BAB I

PENDAHULUAN

1.1. Latar Belakang

Pendidikan merupakan salah satu cara dalam meningkatkan kualitas Sumber daya manusia (SDM) dan aset penting bagi kemajuan bangsa. Melalui pendidikan, SDM yang berkualitas baik dari segi spiritual, intelegensi serta keterampilan dapat disiapkan. Adanya SDM yang berkualitas, sosok-sosok individu ini diharapkan dapat berperan dalam proses pembangunan bangsa dan Negara. Salah satu cara agar dapat membentuk manusia berkualitas ialah berdasarkan tingkat pendidikan yang semakin tinggi (Profil Anak Indonesia, 2019).

Pendidikan Tinggi adalah jenjang pendidikan setelah pendidikan menengah yang mencakup beberapa program pendidikan antara lain program diploma, program sarjana, program magister, program doktor, dan program profesi, serta program spesialis yang dikenal dengan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Selanjutnya, agar seseorang lulusan dari Perguruan Tinggi dikatakan baik apabila dapat lulus tepat waktu atau waktu lama studi tidak lebih dari 8 semester atau 4 tahun untuk jenjang S1 dan dapat memiliki nilai atau IPK yang baik pula. Selain itu, agar dapat memenuhi standar kompetensi lulusan bagi mahasiswa program sarjana (S1) beban wajib yang harus ditempuh adalah paling sedikit 144-160 satuan kredit semester (sks).

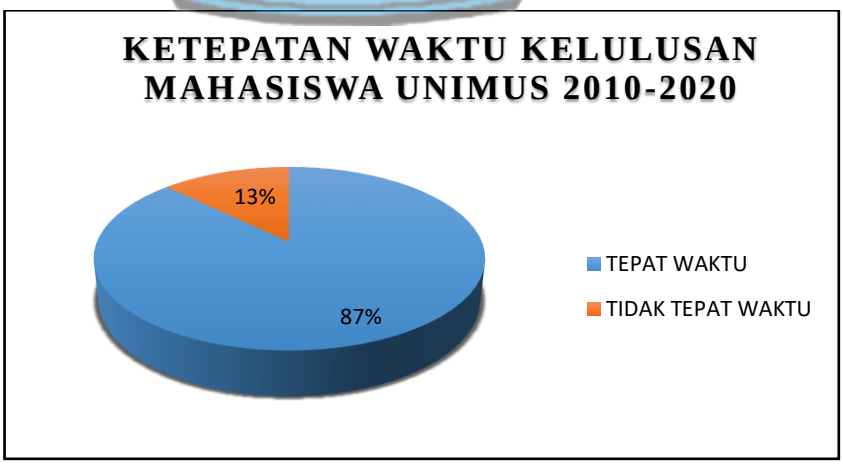
Dalam dunia pendidikan, setiap perguruan tinggi mempunyai kewajiban

untuk mengontrol prestasi belajar setiap mahasiswanya dan dapat menghasilkan lulusan yang berkualitas. Dimana seluruh perguruan tinggi baik dalam negeri maupun luar negeri juga dituntut untuk menjamin mutu lulusan, agar mutu ini dimaksudkan dapat bermanfaat bagi bangsa dan Negara dan juga lulusan perguruan tinggi dapat langsung dimanfaatkan oleh *stakeholders*. Selain itu kualitas lulusan dari perguruan tinggi juga dapat dipengaruhi oleh beberapa faktor, baik faktor *internal* maupun *eksternal*. Faktor *internal* adalah faktor yang berasal dari dalam diri mahasiswa sendiri, seperti kecerdasan, kemampuan belajar, latar belakang keluarga dan lain sebagainya. Faktor *eksternal* adalah faktor yang berasal dari luar diri mahasiswa seperti lingkungan belajar, lingkungan pergaulan, sarana dan prasarana yang dimiliki, serta lain sebagainya. Faktor-faktor *internal* dan *eksternal* tersebut diduga memiliki pengaruh terhadap lamanya seorang mahasiswa dalam menyelesaikan masa studi di jenjang pendidikan yang sedang ditempuh (Erene fajrila, 2018).

Universitas Muhammadiyah Semarang (UNIMUS) merupakan salah satu perguruan tinggi swasta yang berada di Semarang, Jawa Tengah, Indonesia. UNIMUS didirikan pada tanggal 4 agustus 1999 dengan program studi yang memperoleh ijin operasional pada awal pembukaan tahun 1999 sebanyak 14 program studi. Namun seiring berjalannya waktu UNIMUS telah tumbuh berkembang menjadi tempat pembelajaran yang terpilih. Hingga pada akhirnya UNIMUS membuka beberapa program studi baru, pada tahun 2020 ini Universitas Muhammadiyah Semarang memiliki total 8 fakultas, dengan 4 (empat) program Diploma tiga, 1 (satu) program Diploma empat, 2 (dua) program pascasarjana, 1

(satu) program studi profesi ners, 1 (satu) program studi profesi doktor gigi, 1 (satu) program studi profesi dokter dan 15 program sarjana. Setiap tahun Universitas Muhammadiyah Semarang mengadakan perayaan wisuda untuk mahasiswa yang dinyatakan lulus, tetapi juga masih banyak mahasiswa Universitas Muhammadiyah Semarang di 8 fakultas tersebut yang tidak lulus tepat waktu.

Oleh karena itu, Universitas Muhammadiyah Semarang memerlukan tindakan yang tepat untuk mengetahui faktor-faktor yang mempengaruhi lama studi mahasiswa tersebut apakah tepat waktu atau tidak tepat waktu. Ada banyak faktor yang mempengaruhi lama (masa) studi mahasiswa dalam menempuh suatu pendidikan. Faktor yang digunakan dalam penelitian ini memuat penelitian terdahulu oleh Erene Fajrila (2018) yang berjudul perbandingan klasifikasi ketepatan waktu kelulusan mahasiswa menggunakan regresi logistik biner dan naïve bayes classifier yang mengasumsikan bahwa faktor-faktor yang mempengaruhi lama studi mahasiswa antara lain jenis kelamin, asal daerah, jenis SMA, jurusan saat SMA, fakultas, IPK, serta Pekerjaan Orang Tua.



Gambar 1. 1 Ketepatan Waktu Kelulusan Mahasiswa Unimus Tahun 2010-2020

Berdasarkan gambar 1.1 terlihat bahwa klasifikasi data untuk ketepatan waktu kelulusan mahasiswa di UNIMUS ada sebanyak 4640 data atau 87% yang berada dikelas tepat waktu, sedangkan kelas tidak tepat waktu hanya 675 data atau sekitar 13%, sehingga dalam rangka untuk mengetahui lama (masa) studi mahasiswa perlu dilakukan klasifikasi berdasarkan faktor-faktor yang mempengaruhinya. Selain itu, data mahasiswa yang ada cukup besar hingga mencapai ratusan hingga ribuan untuk data kelulusan mahasiswa dalam satu tahunnya, maka penting untuk menggali informasi-informasi berharga dalam data tersebut. Kemudian, untuk membantu dalam menemukan informasi-informasi berharga tersebut diperlukan adanya teknik data mining. Data mining merupakan suatu proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat 4 yang tersimpan di dalam *database* besar (Turban et al, 2005). Salah satu teknik data mining adalah pengklasifikasian. Metode pengklasifikasian menjadi hal yang penting sebagai alat evaluasi dan penarikan kesimpulan dari permasalahan lama (masa) studi mahasiswa UNIMUS. Penentuan kululusan mahasiswa dalam menyelesaikan lama studi tepat waktu atau tidak tepat waktu ini merupakan salah satu contoh klasifikasi.

Teknik klasifikasi bertujuan untuk menemukan suatu fungsi keputusan yang secara akurat memprediksi kelas dari data *testing* yang berasal dari fungsi distribusi yang sama dengan data untuk *training*. Oleh karena itu terdapat dua kondisi himpunan kelas data, yaitu *balance* dan *imbalanced class* data. *Imbalanced class* terjadi ketika satu kelas melebihi jumlah kelas lainnya. Kelas data banyak disebut

kelas mayoritas (kelas negatif) sedangkan kelas data sedikit disebut kelas minoritas (kelas positif). Dalam kondisi seperti data ketepatan waktu kelulusan UNIMUS, sebagian besar *classifier* bias terhadap kelas mayor, karena mesin klasifikasi akan condong memprediksi ke kelas mayor dan mengabaikan kelas minor (Japkowicz dan Stephen, 2002).

Salah satu metode dalam klasifikasi adalah *Support Vector Machine* (SVM). SVM merupakan metode klasifikasi non parametrik yang tidak harus memenuhi asumsi dan distribusi tertentu (Sahitayakti & Fithriasari, 2015). Teknik SVM digunakan untuk menemukan fungsi pemisah (*classifier*) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda. Kelebihan utama SVM adalah dengan *support vectornya* sudah mewakili semua data yang ingin diklasifikasikan berbeda dengan metode lain yang mengharuskan semua data diinput untuk diklasifikasikan sehingga menghasilkan *performance* yang baik. Selain itu SVM juga baik dalam hal prediksi untuk klasifikasi (Sain & Purnami, 2015). Sementara itu kelebihan lainnya yaitu dalam menentukan jarak menggunakan *support vector* sehingga proses komputasi menjadi cepat (Octaviani, et al., 2014). Pada umumnya data dalam dunia nyata jarang yang bersifat *linier separable*, kebanyakan bersifat *non-liniear*. Untuk menyelesaikan masalah *non-liniear*, SVM dimodifikasi dengan memasukkan fungsi kernel (Johra, 2018). Metode ini sudah baik dalam melakukan klasifikasi ketika jumlah kelas dari variabel label dalam data *balanced* akan tetapi jika data yang digunakan *imbalanced*, akan berdampak pada sulitnya mendapatkan model prediksi yang baik dan bermakna karena adanya ketidakcukupan informasi dari kelas minor (Yap dkk,

2014). Metode klasifikasi ini akan bias terhadap kelas mayor dan memiliki kinerja rendah pada kelas minor (Batuwita & Palade, 2012).

Secara umum pendekatan berbasis sampling dibagi menjadi dua yaitu metode *oversampling* dan *undersampling*. Metode *undersampling* menyeimbangkan data dengan cara menghapus beberapa pengamatan pada kelas mayor hingga keseimbangan data *training* yang diinginkan tercapai. Pendekatan berbasis sampling yang kedua yaitu *oversampling*. Metode ini bekerja untuk menyeimbangkan data *training*, jika data yang digunakan untuk membuat model tidak seimbang maka akan meningkatkan kesalahan dalam klasifikasi kelas minor dengan cara meningkatkan jumlah data pada kelas minor. Oleh karena itu, salah satu alternatif paling efektif untuk meningkatkan akurasi model adalah melakukan *Synthetic Minority Oversampling Technique* (SMOTE) pada praproses (Barro, et al., 2013). Selain itu, alternatif lain untuk meningkatkan nilai akurasi kelas *imbalanced* adalah dengan menggunakan metode ensemble. Metode *ensemble* pada prinsipnya mengkombinasikan sekumpulan *classifier* yang dilatih dengan tujuan untuk membuat model klasifikasi (*classifier*) campuran yang terimprovisasi sehingga membuat *classifier ensemble* yang terbentuk lebih akurat dari pada classifier asalnya dalam melakukan suatu pengklasifikasian (Han dkk, 2012).

Boosting (Freud dan Schapire, 1997) dan *Bagging* (Breiman, 1996) merupakan metode *ensemble* yang paling popular digunakan. *Boosting* dan *Bagging* adalah salah satu metode *ensemble* yang berbasis variasi data ensemble, yang terdiri dari memanipulasi data *training* sedemikian rupa sehingga masing-masing *classifier* dilatih dengan data *training* yang berbeda. Metode *Bagging*

didasarkan pada gagasan membuat berbagai sampel dari data *training*. Untuk variasi dari data *training* akan dihasilkan model klasifikasi tertentu, kemudian hasilnya akan diberikan sebagai kombinasi atau gabungan model.

Pada prinsipnya *Boosting* membentuk satu *classifier* yang kuat dengan mengkombinasikan sekumpulan *classifier*.*Boosting* mempertahankan sekumpulan bobot pengamatan pada saat *training* pengamatan dan secara adaptif menyesuaikan (updating) bobot-bobot ini pada akhir tiap iterasi *boosting*. Bobot-bobot dari pengamatan yang salah terklasifikasikan pada saat training akan dinaikkan sementara bobot-bobot pengamatan yang terklasifikasikan dengan benar akan diturunkan nilainya, dengan kata lain *Boosting* memaksa suatu *classifier* untuk memberi perhatian yang lebih pada pengamatan yang salah diklasifikasikan (Li dkk, 2008). Namun, karena desainnya yang berorientasi pada akurasi, algoritma metode *ensemble* yang secara langsung diterapkan ke data yang *imbalanced* tidak bisa menyelesaikan masalah pengklasifikasi. Dengan mengkombinasikan *ensemble* dengan teknik lain untuk mengatasi masalah *imbalanced* data telah dilakukan dan menghasilkan nilai yang positif (Galar dkk, 2011).

Salah satu penelitian mengenai *boosting* dengan *base classifier* SVM dilakukan oleh Winalia Agwil (2015). Secara umum algoritma pada level data dan metode *ensemble* lebih fleksibel karena mereka dapat digunakan secara terpisah dari *base classifier*. Salah satu metode yang popular digunakan akhir-akhir ini yaitu *SMOTE-Boosting* dan *SMOTE-Bagging* yang mengkombinasikan algoritma pada level data yaitu menambahkan algoritma *SMOTE* ditiap iterasinya dengan metode *ensemble* (Freund dan Schapire, 1995).

Penelitian oleh Syauqi Amri (2018) membahas tentang klasifikasi ketepatan kelulusan lama studi mahasiswa yang merupakan hasil akhir pencapaian yang membanggakan dalam menempuh suatu pendidikan pada jenjang tertentu. Tujuan dari penelitian ini untuk mengetahui hasil klasifikasi yang tepat antara *Support Vectore Machine* 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%.

Yongqing dkk (2014) mengembangkan suatu metode untuk mengatasi masalah imbalanced data menggunakan *smote bagging*, hasil penelitiannya menunjukkan bahwa metode SMOTE dapat mempertahankan *specificity* tinggi dan meningkatkan *sensitivity*. Penggabungan metode *smote bagging* dapat menyeimbangkan kembali data yang tidak seimbang dan meningkatkan akurasi dengan teknik *rebalancing* daripada pengelompokkan *ensemble* biasa.

Chawla dkk (2003) menggunakan metode *smote boosting* untuk pengklasifikasian *imbalanced data* dengan rasio *imbalanced* sebesar 71% untuk mayoritas kelasnya dan 29% untuk minoritas kelasnya, hasil penelitiannya menunjukkan bahwa beberapa set data yang tidak seimbang menunjukkan bahwa 86% 14% Sekolah Putus Sekolah 5 algoritma *smote boosting* yang diusulkan dapat menghasilkan prediksi yang lebih baik dari kelas minoritas daripada *AdaBoost*, *AdaCost*. *SMOTEBoost* secara implisit meningkatkan bobot pada kelas minoritas yang salah diklasifikasikan (*false negative*) dalam distribusi jumlah kelas minoritas

akan meningkat menggunakan algoritma SMOTE. Oleh karena itu, dalam ukuran resample *boosting* peningkatan *SMOTEBoost* mampu membuat wilayah keputusan yang lebih luas untuk kelas minoritas dibandingkan dengan *boosting* standar. Chawla dkk menyimpulkan bahwa *SMOTEBoost* dapat membangun pengklasifikasian dan mengurangi bias pada pengklasifikasian. *SMOTEBoost* menggabungkan kekuatan SMOTE untuk meningkatkan nilai recall dan *boosting* untuk meningkatkan nilai *precision*. Secara keseluruhan didapatkan ukuran *performance F-value* yang lebih baik.

Begitu halnya dengan *SMOTE-Bagging* yang menambahkan algoritma SMOTE di tiap prosedur resampling-nya. Tujuan dari adanya SMOTE yaitu untuk menambah *probabilitas* terpilihnya sampel-sampel yang sulit diklasifikasikan yang berasal dari kelas minor ke dalam data training di tiap iterasi sehingga membuat *base classifier* lebih fokus pada pengamatan kelas minor. Hal ini tentunya akan meningkatkan ketepatan klasifikasi pada kelas minoritas. Kemudian SMOTE yang dikombinasikan dengan prosedur *Bagging* memberikan kinerja keseluruhan (G-Mean) mengalami peningkatan (Wang, 2009).

Setelah mempelajari dan memahami beberapa penelitian terdahulu yang berkaitan dengan metode dan objek yang digunakan pada penelitian ini, maka dapat diketahui perbedaan yang dimiliki dari penelitian ini dengan penelitian- penelitian sebelumnya yang terletak pada objek yang digunakan dengan membandingkan beberapa metode klasfifikasi. Objek yang digunakan pada penelitian ini adalah data kelulusan mahasiswa UNIMUS yang diklasifikasikan menggunakan 3 metode klasifikasi yaitu *SMOTE bagging* SVM dan *SMOTE boosting* SVM yang kemudian

hasil dari ketiga metode tersebut dibandingkan metode mana yang menghasilkan akurasi paling baik.

1.2. Rumusan Masalah

Berdasarkan latar belakang diatas maka penulis dapat memaparkan beberapa rumusan masalah yang akan diselesaikan pada penelitian, yaitu:

1. Bagaimana hasil klasifikasi metode *SMOTE Bagging SVM* dan *SMOTE boosting SVM* untuk pengklasifikasian *imbalanced* data kelulusan mahasiswa unimus tahun 2010-2020?
2. Bagaimana metode terbaik dari *SMOTE bagging SVM* dan *SMOTE boosting SVM* untuk pengklasifikasian *imbalanced* data kelulusan mahasiswa unimus tahun 2010-2020?

1.3. Tujuan Penelitian

Berdasarkan rumusan yang telah dipaparkan, maka tujuan dari penelitian adalah sebagai berikut:

1. Mengklasifikasikan metode *SMOTE Bagging* SVM dan *SMOTE Boosting* SVM untuk pengklasifikasian *imbalanced* data kelulusan mahasiswa unimus tahun 2010-2020.

2. Memperoleh metode terbaik dengan menggunakan metode *SMOTE Bagging* SVM dan *SMOTE Boosting* SVM untuk pengklasifikasian *imbalanced* data kelulusan mahasiswa unimus tahun 2010-2020.

1.4. Manfaat Penelitian

Manfaat yang dapatdiperoleh dari penelitian ini adalah:

1. Manfaat Teoritis

   Membantu perkembangan ilmu pengetahuan mengenai metode klasifikasi sehingga dapat digunakan sebagai bahan bacaan dan referensi bagi pembaca dalam melakukan analisis terutama pada metode *SMOTE Bagging* SVM dan *SMOTE boosting* SVM.

2. Manfaat Praktis

a. Bagi Peneliti

   Peneliti mampu menerapkan metode yang sesuai dalam materi yang telah dipelajari serta peneliti memiliki pengetahuan dan wawasan mengenai klasifikasi data kelulusan mahasiswa UNIMUS dengan metode *SMOTE Bagging* SVM dan *SMOTE boosting* SVM.

b. Bagi pihak terkait

   Memberikan pengetahuan bagi pihak terkait diantaranya universitas, masyarakat, mahasiswa UNIMUS yaitu dapat membantu memberikan informasi mengenai lama studi mahasiswa UNIMUS serta faktor-faktor yang mempengaruhinya dan dapat meningkatkan mutu dan kualitas pendidikan di UNIMUS.

1.5. Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah:

1. Metode *Support Vector Machine* digunakan sebagai *base classifier*.
2. Kernel yang digunakan dalam penelitian ini yaitu *linear, polynomial, radial basis function dan sigmoid*.
3. Melakukan klasifikasi pada *imbalanced* data dengan menggunakan metode ensemble *SMOTE bagging* SVM dan *SMOTE boosting* SVM.
4. Pada penelitian ini membagi data *testing* dan data *training* menggunakan 10-*fold cross validation*