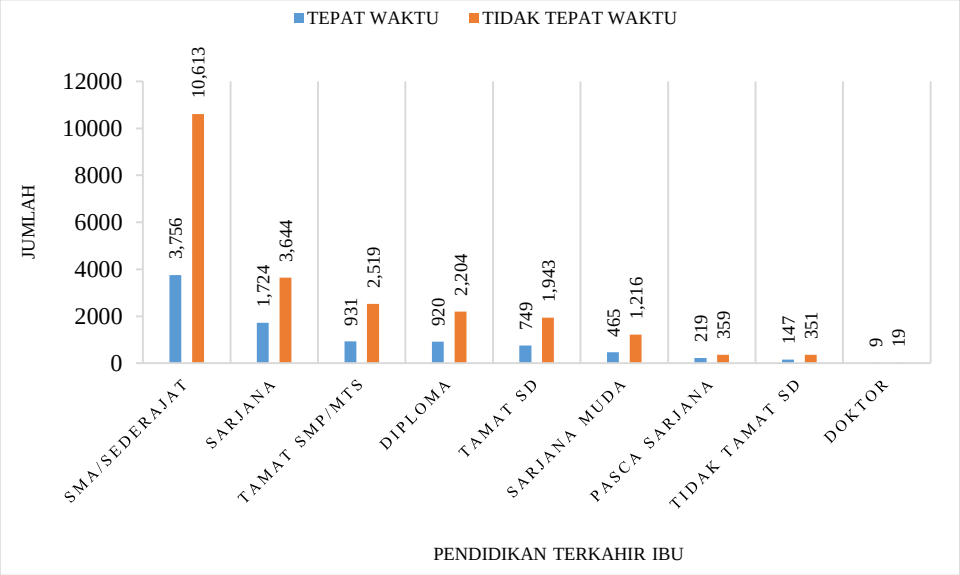
Berdasarkan gambar di atas, jumlah alumni berdasarkan jenis pekerjaan ibu terbanyak adalah pekerjaan ibu rumah tangga, yaitu sebanyak 9.961 alumni, kemudian pekerjaan sebagai pegawai negeri sipil sebanyak 9.551 alumni, kemudian pekerjaan sebagai wiraswasta sebanyak 9.404 alumni, kemudian pekerjaan sebagai pegawai swasta sebanyak 1.800 alumni, kemudian pekerjaan sebagai petani sebanyak 681 alumni, dan pekerjaan lain-lain sebanyak 391 alumni. Lalu jumlah alumni berdasarkan pekerjaan terakhir ayah adalah sebagai berikut ;



Gambar 5.10 *Jumlah Alumni Berdasarkan Pendidikan Terakhir Ayah*

Gambar di atas menjelaskan tentang jumlah alumni berdasarkan pendidikan terakhir ayah. Pendidikan terakhir ayah terbanyak adalah SMA/Sederajat dengan

jumlah sebanyak 11.864 alumni, kemudian diikut dengan pendidikan terakhir sarjana dengan jumlah sebanyak 9.447 alumni. Adapun jumlah alumni berdasarkan pendidikan terakhir ayah yang paling sedikit adalah doktor dengan jumlah sebanyak 193 alumni dan kemudian pendidikan terakhir yang tidak tamat SD adalah sebanyak 320 alumni. Sedangkan untuk pendidikan terakhir ibu adalah sebagai berikut ;



Gambar 5.11 *Jumlah Alumni Berdasarkan Pendidikan Terakhir Ibu*

Berdasarkan gambar di atas, jumlah alumni dengan pendidikan terakhir ibu terbanyak adalah SMA/Sederajat dengan jumlah sebanyak 14.369 alumni, kemudian diikuti oleh sarjana dengan jumlah sebanyak 5.368 alumni. Adapun jumlah alumni berdasarkan pendidikan terakhir paling sedikit adalah doktor dengan jumlah 28 alumni dan kemudian tidak tamat SD dengan jumlah 498 alumni.

5.2 Persiapan Data *Training* dan Data *Testing*

Data yang akan diolah untu analisis terdiri dari 2 kelas, yaitu tepat waktu dan tidak tepat waktu dengan jumlah masing-masing kelas sebanyak 8.920 dan 22.868. perbandingan data antara 2 kelas tersebut sangat jauh, sehingga data dikatakan *imbalance* atau tidak seimbang. Peneliti ingin melihat tentang hasil analisis untuk klasifikasi menggunakan data yang tidak seimbang tersebut (data asli) dan data yang sudah diseimbangkan jumlahnya (*balancing* data). Untuk melakukan analisis klasifikasi, data perlu dibagi terlebih dahulu menjadi 2 yaitu data *training* dan data *testing*. Data *training* digunakan untuk melatih program melakukan analisis

klasifikasi dan data *testing* digunakan untuk menguji apakah hasil analisis dari data *training* sebelumnya memiliki tingkat akurasi yang besar atau tidak, dengan kata lain adalah untuk mengetahui seberapa besar tingkat ketepatan metode dalam melakukan klasifikasi data.

5.2.1 Data Asli

Pembagian jumlah data untuk data *training* dan data *testing* dapat dilihat dari tabel berikut;

Tabel 5.2 *Data Asli*

Tepat waktu	Tidak tepat waktu
8.920	22.868

Tabel 5.3 *Data Training dan Data Testing Pada Data Asli*

Keterangan	Data *Training*	Data *Testing*	Total
Jumlah	23.841	7.947	31.788
Persentase	75%	25%	100%

Berdasarkan tabel 5.3, total data yang digunakan pada penelitian ini adalah 31.788, dengan komposisi 75% untuk data *training*/latih yaitu sebesar 23.841 data, dan sisanya 7.947 untuk data *testing*/uji coba. Pembagian data *training* dan *testing* tersebut dilakukan secara *random*/acak oleh bantuan *software* RStudio dan diapat dalam data *training* yang mengandung kelas tepat waktu sebanyak 6.659 dan kelas yang tidak tepat waktu sebanyak 17.182, sedangkan untuk data *testing* kelas tepat waktu sebanyak 2.261 dan kelas tidak tepat waktu sebanyak 5.686.

5.2.2 *Balancing Data*

Balancing data menggunakan metode *under sampling* menghasilkan jumlah untuk masing-masing kelas yang seimbang, yaitu dengan cara mengambil sampel secara acak (*random sampling*) sebanyak data pada kelas yang lebih sedikit. Pada penelitian ini jumlah kelas yang kecil ada pada kelas tepat waktu dan untuk jumlah kelas yang besar adalah kelas tidak tepat waktu. Sehingga under sampling bekerja

dengan mengambil sampel secara acak pada data kelas yang tidak tepat waktu sebanyak dengan kelas yang tepat waktu.

Tabel 5.4 *Balancing Data*

Tepat waktu	Tidak tepat waktu
8.920	8.920

Tabel 5.5 *Data Training dan Data Testing Setelah Balancing Data*

Keterangan	Data *Training*	Data *Testing*	Total
Jumlah	13.380	4.460	17.840
Persentase	75%	25%	100%

Berdasarkan tabel 5.5, total data yang digunakan pada penelitian ini adalah 17.840, dengan komposisi 75% untuk data *training*/latih yaitu sebesar 13.380 data, dan sisanya 4.460 untuk data *testing*/uji coba. Pembagian data *training* dan *testing* tersebut dilakukan secara *random*/acak oleh bantuan *software* RStudio dan diapat dalam data *training* yang mengandung kelas tepat waktu sebanyak 6.676 dan kelas yang tidak tepat waktu sebanyak 6.704, sedangkan untuk data *testing* kelas tepat waktu sebanyak 2.244 dan kelas tidak tepat waktu sebanyak 2.216.

5.3 Analisis Klasifikasi *Support Vector Machine* (SVM)

Klasifikasi menggunakan metode SVM adalah dengan mencari garis pemisah (*hyperplane*) antar hasil klasifikasi. Data alumni yang digunakan adalah data yang terdiri dari berbagai variabel kategorik dan *numeric*, serta data tersebut tidak *linier*, maka metode SVM yang digunakan dalam penelitian ini adalah menggunakan *Kernel Radian Basis Function* (RBF). Metode RBF ini memerlukan nilai parameter *cost* (C) dan *gamma* yang nilailnya ditentukan oleh peneliti. Oleh karena itu, peneliti di sini menggunakan uji coba nilai C sebanyak 5, yaitu 0.1, 1, 5, 10, dan 50, serta uji coba nilai *gamma* sebanyak 4, yaitu 1, 2, 3,dan 4.

5.3.1 Analisis SVM dengan Data Asli

Berikut ini adalah nilai *error* dari masing-masing nilai C dan *gamma* untuk data asli;

Tabel 5.6 *Nilai Error C dan Gamma Kernel RBF Data Asli*

SVM Kernel RBF				SVM Kernel RBF			
Cost	*Gamma*	*Accuracy*	*Error*	*Cost*	*Gamma*	*Accuracy*	*Error*
0,1	1	0,7239632	0,2760368	0,1	3	0,7207335	0,2792665
1	1	0,7738354	0,2261646	1	3	0,7309258	0,2690742
5	1	0,7658662	0,2341338	5	3	0,7285351	0,2714649
10	1	0,7635173	0,2364827	10	3	0,7278640	0,2721360
50	1	0,7584837	0,2415163	50	3	0,7266895	0,2733105
0,1	2	0,7209851	0,2790149	0,1	4	0,7206915	0,2793085
1	2	0,6347010	0,3652990	1	4	0,7296255	0,2703745
5	2	0,7317231	0,2682769	5	4	0,7279058	0,2720942
10	2	0,7310939	0,2689061	10	4	0,7272767	0,2727233
50	2	0,7294579	0,2705421	50	4	0,7256827	0,2743173

Tabel di atas menunjukan bahwa metode SVM menggunakan *kernel* RBF menghasilkan nilai *error* paling kecil untuk parameter C=1 dan *gamma*=1, sedangkan untuk hasil *error* paling besar adalah untuk C=1 dan gamma=2, sehingga nilai C dan *gamma* yang digunakan untuk nilai parameter di *kernel* RBF adalah yang menghasilkan *error* paling kecil, yaitu C=1 dan *gamma*=1, serta nilai akurasi untuk model pada data *train* adalah sebagai berikut;

Tabel 5.7 *Hasil Prediksi Data Asli SVM Kernel RBF Data Train*

SVM Kernel RBF		
Prediksi	Data *Train*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	5845	317
Tidak Tepat Waktu	814	16865

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *train* adalah sebanyak 6.659, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 5.845 dan prediksi salah sebanyak 814. Kemudian untuk alumni yang tidak tepat waktu dari data *train* adalah sebanyak 17.182, hasil prediksi benar untuk tidak tepat waktu adalah 16.865 dan prediksi salah adalah sebanyak 317.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{5.845+16.865}{23.841} = 0.953 \text{ atau } 95{,}3\%.$$

Setelah mendapatkan nilai C dan *gamma* optimal pada data *train*, peneliti akan memasukan nilai tersebut untuk menguji tingkat akurasinya terhadap data *testing* atau data baru. Berikut ini adalah hasil klasifikasi SVM *kernel* RBF data *testing*;

Tabel 5.8 *Hasil Prediksi Data Asli SVM Kernel RBF Data Test*

SVM Kernel RBF		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	816	386
Tidak Tepat Waktu	1.445	5.300

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.261, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 816 dan prediksi salah sebanyak 1.445. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 5.686, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 5.300 dan prediksi salah adalah sebanyak 386.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{816+5.300}{7.947} \qquad = 0.769 \text{ atau } 76.9\%$$

Hasil akurasi untuk data *testing* diperoleh senilai 0.769 atau sebesar 77%. Hasil akurasi data *testing* sebesar 77% menunjukan bahwa metode SVM dengan *kernel* RBF untuk melakukan prediksi terhadap data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 77% dan prediksi salah sebesar 23% atau bisa dikatakan bisa memprediksi dengan bagus.

Selanjutnya adalah analisis klasifikasi SVM menggunakan kernel sigmoid. Kernel sigmoid ini juga memerlukan nilai parameter *cost* (C) dan *gamma* yang nilailnya ditentukan oleh peneliti seperti pada kernel RBF. Oleh karena itu, peneliti di sini menggunakan uji coba nilai C sebanyak 5, yaitu 0.1, 1, 5, 10, dan 50, serta uji coba nilai *gamma* sebanyak 4, yaitu 1, 2, 3,dan 4, agar sesuai dengan kernel RBF pada penjelasan sebelumnya. Berikut ini adalah nilai *error* yang didapat dari masing-masing nilai C dan *gamma* ;

Tabel 5.9 *Nilai Error C dan Gamma Kernel Sigmoid Data Asli*

SVM Kernel Sigmoid				SVM Kernel Sigmoid			
Cost	*Gamma*	*Accuracy*	*Error*	*Cost*	*Gamma*	*Accuracy*	*Error*
0,1	1	0,6890230	0,3109770	0,1	3	0,6685961	0,3314039
1	1	0,6869676	0,3130324	1	3	0,6680089	0,3319911
5	1	0,6898619	0,3101381	5	3	0,6678830	0,3321170
10	1	0,6902394	0,3097606	10	3	0,6679669	0,3320331
50	1	0,6878484	0,3121516	50	3	0,6678411	0,3321589
0,1	2	0,6831933	0,3168067	0,1	4	0,6664150	0,3335850
1	2	0,6828568	0,3171432	1	4	0,6662892	0,3337108
5	2	0,6826899	0,3173101	5	4	0,6664150	0,3335850
10	2	0,6826899	0,3173101	10	4	0,6664150	0,3335850
50	2	0,6826899	0,3173101	50	4	0,6663731	0,3336269

Tabel di atas menunjukan bahwa metode SVM menggunakan *kernel* Sigmoid menghasilkan nilai *error* paling kecil untuk parameter C=10 dan *gamma*=1, sedangkan untuk hasil *error* paling besar adalah untuk C=1 dan gamma=4, sehingga nilai C dan *gamma* yang digunakan untuk nilai parameter di *kernel* sigmoid adalah yang menghasilkan *error* paling kecil, yaitu C=10 dan *gamma*=1, serta nilai akurasi untuk model pada data *test* adalah sebagai berikut;

Tabel 5.10 *Hasil Prediksi Data Asli SVM Kernel Sigmoid Data Train*

SVM Kernel Sigmoid		
Prediksi	Data *Train*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	3,002	3,657
Tidak Tepat Waktu	3,657	13,525

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *train* adalah sebanyak 6.659, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 3.002 dan prediksi salah sebanyak 3.657. Kemudian untuk alumni yang tidak tepat waktu dari data *train* adalah sebanyak 17.182, hasil prediksi benar untuk tidak tepat waktu adalah 13.525 dan prediksi salah adalah sebanyak 3.657.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{3.002+13.525}{23.841} = 0.693 \text{ atau } 69{,}3\%.$$

Setelah mendapatkan nilai C dan *gamma* optimal pada data *train*, peneliti akan memasukan nilai tersebut untuk menguji tingkat akurasinya terhadap data *testing*. Berikut ini adalah hasil klasifikasi SVM *kernel* Sigmoid data *testing*;

Tabel 5.11 *Hasil Prediksi Data Asli SVM Kernel Sigmoid Data Test*

SVM Kernel Sigmoid		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	998	1216
Tidak Tepat Waktu	1263	4470

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.261, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 998 dan prediksi salah sebanyak 1.263. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 5.686, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 4.470 dan prediksi salah adalah sebanyak 1.216.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{998+4.470}{7.947} \qquad = 0.688 \text{ atau } 68.8\%$$

Hasil akurasi untuk data *testing* diperoleh senilai 0.688 atau sebesar 68%. Hasil akurasi data *testing* sebesar 68% menunjukan bahwa metode SVM dengan *kernel* sigmoid untuk melakukan prediksi terhadap data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 68% dan prediksi salah sebesar 32% atau bisa dikatakan bisa memprediksi dengan cukup bagus.

5.3.2 Analisis SVM dengan *Balancing* Data

Berikut ini adalah nilai *error* dari masing-masing nilai C dan *gamma* untuk *balancing* data;

Tabel 5.12 *Nilai Error C dan Gamma Kernel RBF Data Balanced*

SVM Kernel RBF				SVM Kernel RBF			
Cost	*Gamma*	*Accuracy*	*Error*	*Cost*	*Gamma*	*Accuracy*	*Error*
0.1	1	0.6453662	0.3546338	0.1	3	0.4915546	0.5084454
1	1	0.7624066	0.2375934	1	3	0.6094918	0.3905082
5	1	0.7465620	0.2534380	5	3	0.6119581	0.3880419
10	1	0.7450673	0.2549327	10	3	0.6115097	0.3884903
50	1	0.7390135	0.2609865	50	3	0.6129297	0.3870703
0.1	2	0.4951420	0.5048580	0.1	4	0.4914051	0.5085949
1	2	0.6624066	0.3375934	1	4	0.5793722	0.4206278
5	2	0.6665770	0.3334230	5	4	0.5928999	0.4071001
10	2	0.6631540	0.3368460	10	4	0.5928251	0.4071749
50	2	0.6609865	0.3390135	50	4	0.5936472	0.4063528

Tabel di atas menunjukan bahwa metode SVM menggunakan *kernel* RBF menghasilkan nilai *error* paling kecil untuk parameter C=1 dan *gamma*=1, sedangkan untuk hasil *error* paling besar adalah untuk C=0.1 dan gamma=3, sehingga nilai C dan *gamma* yang digunakan untuk *kernel* RBF adalah yang menghasilkan *error* paling kecil, yaitu C=1 dan *gamma*=1, serta nilai akurasi untuk model pada data *train* dengan bantuan *software* Rstudio adalah sebesar 76.12%.

Setelah mendapatkan nilai C dan *gamma* optimal pada data *train*, peneliti akan memasukan nilai tersebut untuk menguji tingkat akurasinya terhadap data *testing* atau data baru. Berikut ini adalah hasil klasifikasi SVM *kernel* RBF data *testing*;

Tabel 5.13 *Hasil Prediksi Data Balanced SVM Kernel RBF Data Test*

SVM Kernel RBF		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	1,663	508
Tidak Tepat Waktu	553	1,736

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.216, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 1.663 dan prediksi salah sebanyak 553. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 2.244, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 1.736 dan prediksi salah adalah sebanyak 508.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{1.663+1.736}{4.460} = 0.7621 \text{ atau } 76.21\%$$

Hasil akurasi untuk data *testing* diperoleh senilai 0.7621 atau sebesar 76,21%. Hasil akurasi data *testing* sebesar 76,21% menunjukan bahwa metode SVM dengan *kernel* RBF untuk melakukan prediksi terhadap data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 76,21% dan prediksi salah sebesar 23,79% atau bisa dikatakan bisa memprediksi dengan bagus.

Selanjutnya adalah analisis klasifikasi SVM menggunakan kernel sigmoid. Kernel sigmoid ini juga memerlukan nilai parameter *cost* (C) dan *gamma* yang nilailnya ditentukan oleh peneliti seperti pada kernel RBF. Oleh karena itu, peneliti di sini menggunakan uji coba nilai C sebanyak 5, yaitu 0.1, 1, 5, 10, dan 50, serta uji coba nilai *gamma* sebanyak 4, yaitu 1, 2, 3,dan 4, agar sesuai dengan kernel RBF pada penjelasan sebelumnya. Berikut ini adalah nilai *error* yang didapat dari masing-masing nilai C dan *gamma* ;

Tabel 5.14 *Nilai Error C dan Gamma Kernel Sigmoid Data Balanced*

SVM Kernel Sigmoid				SVM Kernel Sigmoid			
Cost	*Gamma*	*Accuracy*	*Error*	*Cost*	*Gamma*	*Accuracy*	*Error*
0,1	1	0,6890230	0,3109770	0,1	3	0,6685961	0,3314039
1	1	0,6869676	0,3130324	1	3	0,6680089	0,3319911
5	1	0,6898619	0,3101381	5	3	0,6678830	0,3321170
10	1	0,6902394	0,3097606	10	3	0,6679669	0,3320331
50	1	0,6878484	0,3121516	50	3	0,6678411	0,3321589
0,1	2	0,6831933	0,3168067	0,1	4	0,6664150	0,3335850
1	2	0,6828568	0,3171432	1	4	0,6662892	0,3337108
5	2	0,6826899	0,3173101	5	4	0,6664150	0,3335850
10	2	0,6826899	0,3173101	10	4	0,6664150	0,3335850
50	2	0,6826899	0,3173101	50	4	0,6663731	0,3336269

Tabel di atas menunjukan bahwa metode SVM menggunakan *kernel* Sigmoid menghasilkan nilai *error* paling kecil untuk parameter C=10 dan *gamma*=1, sedangkan untuk hasil *error* paling besar adalah untuk C=1 dan gamma=4, sehingga nilai C dan *gamma* yang digunakan untuk nilai parameter di *kernel* sigmoid adalah yang menghasilkan *error* paling kecil, yaitu C=10 dan *gamma*=1, serta nilai akurasi untuk model pada data *train* dengan bantuan *software* Rstudio adalah 60,1%.

Setelah mendapatkan nilai C dan *gamma* optimal pada data *train*, peneliti akan memasukan nilai tersebut untuk menguji tingkat akurasinya terhadap data *testing*. Berikut ini adalah hasil klasifikasi SVM *kernel* Sigmoid data *testing*;

Tabel 5.15 *Hasil Prediksi Data Balanced SVM Kernel Sigmoid Data Test*

SVM Kernel Sigmoid		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	1,326	864
Tidak Tepat Waktu	890	1,380

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.216, hasil prediksi SVM benar untuk tepat waktu adalah sebanyak 1.326 dan prediksi salah sebanyak 890. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 2.244, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 1.380 dan prediksi salah adalah sebanyak 864.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{1.326+1.380}{4.460} = 0.6067 \text{ atau } 60{,}67\%$$

Hasil akurasi untuk data *testing* diperoleh senilai 0.6067 atau sebesar 60,7%. Hasil akurasi data *testing* sebesar 60,7% menunjukan bahwa metode SVM dengan *kernel* sigmoid untuk melakukan prediksi terhadap data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 60,7% dan prediksi salah sebesar 39,3% atau bisa dikatakan bisa memprediksi dengan cukup bagus.

5.4 Analisis Klasifikasi *Random Forest*

Analisis klasifikasi *Random Forest* dilakukan dengan menentukan berapa banyak pohon yang akan terbentu (*ntree*) dan menentukan berapa banyak *random sample* yang diambil untuk setiap percobaan (*mtry*). Peneliti menentukan nilai *ntree* dan *mtry* dengan melakukan uji coba untuk mendapatkan nilai optimal dari *ntree* dan *mtry*. Penentuan nilai *mtry* dilakukan dengan 3 cara, yaitu ;

1. $Mtry = \frac{\sqrt{\text{banyak variabel}}}{2}$

   $= \frac{\sqrt{10}}{2} = 1.94 = 2$

2. $Mtry = \sqrt{\text{banyak variabel}}$

   $= \sqrt{10} = 3{,}16 = 3$

3. $Mtry = \sqrt{\text{banyak variabel}} \times 2$

$= \sqrt{10}$ x 2 = 3,16 x 2 = 6,32 = 6

Dari rumus di atas, didapat bahwa nilai *mtry* adalah 2, 3, dan 6 yang selanjutnya digunakan untuk melakukan klasifikasi.

5.4.1 Analisis Random Forest dengan Data Asli

Berikut ini adalah nilai *error* untuk masing-masing nilai *mtry* dengan menggunakan *ntree* sebesar 500 pada data asli ;

Tabel 5.16 *Nilai Error Masing-masing Mtry Data Asli*

mtry	*error*
2	0,1972652
3	0,2006208
6	0,2035569

Dengan melihat tabel di atas, dapat diketahui bahwa nilai *mtry* terbaik atau optimum adalah sebesar 2, karena nilai *error* terkecil. Setelah menemukan nilai *mtry* optimum, selanjutnya mencari nilai *ntree* optimum dengan memasukan nilai *mtry* sebesar 2. Nilai *ntree* yang akan dicoba ada 5, yaitu 25, 50, 100, 500, dan 1000. Hasil dari nilai *error* yang didapat adalah sebagai berikut ;

Tabel 5.17 *Nilai Error Untuk Masing-Masing Ntree Data Asli*

ntree	*error*
25	0,0424
50	0,0354
100	0,0359
500	0,0347
1000	0,0354

Dapat dilihat dari tabel di atas, bahwa nilai *ntree* optimum atau terbaik adalah sebesar 500 dengan menghasilkan nilai *error* terkecil yaitu 0.0347 atau sebesar 3.47%. Setelah mendapatkan nilai *mtry* dan *ntree* optimum, maka kemudian nilai tersebut digunakan untuk menentukan prediksi *Random Forest* untuk data *training* dan kemudian akan diuji tingkat akurasinya terhadap data *testing* atau data baru. Berikut ini adalah hasil klasifikasi *Random Forest* pada data *testing* ;

Tabel 5.18 *Hasil Prediksi Data Asli Random Forest Data Testing*

Random Forest		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	1.217	534
Tidak Tepat Waktu	1.044	5.152

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.261, hasil prediksi benar untuk tepat waktu adalah sebanyak 1.217 dan prediksi salah sebanyak 1.044. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 5.686, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 5.152 dan prediksi salah adalah sebanyak 534.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{1.217+5.152}{7.947} \quad = 0{,}801$$

Hasil akurasi untuk data *testing* diperoleh senilai 0,801 atau sebesar 80%. Hasil akurasi data *testing* sebesar 80% menunjukan bahwa metode *Random Forest* untuk melakukan prediksi dari data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 80% atau bisa dikatakan bisa memprediksi dengan bagus.

5.4.2 Analisis Random Forest dengan Data *Balanced*

Berikut ini adalah nilai *error* untuk masing-masing nilai *mtry* dengan menggunakan *ntree* sebesar 500 pada data *balanced* ;

Tabel 5.18 *Nilai Error Masing-masing Mtry Data Balanced*

mtry	*error*
2	0,2245142
3	0,2266816
6	0,2321375

Dengan melihat tabel di atas, dapat diketahui bahwa nilai *mtry* terbaik atau optimum adalah sebesar 2, karena nilai *error* terkecil. Setelah menemukan nilai *mtry* optimum, selanjutnya mencari nilai *ntree* optimum dengan memasukan nilai *mtry* sebesar 2. Nilai *ntree* yang akan dicoba ada 5, yaitu 25, 50, 100, 500, dan 1000. Hasil dari nilai *error* yang didapat adalah sebagai berikut ;

Tabel 5.20 *Nilai Error Untuk Masing-Masing Ntree Data Balanced*

ntree	*error*
25	0.0416
50	0.0383
100	0.0328
500	0.0315
1.000	0.0309

Dapat dilihat dari tabel di atas, bahwa nilai *ntree* optimum atau terbaik adalah sebesar 1.000 dengan menghasilkan nilai *error* terkecil yaitu 0.0309 atau sebesar 3.09%. Setelah mendapatkan nilai *mtry* dan *ntree* optimum, maka kemudian nilai tersebut digunakan untuk menentukan prediksi *Random Forest* untuk data *training* dan kemudian akan diuji tingkat akurasinya terhadap data *testing* atau data baru. Berikut ini adalah hasil klasifikasi *Random Forest* pada data *testing* ;

Tabel 5.21 *Hasil Prediksi Data Balanced Random Forest Data Testing*

Random Forest		
Prediksi	Data *Test*	
	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	1,605	339
Tidak Tepat Waktu	611	1,905

Berdasarkan tabel di atas, banyaknya alumni yang tepat waktu dari data *testing* adalah sebanyak 2.216, hasil prediksi benar untuk tepat waktu adalah sebanyak 1.605 dan prediksi salah sebanyak 611. Kemudian untuk alumni yang tidak tepat waktu dari data *testing* adalah sebanyak 2.244, hasil prediksi benar untuk tidak tepat waktu adalah sebanyak 1.905 dan prediksi salah adalah sebanyak 339.

$$\text{Akurasi} = \frac{\sum(\text{prediksi benar})}{\sum(\text{semua prediksi})}$$

$$= \frac{1.605+1.905}{4.460} \quad = 0{,}7899 \text{ atau } 78{,}69\%$$

Hasil akurasi untuk data *testing* diperoleh senilai 0,7869 atau sebesar 78,7%. Hasil akurasi data *testing* sebesar 78,7% menunjukan bahwa metode *Random Forest* untuk melakukan prediksi dari data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 78,7% atau bisa dikatakan bisa memprediksi dengan bagus.

5.4.3 *Importance Variable*

Pada analisis *random forest* menggunakan data asli dan data *balanced* menghasilkan akurasi terbesar di data asli, maka dari itu untu kmenentukan *importance variable* dilakukan menggunakan data asli.

Tabel 5.22 *Importance Variable Random Forest*

Variabel	*Mean Decrease Accuracy*
IPK	231.69
Prodi	157.80
Jenis Kelamin	28.93
Jurusan SMA	24.73
Pekerjaan Ayah	15.41
Pekerjaan Ibu	11.64
Pendidikan Ibu	11.32
Propinsi	10.82
Pendidikan Ayah	9.37

Tabel di atas menjelaskan tentang *importance variable* dari klasifikasi *random forest* yang telah dilakukan. *Importance variable* menunjukan hubungan *variable* itu sendiri dalam mempengaruhi hasil analisis/prediksi. Semakin besar angka *importance variable*, menunjukan semakin besar juga peran *variable* tersebut dalam mempengaruhi hasil analisis. Adapun yang paling besar peran *variable*nya secara berurutan adalah IPK, Prodi, Jenis Klmn, Jurusan SMA, Pekerjaan Ayah, Pekerjaan Ibu, Pendidiikan Ibu, Provinsi, dan Pendidikan Ayah.

5.5 Perbandingan Metode *SVM* dan *Random Forest*

Setelah dilakukan analisis klasifikasi menggunakan SVM kernel RBF dan kernel sigmoid, serta *Random Forest*, didapatkanlah nilai akurasi dari masing-masing metode dan berbeda-beda. Berikut ini adalah perbandingan tingkat akurasi dari 2 metode di atas menggunakan data asli ;

Tabel 5.23 *Perbandingan Akurasi SVM dan Random Forest Data Asli*

Support Vector Machine		*Random Forest*
Kernel RBF	Kernel Sigmoid	
77%	68%	80%

Tabel 5.24 *Perbandingan Akurasi SVM dan Random Forest Data Balanced*

Support Vector Machine		*Random Forest*
Kernel RBF	Kernel Sigmoid	
77,21%	60,67%	78,7%

Tabel 5.7 di atas menjelaskan tentang perbandingan tingkat akurasi dari metode SVM dan *Random Forest* antara data asli dan data *balanced*, mendapatkan bahwa data terbaik adalah data asli dengan hasil klasifikasi data alumni untuk mengetahui ketepatan lama studi adalah klasifikasi menggunakan metode *Random Forest* dengan nilai akurasi sebesar 80%, dimana metode *Random Forest* tersebut menggunakan nilai *mtry*=2 dan *ntree*=500.

BAB VI
KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut :

1. Banyaknya alumni Universitas Islam Indonesia dari tahun kelulusan 2000-2017 terdapat 31.788 alumni yang mana alumni tepat waktu sebanyak 28% dan yang tidak tepat waktu sebanyak 78%. Jurusan dengan mahasiswa tepat waktu terbesar adalah jurusan Ekonomi Islam sebesar 61,5% dan jurusan dengan mahasiswa tepat waktu terkecil adalah Tehnik Sipil sebesar 3,9%.
2. Klasifikasi menggunakan metode SVM dibagi menjadi 2, yaitu SVM kernel RBF dan SVM kernel sigmoid. Akurasi SVM kernel RBF dengan parameter optimum C=1 dan *gamma* =1 adalah sebesar 77%. Akurasi SVM kernel sigmoid dengan parameter optimum C= 10 dan *gamma*= 1 adalah sebesar 68%. Lalu tingkat akurasi untuk metode *Random Forest* dengan nilai optimum m= 2 dan k=500 sebesar 80%. Semua nilai akurasi tersebut didapat menggunakan analisis pada data asli yang berupa *imbalance* data.
3. Penelitian ini menghasilkan metode terbaik untuk melakukan analisis klasifikasi data ketepatan lama studi alumni Universitas Islam Indonesia tahun kelulusan 2000-2017 adalah metode *Random Forest* dengan nilai optimum m= 2 dan k=500 sebesar 80%, yang berarti ketepatan metode klasifikasi *Random Forest* untuk data ini sudah sangat bagus.

6.2 Saran

Saran yang dapat diberikan dari hasil penelitian ini adalah :

1. Pada penelitian selanjutnya diharapkan peneliti mampu mengetahui permasalahan terbesar yang menyebabkan mahasiswa tidak tepat waktu sehingga datap dicari solusi untuk meningkatkan ketepatan studi.
2. Untuk Universitas Islam Indonesia, setelah mendapat kesimpulan bahwa banyak mahasiswa yang lama studinya tidak tepat waktu maka disarankan untuk membuat aplikasi (*software*) untuk mengingatkan tentang lama studi mahasiswa dan batasannya, serta *software* tersebut memberikan rekomendasi apa yang perlu dilakukan oleh mahasiswa agar lama studi bisa tepat waktu (misalnya rekomendasi mengulang mata kuliah A di semester 5, atau memberitahu untuk melakukan *test* CEPT).

DAFTAR PUSTAKA

Amalia, H. (2018). Perbandingan Metode Data Mining SVM dan NN Untuk Klasifikasi Penyakit Ginjal Kronis. *Jurnal PILAR Nusa Mandiri Vol.14 No.1*.

Analytics Vidhya. (2016). *A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)*. Diambil kembali dari https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

Analytics Vidhya Content Team. (2016). *Practical Guide to deal with Imbalanced Classification Problems in R*. Diambil kembali dari https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Assaffat. (2015). Analisis Akurasi Support Vector Machine Dengan Fungsi Kernel Gaussian RBF Untuk Prakiraan Beban Listrik Harian Sektor Industri. *Jurnal Momentum*, Vol. 11 (2).

Breiman, L. (2001). *Random Forest Machine Learning.* Belanda: Kluwer Academic Publisher.

Chandra, A. (2017). *Perbedaan Supervised dan Unsupervised Learning*. Diambil kembali dari https://www.datascience.or.id/detail_artikel/52/supervised-and-unsupervised-learning

Daqiqil, I., Astread, & Mahdiyah, E. (2017). Deteksi dan Perbandingan Kinerja Algoritma Random Forest dan Boosted C5.0. *Jurnal Teknologi dan Sistem Informasi*.

Estoatnowo, D. (2016). Klasifikasi Status Kelulusan Mahasiswa Menggunakan Metode CHAID dan Algoritma C4.5 (Studi Kasus : Mahasiswa Jurusan Statistika Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia). *Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Islam Indonesia.* Yogyakarta.

Gorunescu, F. (2011). *Data Mining : Concepts, Models, and Techniques.* New York: Springer.

Hastie, Tibshirani, & Friedman. (2008). *The Elements of Statistical Learning: Data-mining, Inference and Prediction. Second Edition.* New York: Springer-Verlag.

Jain, K. (2015). *Machine Learning Basics for a Newbie*. Diambil kembali dari https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/

Kementerian Riset, Teknologi dan Pendidikan Tinggi. (2018). *Grafik Jumlah Perguruan Tinggi*. Diambil kembali dari https://forlap.ristekdikti.go.id/perguruantinggi/homegraphpt

Naradhipa, A. R., & Purwarianti, A. (2011). Sentiment Classification for Indonesian Message in Social Media. *International Conference on Electrical Engineering and Informatics*.

Nugraha, Y. S., & Emiliyawati, N. (2017). Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest. *Jurnal Teknik Elektro Vol. 9 No.1*.

Nugroho, A. S., Wranto, A. B., & Handoko, D. (2003). Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika. *IlmuKomputer.com*.

Octaviani, P. A., Wilandari, Y., & Ispriyanti, D. (2014). Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. *Jurnal Gaussian Vol.3 No.4*, 811-820.

Perdana, R. R., Soelaiman, R., & Fatichah, C. (2017). Implementasi Ekstraksi Fitur untuk Pengelompokan Berkas Musik Berdasarkan Kemiripan Karakteristik Suara. *Jurnal Teknik ITS Vol.6 No.1*.

PERMENDIKNAS. (2000). *Keputusan Menteri Pendidikan Nasional Republik Indonesia No.232.* Jakarta: Depdiknas.

Pratiwi, Y. R. (2017). Perbandingan Analisis Sentimen Pada Pertalite Melalui Jejaring Sosial Twitter dengan Menggunakan Metode Support Vector Machine dan Maximum Entropy. *Skipsi. Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.* Yogyakarta.

Rachman, F., & Purnami, S. W. (2012). Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer dengan Menggunakan Regresi Logistik Ordinal dan Support Vector Machine (SVM). *Jurnal Sains dan Seni ITS Vol.1 No.1*.

Redaksi. (2014). *7 Tahap Menyusun Data Collection Plan untuk Perbaikan di Lini Produksi*. Diambil kembali dari Shift Indonesia: http://shiftindonesia.com/7-tahap-menyusun-data-collection-plan-untuk-perbaikan-di-lini-produksi/

Republik Indonesia. (2012). *Undang-Undang Republik Indonesia Nomor 12 pasal 1.* Jakarta: Sekretariat Negara.

Republik Indonesia. (2012). *Undang-Undang Republik Indonesia Nomor 12 pasal 4.* Jakarta: Sekretariat Negara.

Republik Indonesia. (2012). *Undang-Undang Republik Indonesia Nomor 12 pasal 5.* Jakarta: Sekretariat Negara.

Sartono, B., & Syafitri, U. D. (2010). Metode Pohon Gabungan: Solusi Pilihan Untuk Mengatasi Kelemahan Pohon Regresi. *Forum Statistika dan Komputasi*, 1-7.

Sasongko, T. B. (2016). Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA). *Jurnal Teknik Informatika dan Sistem Informasi*, Volume 2 Nomor 2.

Sastrawan, A. S., Baizal, Z. A., & Bijaksana, M. A. (2010). Analisis Pengaruh Metode Combine Sampling Dalam Churn Prediction Untuk Perusahaan Telekomunikasi. *Seminar Nasional Informatika 2010 (semnasIF 2010) UPN "Veteran" Yogyakarta*.

Subhan, A., & Fanani, A. Z. (2017). Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian dengan Menggunakan Algoritma Naice Bayes. *Skripsi. Fakultas Teknik Informatika Universitas Dian Nuswantoro.* Semarang.

Tawakal, M. I. (2015). *Apa yang dimaksud dengan Data Cleansing (Data Scrubbing) ?* Diambil kembali dari https://www.dictio.id/t/apa-yang-dimaksud-dengan-data-cleansing-data-scrubbing/15064

Untari, D. (2014). *Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5*. Semarang: Program Studi Teknik Informatika - Fakultas Ilmu Komputer - Universitas Dian Nuswantoro.

Zhu, M. (2008). Kernels and Ensembles: Perspectives on Statistical Learning. Dalam *The American Statistician 62: 97 – 109.*

LAMPIRAN

Lampiran 1. *Data Penelitian*

No	Jenis Klmn	Prodi	Propinsi	Pekerjaan Ayah	Pekerjaan Ibu	Pendidikan Ayah	Pendidikan Ibu	Jurusan Sma	Tgl Terdaftar	Tgl Yudicium	IPK	KET
1	L	Manajemen	JAWA TENGAH	PENSIUN	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	26-Mar-2005	3.01	TIDAK TEPAT WAKTU
2	P	Manajemen	D.I. YOGYAKARTA	USAHA SENDIRI	PEG SWASTA	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPA	1-Sep-2000	24-Mar-2007	2.92	TIDAK TEPAT WAKTU
3	L	Manajemen	D.I. YOGYAKARTA	USAHA SENDIRI	USAHA SENDIRI	SARJANA MUDA	SARJANA	IPS	1-Sep-2000	4-Dec-2004	3.08	TIDAK TEPAT WAKTU
4	L	Manajemen	JAWA TENGAH	PEG NEGERI SIPIL	PEG NEGERI SIPIL	SARJANA MUDA	DIPLOMA	IPA	1-Sep-2000	27-Jul-2005	3.02	TIDAK TEPAT WAKTU
5	P	Manajemen	JAWA BARAT	PEG SWASTA	WIRASWASTA	TAMAT SMA/SMK/SMEA/STM	TIDAK TAMAT SD	IPS	4-Sep-2000	10-Apr-2004	3.55	TEPAT WAKTU
6	P	Manajemen	RIAU	PEG NEGERI SIPIL	PEG SWASTA	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-May-2004	3.6	TEPAT WAKTU
7	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	25-Sep-2004	3.57	TEPAT WAKTU
8	P	Manajemen	D.I. YOGYAKARTA	PENSIUN	PEG SWASTA	SARJANA	PASCA SARJANA	IPS	1-Sep-2000	27-Sep-2005	3.25	TIDAK TEPAT WAKTU
9	L	Manajemen	JAWA TENGAH	PEG NEGERI SIPIL	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	3-Dec-2005	2.9	TIDAK TEPAT WAKTU
10	P	Manajemen	D.I. YOGYAKARTA	USAHA SENDIRI	USAHA SENDIRI	TAMAT SMA/SMK/SMEA/STM	TAMAT SMP/MTS	IPS	1-Sep-2000	26-Mar-2005	3.16	TIDAK TEPAT WAKTU
11	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	28-May-2005	3.16	TIDAK TEPAT WAKTU
12	L	Manajemen	D.I. YOGYAKARTA	USAHA SENDIRI	USAHA SENDIRI	SARJANA	DIPLOMA	IPA	1-Sep-2000	23-Sep-2006	2.96	TIDAK TEPAT WAKTU
13	L	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	SARJANA	IPS	4-Sep-2000	27-May-2004	3.38	TEPAT WAKTU
14	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	PEG NEGERI SIPIL	SARJANA	SARJANA	IPS	4-Sep-2000	27-Jul-2004	3.43	TEPAT WAKTU
15	P	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	PEG NEGERI SIPIL	SARJANA MUDA	SARJANA	IPS	1-Sep-2000	27-Jul-2005	2.85	TIDAK TEPAT WAKTU
16	P	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	WIRASWASTA	DIPLOMA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-Jul-2004	3.53	TEPAT WAKTU
17	L	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	PEG SWASTA	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	27-Jan-2005	3.1	TIDAK TEPAT WAKTU
18	P	Manajemen	JAWA TENGAH	PEG NEGERI SIPIL	TIDAK BEKERJA	SARJANA	SARJANA	IPS	4-Sep-2000	27-Jul-2004	3.23	TEPAT WAKTU
19	L	Manajemen	JAWA TENGAH	PEG NEGERI SIPIL	TIDAK BEKERJA	PASCA SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-Jul-2004	3.55	TEPAT WAKTU
20	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	PEG NEGERI SIPIL	DOKTOR	SARJANA	IPA	4-Sep-2000	27-Jul-2004	3.69	TEPAT WAKTU
21	L	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	PEG NEGERI SIPIL	SARJANA	SARJANA	IPA	1-Sep-2000	4-Dec-2004	3.08	TIDAK TEPAT WAKTU
22	P	Manajemen	JAWA TENGAH	USAHA SENDIRI	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-May-2004	3.05	TEPAT WAKTU
23	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	SARJANA	SARJANA	IPA	1-Sep-2000	26-Mar-2005	2.81	TIDAK TEPAT WAKTU
24	P	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	TIDAK BEKERJA	SARJANA MUDA	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	27-Jul-2004	3.69	TEPAT WAKTU
25	L	Manajemen	JAWA TENGAH	USAHA SENDIRI	USAHA SENDIRI	TAMAT SMA/SMK/SMEA/STM	TAMAT SMP/MTS	IPS	1-Sep-2000	26-May-2007	3.09	TIDAK TEPAT WAKTU
26	L	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	WIRASWASTA	DOKTOR	SARJANA MUDA	IPA	1-Sep-2000	25-Mar-2006	2.99	TIDAK TEPAT WAKTU
27	P	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	27-Jul-2004	3.31	TEPAT WAKTU
28	L	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	DIPLOMA	IPA	4-Sep-2000	25-Sep-2004	3.34	TEPAT WAKTU
29	L	Manajemen	JAWA TENGAH	USAHA SENDIRI	USAHA SENDIRI	TAMAT SMP/MTS	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	25-Sep-2004	3.25	TEPAT WAKTU
30	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	PASCA SARJANA	SARJANA	IPA	4-Sep-2000	27-May-2004	3.59	TEPAT WAKTU
31	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	PASCA SARJANA	TAMAT SMA/SMK/SMEA/STM	IPA	1-Sep-2000	4-Dec-2004	3.37	TIDAK TEPAT WAKTU
32	L	Manajemen	JAWA TIMUR	PENSIUN	USAHA SENDIRI	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	25-Sep-2004	3.52	TEPAT WAKTU
33	P	Manajemen	D.I. YOGYAKARTA	WIRASWASTA	WIRASWASTA	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	25-Sep-2004	3.57	TEPAT WAKTU
34	P	Manajemen	SUMATERA SELATAN	USAHA SENDIRI	TIDAK BEKERJA	DIPLOMA	DIPLOMA	IPS	4-Sep-2000	27-Jul-2004	3.27	TEPAT WAKTU
35	P	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	TIDAK BEKERJA	PASCA SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	4-Dec-2004	3.32	TIDAK TEPAT WAKTU
36	L	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	WIRASWASTA	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	25-Sep-2004	3.32	TEPAT WAKTU
37	L	Manajemen	SUMATERA SELATAN	PEG NEGERI SIPIL	WIRASWASTA	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	27-Sep-2005	2.74	TIDAK TEPAT WAKTU
38	P	Manajemen	RIAU	PEG NEGERI SIPIL	TIDAK BEKERJA	SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	25-Sep-2004	3.35	TEPAT WAKTU
39	P	Manajemen	D.I. YOGYAKARTA	PEG SWASTA	PEG SWASTA	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-Jul-2004	3.19	TEPAT WAKTU
40	P	Manajemen	SUMATERA BARAT	PEG NEGERI SIPIL	PEG NEGERI SIPIL	TAMAT SMA/SMK/SMEA/STM	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	27-Jul-2004	3.33	TEPAT WAKTU
41	L	Manajemen	D.I. YOGYAKARTA	PEG NEGERI SIPIL	PEG NEGERI SIPIL	SARJANA	PASCA SARJANA	IPA	4-Sep-2000	25-Sep-2004	3.31	TEPAT WAKTU
42	P	Manajemen	JAWA TENGAH	PENSIUN	TIDAK BEKERJA	TAMAT SMP/MTS	TAMAT SMA/SMK/SMEA/STM	IPA	4-Sep-2000	25-Sep-2004	3.31	TEPAT WAKTU
43	P	Manajemen	JAWA BARAT	PEG NEGERI SIPIL	TIDAK BEKERJA	PASCA SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	4-Sep-2000	27-May-2004	3.63	TEPAT WAKTU
44	L	Manajemen	KALIMANTAN TIMUR	PEG SWASTA	TIDAK BEKERJA	PASCA SARJANA	TAMAT SMA/SMK/SMEA/STM	IPS	1-Sep-2000	27-Sep-2005	2.88	TIDAK TEPAT WAKTU
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
31787	P	Farmasi	D.I. YOGYAKARTA	USAHA SENDIRI	TIDAK BEKERJA	TAMAT SMP/MTS	TAMAT SD	IPA	1-Sep-1999	27-Jan-2004	3.21	TIDAK TEPAT WAKTU
31788	P	Farmasi	D.I. YOGYAKARTA	WIRASWASTA	WIRASWASTA	TAMAT SD	TAMAT SD	LAIN-LAIN	1-Sep-1999	27-Jan-2004	3.33	TIDAK TEPAT WAKTU

Lampiran 2. *Syntax SVM dan Random Forest*

```
memory.limit()
memory.limit(size = 125000) #untuk menambah kapasitas R memory

install.packages("ROSE")
install.packages("e1071")
install.packages("caret")
install.packages("randomForest")
library(ROSE)
library(e1071)
library(randomForest)
library(caret)

#persiapan data
setwd("D:/kuliah/SKRIP SWEEET")
alumni=read.csv("D:/kuliah/SKRIP SWEEET/DATA/ANALISIS_FIX_FIX.csv")
alumni_fix<-alumni[,3:12]
table(alumni_fix$KET)

#Balancing data menggunakan metode under sampling
under <- ovun.sample(KET~., data = alumni_fix, method = "under", N = 17840,
seed = 12)$data
write.csv(under,"E:/Under_Sampling.csv")
table(under$KET)
str(under)

#training dan testing
nrow(under)
```

```
n=round(0.75*nrow(under));n
set.seed(12345);index= sample(seq_len(nrow(under)), size=n)
train=under[index,];table(train$KET)
test=under[-index,];table(test$KET)

#1. SVM
#tuning SVM Kernel RBF
tune.svmR10<-tune(svm,KET~., data =train,kernel="radial",
            ranges =list(cost=c(0.1,1,5,10,50),gamma=c(1,2,3,4)),tunecontrol =
tune.control(cross=10))
summary(tune.svmR10)

tune.svmR5<-tune(svm,KET~., data =train,kernel="radial",
             ranges =list(cost=c(0.1,1,5,10,50),gamma=c(1,2,3,4)),tunecontrol =
tune.control(cross=5))
summary(tune.svmR5)
#SVM RBF OPTIMAL DI K=,C=1,GAMMA=1
svmR=svm(KET~.,data=train, cost=1, gamma=1, kernel="radial",cross=5)
summary(svmR)
#prediksi dan akurasi data train
pred.train.RBF<- predict(svmR,train)
confusionMatrix(pred.train.RBF,train$KET)
#prediksi dan akurasi data test
pred.test.RBF<- predict(svmR,test)
confusionMatrix(pred.test.RBF,test$KET)

#Tune SVM kernel Sigmoid
tune.svmS5<-tune(svm,KET~., data =train,kernel="sigmoid",
```

```
        ranges =list(cost=c(0.1,1,5,10,50),gamma=c(1,2,3,4)),tunecontrol =
tune.control(cross=5))
summary(tune.svmS5)

tune.svmS10<-tune(svm,KET~., data =train,kernel="sigmoid",
        ranges =list(cost=c(0.1,1,5,10,50),gamma=c(1,2,3,4)),tunecontrol =
tune.control(cross=10))
summary(tune.svmS10)
#SVM SGMD OPTIMAL DI K=5,C=10,GAMMA=1
svmS=svm(KET~.,data=train, cost=10, gamma=1, kernel="sigmoid",cross=5)
summary(svmS)
#prediksi dan akurasi SVM data train
pred.train.SGMD<- predict(svmS,train)
confusionMatrix(pred.train.SGMD,train$KET)
#prediksi dan akurasi SVM data test
pred.test.SGMD<- predict(svmS,test)
confusionMatrix(pred.test.SGMD,test$KET)

#Importance Variable
IV <- t(svmR$coefs) %*% svmR$SV   # weight vectors
IV <- apply(IV, 2, function(v){sqrt(sum(v^2))})  # weight
IV <- sort(IV, decreasing = T)
View(IV)
IV

#2. RANDOM FOREST
tune.rf= tuneRF(x=train[,-10],y=train[,10],stepFactor=2) #optimal di m=2 (error
terkecil)
tune.rf
```

```
#masukin m optimum buat ncari ntree optimum (ntree optimum di 500)
alumni.rf1 <- randomForest(KET~., data =train, importance=T,ntree=25,mtry=2)
alumni.rf2 <- randomForest(KET~., data =train, importance=T,ntree=50,mtry=2)
alumni.rf3 <- randomForest(KET~., data =train, importance=T,ntree=100,mtry=2)
alumni.rf4 <- randomForest(KET~., data =train, importance=T,ntree=500,mtry=2)
alumni.rf5 <- randomForest(KET~., data =train,
importance=T,ntree=1000,mtry=2)
#uji coba prediksi training data, mencari ntree optimum (ntree optimum di 500)
pred.trainRF1<- predict(alumni.rf1,train)
pred.trainRF2<- predict(alumni.rf2,train)
pred.trainRF3<- predict(alumni.rf3,train)
pred.trainRF4<- predict(alumni.rf4,train)
pred.trainRF5<- predict(alumni.rf5,train)
accuracy.trainRF1 <-confusionMatrix(pred.trainRF1,train$KET)
accuracy.trainRF2 <-confusionMatrix(pred.trainRF2,train$KET)
accuracy.trainRF3 <-confusionMatrix(pred.trainRF3,train$KET)
accuracy.trainRF4 <-confusionMatrix(pred.trainRF4,train$KET)
accuracy.trainRF5 <-confusionMatrix(pred.trainRF5,train$KET)
accuracy.trainRF1
accuracy.trainRF2
accuracy.trainRF3
accuracy.trainRF4
accuracy.trainRF5

#ntree optimum di 500

#prediksi dan akurasi RF data train menggunakan ntree=500 (alumni.rf4)
pred.train.RF<- predict(alumni.rf4,train)
```

```
confusionMatrix(pred.train.RF,train$KET)
#prediksi dan akurasi RF data test menggunakan ntree=500 (alumni.rf4)
pred.test.RF<- predict(alumni.rf4,test)
confusionMatrix(pred.test.RF,test$KET)
#importance variable random forest
View(importance(alumni.rf4))
varImpPlot(alumni.rf5,pch=19, main = "Importance Variable of alumni.rf5")
```