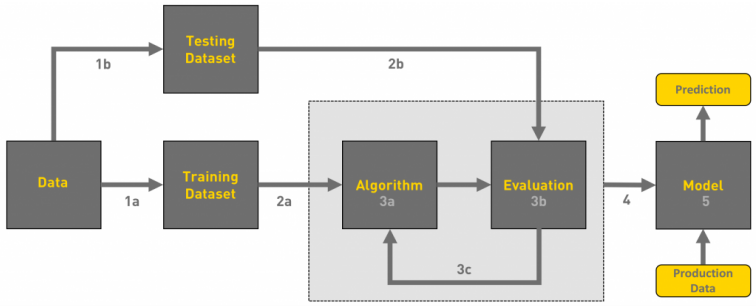
BAB I

PENDAHULUAN

1.1 Latar Belakang

Seiring dalam perkembangan zaman, teknologi pun juga mengalami kemajuan yang tidak dapat dihindari. Segala hal yang saat ini dinikmati dan membantu banyak orang di dunia tidak lain adalah hasil dari perkembangan teknologi yang selalu bergerak maju. Salah satu dari sekian banyak ilmu yang mempunyai peranan penting dalam kemajuan teknologi adalah ilmu matematika dan statistika. Ketika ilmu-ilmu ini dipakai dalam dunia nyata seringkali ilmu-ilmu ini berhubungan dengan data. Data pada kenyataannya banyak membantu kehidupan manusia dalam berbagai sektor seperti kesehatan, ekonomi, pendidikan, dan tentunya teknologi. Berbagai data yang terkumpul kemudian akan dipahami dan dianalisis sehingga menghasilkan sesuatu yang dapat bermanfaat dan praktikal. Seiring berjalannya waktu orang berlomba-lomba untuk mencitakan metode yang terbaik untuk dapat mengolah data menjadi semakin efisien dan akurat, hal ini lah yang membuat ada sangat banyak metode pengolahan data [1]. Contoh yang dapat dilihat adalah bagaimana mendeteksi orang yang depresi, mengatasi masalah penyebaran *Coronavirus Disease* (COVID-19), dan masih banyak lagi [2].

Proses pengumpulan, penambangan, dan penyeleksian data dari sebuah data yang besar ini lah kemudian disebut sebagai *data mining* yang nantinya akan dianalisis sehingga dapat menghasilkan kesimpulan dan informasi yang berguna. Metode yang digunakan untuk menganalisis data-data yang telah diperoleh dikenal sebagai *machine learning* yang kemudian tugasnya adalah membantu manusia dalam menemukan pola-pola dari data yang sangat kompleks sehingga bermanfaat. Pada Gambar 1.1 dijelaskan bagaimana proses kerja *machine learning* secara singkat. Pertama, data yang diperoleh dari sumber yang digunakan akan dibagi menjadi dua bagian dengan proporsi yang ditentukan oleh peneliti, yaitu data *training* dan data *testing*, lalu data-data yang berada di data *training* akan diolah sedemikian rupa sehingga menghasilkan model yang kemudian dapat memprediksi hasil dari sebuah masalah yang ada di lapangan [3].



Gambar 1.1 Proses Kerja *Machine Learning*

Sumber: *Workflow of a Machine Learning project* [4]

Pada umumnya *machine learning* memiliki dua pendekatan dasar yaitu *supervised learning* dan *unsupervised learning*. Metode *supervised learning* menggunakan data-data yang sudah memiliki label, data-data seperti ini bertujuan untuk melatih algoritma untuk kemudian dapat mengklasifikasi data dan memprediksi hasil secara akurat. Masalah dalam *supervised learning* umumnya dibagi menjadi dua yaitu masalah klasifikasi dan masalah regresi. Sebaliknya, metode *unsupervised learning* adalah bagian dari *machine learning* yang menganalisis data-data yang tidak memiliki label. Oleh karena itu, metode *unsupervised learning* umumnya digunakan untuk mengelompokkan data atau sering disebut sebagai *clustering* [5].

Tujuan akhir yang harus dapat diselesaikan oleh *machine learning* dan *data mining* adalah untuk membuat model yang baik dari sebuah *dataset*. Proses untuk membuat model ini biasa disebut sebagai *learning* atau *training* yang akan dicapai dengan menggunakan *learning algorithm*, kemudian *learning model* ini akan disebut sebagai *learners*. Ada beberapa algoritma yang populer dan banyak digunakan sebagai penelitian seperti *decision tree*, *regression*, *K-nearest neighbor*, *support vector machine*, *neural networks*, dan *linear discriminant analysis*. Metode *ensemble* melatih beberapa *learners* untuk memecahkan masalah yang sama, proses singkatnya adalah dengan membangun sebuah gabungan dari *learners* dan menyatukan semuanya, *learners* yang digabungkan menjadi satu ini biasa dikenal dengan istilah *base learners* atau *weak learners*. *Base learners* dapat dipilih dari beberapa metode populer yang tadi sudah disebutkan, kemudian metode *ensemble* akan meningkatkan performa *base learners* yang hanya sedikit lebih baik daripada tebakan acak menjadi sebuah *strong learners* yang kemudian dapat menghasilkan prediksi yang sangat akurat [5].

Netflix pernah membuat kompetisi untuk menyelesaikan masalah *supervised learning* dengan hadiah sebesar USD1.000.000 bagi yang dapat meningkatkan

performa *netflix classifier* sebanyak 10%. Setelah tiga bulan berjalan didapati ada beberapa tim di puncak klasemen yang berhasil meningkatkan performa sebanyak 5%, sampai pada akhirnya ada tim terbaik yang berhasil meningkatkan performa sebesar 8.5% dan hasil akhir menyatakan bahwa beberapa tim yang menduduki puncak klasemen memakai metode *ensemble* untuk meningkatkan performa klasifikasi pada tanggal 21 September 2009 [1].

Terdapat dua jenis masalah klasifikasi yaitu *binary classification* dan klasifikasi *multi-class classification*. Penelitian ini akan menggunakan masalah *multi-class classification* dalam membandingkan performa *ensemble* learning. Algoritma populer yang akan dipakai dalam penelitian ini adalah *logistic regression* sebagai *base model*. Kemudian akan dibandingkan metode *ensemble* yang terpilih yaitu *bootstrap aggregating* (*bagging*) dengan *boosting*. Data-data yang digunakan dalam penelitian ini memiliki variabel target dengan klasifikasi *multi-class* yang akan diperoleh dari situs *kaggle.com*

1.2 Rumusan Masalah

Berdasarkan latar belakang yang ada, akan dijawab masalah-masalah berikut.

1. Bagaimana perbandingan performa antara model *ensemble Bagging* dan *AdaBoost* untuk kasus klasifikasi *multi-class*?
2. Bagaimana perbandingan waktu pembentukan model antara model *ensemble Bagging* dan *AdaBoost* untuk kasus klasifikasi *multi-class*?
3. Bagaimana pengaruh pemilihan metrik evaluasi dalam membandingkan performa model untuk dataset dengan kelas yang tidak seimbang?

1.3 Tujuan Penulisan

Tujuan utama dari penelitian ini adalah sebagai berikut.

1. Mengetahui perbandingan performa antara model *ensemble Bagging* dan *AdaBoost* untuk kasus klasifikasi *multi-class*.
2. Mengetahui perbandingan waktu pembentukan model antara model *ensemble Bagging* dan *AdaBoost* untuk kasus klasifikasi *multi-class*.
3. Mengetahui pengaruh pemilihan metrik evaluasi dalam membandingkan performa model untuk dataset dengan kelas yang tidak seimbang.

1.4 Batasan Masalah

Batasan masalah dan asumsi yang digunakan dalam mencapai tujuan penelitian ini adalah sebagai berikut.

1. Data diambil dari situs *kaggle.com* dan diasumsikan valid.
2. Data yang diambil merupakan data dengan klasifikasi *multi-class*.
3. Pengolahan data menggunakan Python, Minitab dan SPSS.

1.5 Manfaat Penulisan

Manfaat dari penelitian ini dapat diklasifikasikan menjadi dua bagian, yaitu manfaat teoretis dan manfaat praktis.

1.5.1 Manfaat Teoretis

1. Dapat mengetahui bagaimana proses pengklasifikasian menggunakan *bagging* dan *boosting*.
2. Dapat mengetahui performa metode *ensemble* dalam mengklasifikasi data *multi-class*.

1.5.2 Manfaat Praktis

1. Membantu pembaca mengetahui praktik menggunakan metode *bagging* dan *boosting* dengan bantuan perangkat lunak python.
2. Menjadi bahan kajian bagi peneliti lain untuk pemakaian metode *bagging* dan *boosting* dalam mengklasifikasi data.

1.6 Struktur Penulisan

Penelitian ini akan ditulis berdasarkan struktur sebagai berikut:

1. pada Bab I akan dijelaskan latar belakang berisi penjelasan mengenai *machine learning* dan *data mining*, juga beberapa metode yang akan digunakan, lalu akan dijelaskan rumusan masalah, tujuan, batasan masalah, serta manfaat,
2. pada Bab II akan dipaparkan semua teori dan definisi yang akan digunakan seperti *logistic regression* sebagai *base model* yang akan dipakai, *maximum likelihood estimation*, metode *bagging* serta *boosting*, dan evaluasi model.

Teori-teori ini kemudian akan menjadi landasan berpikir dan perhitungan untuk mencapai tujuan dari Skripsi ini. Bab ini juga akan menjelaskan penelitian ilmiah yang relevan dan literatur yang terkait dengan Skripsi ini,

3. pada Bab III akan dijelaskan langkah-langkah yang harus dilakukan untuk melakukan proses pengerjaan dari data mentah, membersihkan data atau *preprocessing* data, sampai memodelkan menggunakan metode *logistic regression*, *bagging* dan *boosting*. Selain itu juga akan dibahas mengenai data-data yang akan dipakai dalam penelitian ini,
4. pada Bab IV akan dilakukan pengolahan data, pemodelan, analisis hasil, visualisasi, dan pembahasan terhadap langkah-langkah yang dijelaskan pada bab III. Bab ini akan menguraikan penjelasan proses penggunaan *bagging* dengan basis *logistic regression* dan *boosting* dengan basis *logistic regression* untuk menghasilkan model yang dapat mengklasifikasikan data-data *multi-class* serta tolak ukur dalam membandingkan performa kedua model yang dihasilkan pada bagian analisis hasil,
5. pada Bab V akan diberikan kesimpulan yang didapat dari hasil pengolahan data dan analisis yang telah dilakukan. Pada bab ini juga akan diberikan saran dalam melakukan analisis yang lebih efektif.