

## **BAB III**

### **METODOLOGI PENELITIAN**

#### **3.1 Objek Dan Subjek Penelitian**

Subjek pada penelitian ini adalah tentang meningkatnya penipuan pekerjaan dan agensi pekerjaan yang dilansir dari *Federal Trade Commission* yang memberikan data hasil laporan yang diterima. Dan yang menjadi objek pada penelitian ini adalah *dataset* tentang *fake job postings* yang diambil dari situs penyedia *dataset*, Kaggle.

#### **3.2 Alat Dan Bahan Penelitian**

Alat dan Bahan yang digunakan dalam penelitian ini adalah:

##### **3.2.1 Perangkat Keras**

Perangkat keras yang digunakan karena kebutuhan dalam membuat penelitian ini yaitu sebuah laptop dengan spesifikasi:

1. Prosesor Intel (R) Core (TM) i3-3110M
2. RAM 4 GB
3. Penyimpanan SSD 250 GB
4. Layar 14" HD
5. Intel(R) HD Graphics 4000

### **3.2.2 Perangkat Lunak**

Perangkat lunak yang digunakan dalam membuat penelitian ini adalah sebagai berikut:

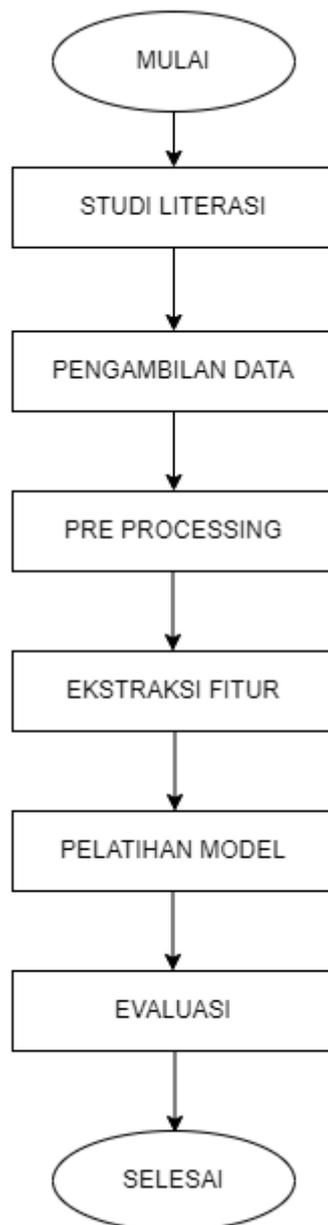
1. Sistem Operasi Windows 8.1
2. Google Chrome

### **3.2.3 Bahan**

Bahan untuk membuat penelitian ini ialah sebuah data set yang didapat dari *Kaggle* bernama *fake\_job\_postings* dan sudah dalam bentuk *Comma-Separated Values* atau CSV.

### 3.3 Diagram Alir Penelitian

Pada tahapan penelitian ini yaitu menguraikan seluruh kegiatan yang akan dilakukan oleh penulis selama penelitian berlangsung, dan digunakan sebagai pedoman dalam pelaksanaan penelitian agar hasil yang dicapai tidak menyimpang dari tujuan. Berikut adalah tahapan - tahapan yang dilalui:



### 3.3.1 Studi Literasi

Studi literatur ini dilakukan untuk mengetahui penelitian – penelitian yang terjadi pada bidang klasifikasi, pada jurnal yang ditulis oleh Indri Monika Parapat, Muhammad Tanzil Furqon, dan Sutrisno [5] tentang klasifikasi menggunakan *Support Vector Machine* sebagai algoritma klasifikasinya dengan hasil akhir dari penelitiannya menghasilkan rata-rata akurasi tertinggi sebesar 63,11%  $\lambda = 10$ ,  $C = 1$ ,  $itermax = 200$  dan juga menggunakan *kernel polynomial*. Kemudian pada jurnal yang ditulis oleh Tryse Rezza Biantong, Muhammad Tanzil Furqon, dan Arief Andy Soebroto [6] tentang implementasi metode *Support Vector Machine* untuk klasifikasi jenis penyakit malaria mendapatkan hasil dengan menggunakan *K-fold Cross Validation* dengan nilai  $K-fold=10$ . Hasil akurasi terbaik yang dihasilkan oleh sistem ini adalah 72% dengan nilai parameter  $\lambda=0.1$ ,  $\sigma=1$ ,  $\gamma=0.001$ ,  $C=0.1$ ,  $\epsilon=1.10^{-5}$ ,  $itermax = 50$  pada data rasio 80% data *training* : 20% data *testing*. Lalu Daisy Kurniawaty, Imam Cholissodin, dan Putra Pandu Adikara [7] tentang klasifikasi gangguan jiwa *Skizofrenia* menggunakan algoritma *Support Vector Machine* mendapatkan hasil pengujian akhir menghasilkan rata-rata akurasi optimal yaitu 100% dengan menggunakan rasio perbandingan 90%:10%, nilai  $\gamma = 0,00001$ ,  $\lambda = 3$ , nilai  $C = 0,01$ , *kernel polynomial of degree*, dan iterasi maksimal adalah 1000.

### 3.3.2 Pengambilan Data

Data yang dipakai dalam penelitian ini didapatkan dari sebuah *website open source* dan juga komunitas para *data scientists* dan *machine learning engineers* yang mana merupakan satu anak perusahaan dari Google yaitu Kaggle.

### **3.3.3 Pre Processing**

Lancar atau tidaknya proses klasifikasi di tentukan oleh seberapa rumit data yang dipakai, data yang digunakan itu sendiri belum tentu bisa langsung dipakai untuk proses klasifikasi di mana ada satu tahapan yang harus dilalui terlebih dahulu ialah tahap *pre processing*. Ditahap ini data yang sudah diperoleh harus diolah sedemikian rupa terlebih dahulu agar bisa dipakai. Berikut merupakan tahapan dari *pre processing* yang dipakai dalam penelitian ini:

#### **3.3.3.1 Cleanning**

Tahap pertama yang perlu dilakukan ketika akan preprocessing data adalah data cleaning atau membersihkan data. Artinya, data mentah yang telah diperoleh perlu diseleksi kembali. Kemudian, hapus atau hilangkan data-data yang tidak lengkap, tidak relevan, dan tidak akurat. Dengan melakukan tahap ini, Anda akan menghindari kesalahpahaman ketika menganalisis data tersebut.

#### **3.3.3.2 Splitting**

*Data Splitting* atau Pemisahan Data adalah di mana data dibagi menjadi 2 atau lebih subset. Umumnya, dengan 2 bagian terpisah, bagian pertama biasanya digunakan untuk mengevaluasi atau menguji data dan data yang lainnya dipakai untuk melatih model. Pemisahan Data adalah bagian penting dari *data science*, umumnya untuk membuat model dengan dasar sebuah data.

#### **3.3.3.3 One Hot Encoding**

*One hot encoding* adalah sebuah proses yang biasanya berada pada bagian *pre-processing* yang bertujuan untuk mengubah data *categorikal* integer menjadi *boolean* (*true false*) dimana setiap data kategori unik akan di-expand menjadi kolom atau parameter baru.

#### **3.3.3.4 Tokenizer**

*Tokenizer* adalah proses untuk membagi teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token - token/bagian-bagian tertentu.

#### **3.3.4 Ekstraksi Fitur**

Metode *Term Frequency Inverse Document Frequency* (TF-IDF) digunakan secara umum untuk menghubungkan kata-kata (term) dengan dokumen atau kalimat dengan memberikan bobot atau nilai pada setiap kata tersebut. Konsep utama dalam metode TF-IDF adalah menggabungkan frekuensi kata dalam sebuah dokumen (*Term Frequency*) dengan kebalikan frekuensi kata tersebut di seluruh dokumen (*Inverse Document Frequency*). Dalam perhitungan bobot menggunakan metode TF-IDF, terlebih dahulu dihitung nilai *Term Frequency* (TF) untuk setiap kata dengan memberikan bobot pada masing-masing kata tersebut [15]

#### **3.3.5 Pelatihan Model**

Latih model SVM pada data pelatihan dengan menggunakan *kernel* yang pada penelitian ini menggunakan *kernel Linear* dan *gamma* dengan nilai *auto*. Proses pelatihan ini bertujuan untuk menemukan *hyperplane* (garis/permukaan) yang memaksimalkan margin antara kelas dan mencapai pemisahan yang optimal.

### **3.3.6 Evaluasi**

Pada langkah evaluasi, metode yang digunakan adalah *Compute Area Under the Receiver Operating Characteristic Curve* atau ROC AUC. Evaluasi ini bertujuan untuk mengukur nilai akurasi (*accuracy*), presisi (*precision*), recall, dan skor F1 (F1-Score) dalam mengevaluasi performa model klasifikasi *Support Vector Machine* dalam mengklasifikasikan *fake job postings*. Dengan menganalisis hasil evaluasi ini kita dapat membuat kesimpulan yang lebih kuat dari penelitian ini.