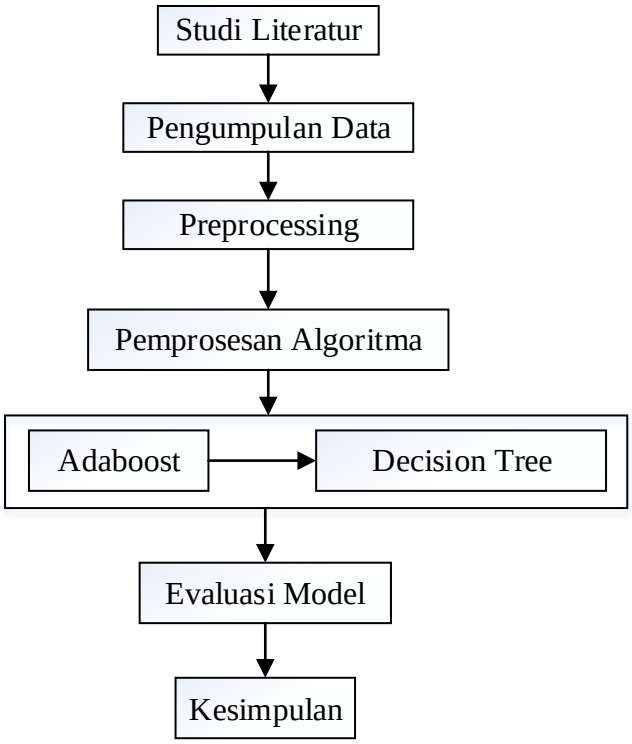
# BAB III
# METODOLOGI PENELITIAN

## 3.1. Metodologi Penelitian

Pada bab ini akan membahas langkah-langkah dari proses penelitian yang akan dikerjakan, dimulai dalam melakukan analisa dan mencari sebuah dataset melalui scraping data untuk memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan. Pembuatan alur penelitian merupakan sebagai tahapan dalam melakukan penelitian ini. Berikut adalah alur penelitian yang akan dilakukan dalam penelitian ini




Gambar 3. 1. Alur Penelitian

## 3.2. Studi Literatur

Tahapan yang pertamakali dilakukan yaitu studi lieratur atau studi awal yang merupakan tahap yang pertama kali dilakukan dalam penelitian. Pada tahap ini, peneliti melakukan observasi atau pemahaman penelitian yang meliputi tujuan dan persyaratan proyek dengan jelas dalam hal bisnis atau unit penelitian secara keseluruhan, menterjemahkan tujuan dan batasan ke dalam perumusan definisi masalah dalam melakukan prediksi dengan penerapan machine learning. Dalam studi literatur ini juga dilakukan pencarian terhadap sumber-sumber teori yang relevan dengan topik penelitian dari berbagai sumber seperti buku, penelitian terkait yang ada dalam suatu jurnal, internet, dan lain sebagainya yang mendukung proses penelitian dengan tujuan untuk memperkuat permasalahan.

## 3.3. Pengumpulan data

Pada tahap awal yaitu pengumpulan data dilakukan dengan menggunakan data public. Data yang digunakan pada data publik tersebut berasal dari website kaggle.com tentang *instagram fake spammer genuine accounts.* berikut linknya https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts.

## 3.4. Preprocessing

Tahap ini dilakukan untuk mengolah data mentah agar data terbebas dari kesalahan. Banyaknya data ulasan yang didapat banyak yang menggunakan kata yang tidak berstruktur. Tujuan dari preprocessing data adalah untuk memastikan data yang akan digunakan pada model machine learning sudah berkulitas dan tidak ada lagi data noise untuk proses klasifikasi. Preprocessing berguna untuk mengekstrak informasi dari ulasan, mengubah kata-kata yang tidak terstruktur itu menjadi bentuk standar. Dalam tahapan preprocessing juga akan disiapkan preprocessing data agar nantinya dapat digunakan dengan Data Transformation dan Split Validation dalam pembagian datanya. Karena kualitas data dapat mempengaruhi akurasi itu sendiri. Dalam data transformation dilakukan tahap

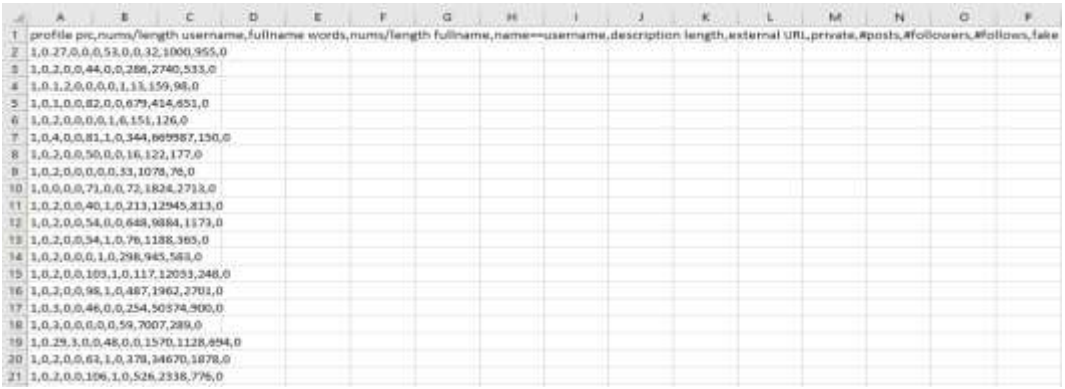
pre-processing data yaitu normalisasi data dengan cara MinMax Normalization yang bertujuan untuk membuat beberapa variabel memiliki rentang nilai yang sama, tidak terlalu besar atau terlalu kecil, dengan begitu dapat mempermudah analisis statistik. Sedangkan dalam split validation akan dibagi menjadi data training dan data testing menggunakan perbandingan yang diinginkan. Dalam penelitian ini akan digunakan perbandingan data pada split validation sebesar 80:20. Penggunaan perbandingan yang berbeda juga dapat mempengaruhi hasil dari akurasi yang didapatkan nantinya. Berikut adalah langkah-langkah dalam dalam preprocessing data.

a. Seleksi Dataset

Pada proses seleksi dataset adalah untuk memilih data keseluruhan dari dataset yang digunakan. Dataset yang didapatkan berasal dari dataset public yaitu Kaggle.com tentang fake account instagram. Dari hasil yang di dapatkan akan digunakan data sebanyak 1.152 data fake account pada Instagram diamana setiap satu datanya adalah menandakan satu informasi akun Instagram yang meliputi seperti profile picture kemudian apakah akun di privasi, jumlah followers dan lainya untuk perhitungan data mining.

b. Integrasi Dataset

Pada proses integrase dataset adalah pemrosesan dari data mentah dalam dataset yang akan diolah kedalam data yang siap untuk dilakukan preprocessing. Pada data awal yang merupakan data mentah adalah sebagai berikut.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | profile pic,nums/length username,fullname words,nums/length fullname,name==username,description length,external URL,private,#posts,#followers,#follows,fake | | | | | | | | | | | | | | | |
| 2 | 1,0.27,0,0,0,53,0,0,32,1000,955,0 | | | | | | | | | | | | | | | |
| 3 | 1,0,2,0,0,44,0,0,286,2740,533,0 | | | | | | | | | | | | | | | |
| 4 | 1,0.1,2,0,0,0,0,1,13,159,98,0 | | | | | | | | | | | | | | | |
| 5 | 1,0,1,0,0,82,0,0,679,414,651,0 | | | | | | | | | | | | | | | |
| 6 | 1,0,2,0,0,0,0,1,6,151,126,0 | | | | | | | | | | | | | | | |
| 7 | 1,0,4,0,0,81,1,0,344,669587,150,0 | | | | | | | | | | | | | | | |
| 8 | 1,0,2,0,0,50,0,0,16,122,177,0 | | | | | | | | | | | | | | | |
| 9 | 1,0,2,0,0,0,0,0,33,1078,76,0 | | | | | | | | | | | | | | | |
| 10 | 1,0,0,0,0,71,0,0,72,1824,2713,0 | | | | | | | | | | | | | | | |
| 11 | 1,0,2,0,0,40,1,0,213,12945,813,0 | | | | | | | | | | | | | | | |
| 12 | 1,0,2,0,0,54,0,0,648,9884,1173,0 | | | | | | | | | | | | | | | |
| 13 | 1,0,2,0,0,54,1,0,76,1188,365,0 | | | | | | | | | | | | | | | |
| 14 | 1,0,2,0,0,0,1,0,298,945,583,0 | | | | | | | | | | | | | | | |
| 15 | 1,0,2,0,0,103,1,0,117,12053,248,0 | | | | | | | | | | | | | | | |
| 16 | 1,0,2,0,0,98,1,0,487,1962,2701,0 | | | | | | | | | | | | | | | |
| 17 | 1,0,3,0,0,46,0,0,254,50374,900,0 | | | | | | | | | | | | | | | |
| 18 | 1,0,3,0,0,0,0,0,59,7007,289,0 | | | | | | | | | | | | | | | |
| 19 | 1,0.29,3,0,0,48,0,0,1570,1128,694,0 | | | | | | | | | | | | | | | |
| 20 | 1,0,2,0,0,63,1,0,378,34670,1078,0 | | | | | | | | | | | | | | | |
| 21 | 1,0,2,0,0,106,1,0,526,2338,776,0 | | | | | | | | | | | | | | | |

Gambar 3. 2. Dataset sebelum terintegrasi

Data diatas adalah data sebelum dilakukan integrasi yang kemudian akan dilakukan integrasi seperti pada gambar 3.3 dimana data telah di integrasi dilakukan pemisahan data terhadap setiap atribut yang dipisahkan oleh koma dengan cara pemisahan text ke kolom dimana yang data awalnya hanya dipisahkan oleh koma kemudian data dapat di definisikan untuk dilakukan preprocessing data pada dataset. Berikut adalah dataset yang telah dilakukan integrase.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | profile pic | nums/l | fullname | nums/lengtl | name==us | descriptio | external URL | private | #posts | #followers | #follows | fake |
| 2 | 1 | 0.27 | 0 | 0 | 0 | 53 | 0 | 0 | 32 | 1000 | 955 | 0 |
| 3 | 1 | 0 | 2 | 0 | 0 | 44 | 0 | 0 | 286 | 2740 | 533 | 0 |
| 4 | 1 | 0.1 | 2 | 0 | 0 | 0 | 0 | 1 | 13 | 159 | 98 | 0 |
| 5 | 1 | 0 | 1 | 0 | 0 | 82 | 0 | 0 | 679 | 414 | 651 | 0 |
| 6 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 6 | 151 | 126 | 0 |
| 7 | 1 | 0 | 4 | 0 | 0 | 81 | 1 | 0 | 344 | 669987 | 150 | 0 |
| 8 | 1 | 0 | 2 | 0 | 0 | 50 | 0 | 0 | 16 | 122 | 177 | 0 |
| 9 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 33 | 1078 | 76 | 0 |
| 10 | 1 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 72 | 1824 | 2713 | 0 |
| 11 | 1 | 0 | 2 | 0 | 0 | 40 | 1 | 0 | 213 | 12945 | 813 | 0 |
| 12 | 1 | 0 | 2 | 0 | 0 | 54 | 0 | 0 | 648 | 9884 | 1173 | 0 |
| 13 | 1 | 0 | 2 | 0 | 0 | 54 | 1 | 0 | 76 | 1188 | 365 | 0 |
| 14 | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 298 | 945 | 583 | 0 |
| 15 | 1 | 0 | 2 | 0 | 0 | 103 | 1 | 0 | 117 | 12033 | 248 | 0 |
| 16 | 1 | 0 | 2 | 0 | 0 | 98 | 1 | 0 | 487 | 1962 | 2701 | 0 |
| 17 | 1 | 0 | 3 | 0 | 0 | 46 | 0 | 0 | 254 | 50374 | 900 | 0 |
| 18 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 59 | 7007 | 289 | 0 |
| 19 | 1 | 0.29 | 3 | 0 | 0 | 48 | 0 | 0 | 1570 | 1128 | 694 | 0 |
| 20 | 1 | 0 | 2 | 0 | 0 | 63 | 1 | 0 | 378 | 34670 | 1878 | 0 |
| 21 | 1 | 0 | 2 | 0 | 0 | 106 | 1 | 0 | 526 | 2338 | 776 | 0 |
| 22 | 1 | 0 | 2 | 0 | 0 | 40 | 0 | 0 | 228 | 3516 | 999 | 0 |
| 23 | 1 | 0 | 1 | 0 | 0 | 35 | 1 | 1 | 35 | 1809 | 416 | 0 |

Gambar 3. 3 Dataset yang telah di integrase

Pada integrasi data juga dilakukan pengecekan pada tipe data pada setiap atribut diamana pada tahap ini dilakukan agar memepermudah dalam memeasukan data pada tool pemrosesan dengan mengecek tipe data apakah data integer, real, binominal, polynominal dan lainya.

c. Missing Data

Pada tahapan Missing data dilakukan agar data atau informasi yang hilang dalam suatu data dapat diketahui dengan baik. Dari yang hilang ataupun tidak tersedia mengenai subyek penelitian dalam atribut ataupun variable data tersebut. Dalam penelitian ini akan dilakukan pengecekan data hilang pada dataset. Pengecekan

dilakukan dengan pengecekan pada kolom dataset mana yang terdapat data yang hilang agar pada proses mining tidak ada data yang missing. Berikut adalah hasil pengecekan data missing pada dataset fake account pada Instagram.

```
profile pic             0
nums/length username    0
fullname words          0
nums/length fullname    0
name==username          0
description length      0
external URL            0
private                 0
#posts                  0
#followers              0
#follows                0
fake                    0
dtype: int64
```

Gambar 3. 4. Pengecekan Missing pada Dataset

Pada pengecekan diatas dapat diketahui bahwa tidak ada data yang missing atau hilang dalam dataset sehingga data siap dialakukan pemrosesan pada tahap selanjutnya.

d. Transformasi Data

Pada tahapan transformasi data dilakukan untuk mengubah skala pengukuran data asli ke dalam bentuk lain untuk tujuan Analisa data. Dalam penelitian ini akan dilakukan pengelompokan atribut berdasarkan dataset sehingga perubahan data yang dilakukan siap diolah dalam pengolahan data. Pengolahaan data dalam penelitian ini menggunakan bantuan tools rapidminer. Atribut pada dataset berjumlah 11 atribut dimana terdapat 1 atribut yang dijadikan label yaitu atribut fake sebagai acuan apakah data tersebut menyatakan positif fake akun atau negative fake akun. Dalam kriteria atribut ini juga menggunakan range nilai dalam menentukan value untuk pembagian jumlah kasusnya, sebagai contoh pada atribut followers yang menyatakan jumlah followers dimana followers pada dataset yang jumlahnya tertinggi adalah 15 juta followers pada deskripsi acuan mengambil nilai range antara >98 dan <98 followers. Pengambilan range ini berdasarkan pembelajaran learning pada pohon keputusan dimana tree teratas dalam pohon

keputusan mendapatkan pembagian range <98 dan >98 pada atribut followers dan sebagai penentuanya jika range <98 maka dinyatakan sebagai fake account. Berikut adalah penjelasannya dalam table 3.1.

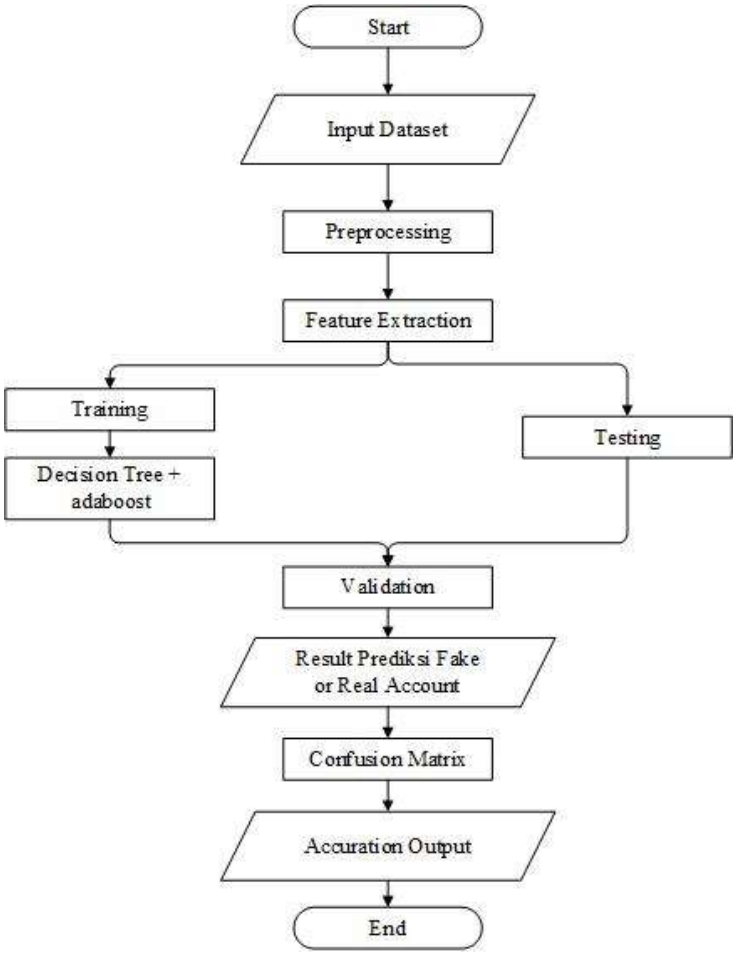
Tabel 3. 1 Deskripsi Atribut pada Dataset

| **Nama Atribut** | **Deskripsi Atribut** | **Kriteria/Keterangan** |
|---|---|---|
| Profile Pic | Ada atau tidaknya Foto Profil pada setiap akun | - 0<br>- 1 |
| nums/length username | Panjang nama pengguna | - >0.47<br>- <0.47 |
| fullname words | Nama lengkap pada deskripsi akun | - >1<br>- <1 |
| nums/length fullname | Panjang nama lengkap | - >0.36<br>- <0.36 |
| name username | Nama pengguna ( ada atau tidak) | - 1<br>- 2 |
| description length | Panjang deskripsi | - >65<br>- <65 |
| external URL | Url eksternal didalam deskripsi (ada atau tidak) | - 1<br>- 0 |
| private | Akun di privasi atau tidak | - 1<br>- 0 |
| Posts | Post pada media sosial | - >50<br>- <50 |
| Followers | Pengikut yang mengikut akun tersebut | - >98<br>- <98 |
| Follows | Akun yang di ikuti | - >342<br>- <342 |
| Fake | Status akun palsu atau tidak | - 1<br>- 0 |

Dalam pemrosesan preprocessing ini tidak terdapat data missing dan data hilang sehingga preprocessing hanya pada pengelompokan atribut dan data transformasi untuk persiapan pemrosesan dengan algoritma yang akan digunakan.

### 3.5. Pemrosesan Algoritma

Pada tahapan pemrosesan algoritma setelah dilakukan proprocessing pada dataset yang telah dilakukan dan juga pemodelan data sehingga nantinya data dapat dilakukan pemrosesan algoritma dengan menggunakan teknik bosting pada Decision Tree. *Model training* diawali dengan melatih data sebelum dilakukan bosting data oleh adaboost, kemudian akan dilakukan inisiasi kernel untuk pemrosesan Decision Tree sehingga menghasilkan model dari data latih yang kemudian model tersebut di uji berdasarkan data uji yang disiapkan kemudian akan dilakukan cross validation untuk menguji nilai akurasi menggunakan Confussion Matrix. Sehingga nantinya dapat dikategorikan fake account dan real account berdsarkan nilai 0 dan 1 berdasarkan dataset yang telah disediakan. Berikut adalah alur pemrosesan algoritma.




Gambar 3. 5. Alur Pemrosesan Algoritma

Proses diatas merupakan alur dari pemrosesan pada algoritma. Pada perhitungan algoritma menggunakan decision tree terdapat perhitungan manual agar nantinya hasil dapat tersinkronisasi dengan perhitungan menggunakan bantuan tool rapidminer. Berikut adalah perhingan manual dengan pada salah satu atribut.

Tabel 3. 2. Perhitungan Algoritma Decision Tree pada atribut Profile pic

| No | Attribut | Keterangan | Jumlah Kasus | Fake (Yes) | Fake (No) | Entrophy | Info Gain | Split Info | Gain Ratio |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Total | 1152 | 580 | 572 | 0,999965 | | | |
| | | | | | | | | | |
| 2 | profile pic | | | | | | | | |
| | | 1 | 808 | 240 | 568 | 0,87763 | 0,35711 | 0,879586 | 0,406 |
| | | 0 | 344 | 340 | 4 | 0,091402 | | | |

Pada pencarian perhitungan entropy total didapatkan dengan perhitungan sebagai berikut.

$$= -\left(\left(\frac{Fake\ (yes)}{Jumlah\ kasus}\right) X\ log_2\left(\frac{Fake\ (yes)}{Jumlah\ kasus}\right)\right)$$

$$-\left(\left(\frac{Fake\ (no)}{Jumlah\ kasus}\right) X\ log_2\left(\frac{Fake\ (no)}{Jumlah\ kasus}\right)\right)$$

$$= \text{Nilai entropy total}$$

Kemudian dalam mencari perhitungan entropy pada profile pic yaitu sama dengan perhitunfan mencari entropy total namun menggunakan jumlah data pada atributnya. Setelah into untuk mencari nilai info gainya adalah sebagai berikut.

$$= (\text{Nilai entropy total}) - \left(\left(\frac{Jumlah\ kasus\ atribut}{jumlah\ kasus\ total}\right) X\ nilai\ entropy\ atribut\right)$$

$$= \text{nilai info gain}$$

Setelah perhitungan nilai info gain adalah perhitungan pada nilai split info. Berikut adalah perhitungan pada split info dalam atribut profil pic

$$\left(-\left(\frac{Jumlah\ kasus\ atribut}{jumlah\ kasus\ total}\right) X\ log_2\left(\frac{jumlah\ kasus\ atribut}{jumlah\ kasus\ total}\right)\right) = nilai\ split\ info$$

Setelah didapt nilai split info maka selanjutnya adalah menghitung nilai gain ratio dengan membagi nilai seperti berikut.

$$= \frac{info\ gain}{split\ info} = nilai\ Gain\ Ratio$$

Maka ditemukan gain ratio pada atribut tersebut dan begitu seterusnya pada setiap atribut dalam perhitungan pada pemrosesan algoritma dengan pohon keputusan.

### 3.6. Evaluasi Model

Setelah model selesai dibuat, model digunakan untuk prediksi fake account instagram. Dimana pada tahapan modeling akan dilakukan pembelajaran learning untuk memprediksi akun Instagram asli atau palsu berdasarkan pada data latih sesuai dengan salah satu algoritma machine learning yaitu decision tree yang memprediksi akun palsu atau asli dengan penerapan pohon keputusan dalam memprediksi akun asli dan palsu. Kemudian akan ditemukan parameter yang digunakan untuk menentukan akun palsu dan asli berdasarkan pohon keputusan dari decision tree. Setelah dilakukan pemodelan dengan menggunakan metode machine learning dengan menggunakan decision tree, maka akan dilakukan validasi hasil dengan memvalidasi data testing dan training yang telah dilatih sebelumnya. Pada validasi juga digunakan confusion matrix dan juga Cross Validation untuk mengukur performa dari pemrosesan algoritmanya.

***a. Confusion Matrix***

Dalam penelitian ini, digunakan teori Confusion Matrix sebagai salah satu alat evaluasi kinerja model yang telah dikembangkan. Confusion Matrix adalah sebuah tabel yang digunakan dalam pemrosesan statistik dan machine learning untuk mengukur performa sebuah model klasifikasi. Confusion Matrix biasanya digunakan dalam konteks pengukuran performa model dalam klasifikasi biner, yang merupakan tugas klasifikasi yang

mengelompokkan data ke dalam dua kategori.

Ada empat elemen utama dalam Confusion Matrix:

1. True Positives (TP): Ini mengacu pada jumlah data yang benar-benar diklasifikasikan dengan benar sebagai positif oleh model. Dalam konteks penelitian ini, ini akan merujuk pada data yang benar-benar sesuai dengan kriteria yang ditetapkan.
2. True Negatives (TN): Ini mengacu pada jumlah data yang benar-benar diklasifikasikan dengan benar sebagai negatif oleh model. Dalam penelitian ini, ini akan merujuk pada data yang benar-benar tidak sesuai dengan kriteria yang ditetapkan.
3. False Positives (FP): Ini mengacu pada jumlah data yang salah diklasifikasikan sebagai positif oleh model, padahal seharusnya negatif. Ini juga dikenal sebagai kesalahan Type I.
4. False Negatives (FN): Ini mengacu pada jumlah data yang salah diklasifikasikan sebagai negatif oleh model, padahal seharusnya positif. Ini juga dikenal sebagai kesalahan Type II.

Dengan menggunakan nilai-nilai di atas, berbagai metrik evaluasi dapat dihitung, termasuk:

1. Accuracy: Mengukur sejauh mana model dapat mengklasifikasikan data dengan benar, dinyatakan sebagai (TP + TN) / (TP + TN + FP + FN).
2. Precision: Mengukur sejauh mana data yang diklasifikasikan sebagai positif oleh model benar-benar positif, dinyatakan sebagai TP / (TP + FP).
3. Recall (Sensitivity atau True Positive Rate): Mengukur sejauh mana model dapat mendeteksi semua data yang seharusnya positif, dinyatakan sebagai TP / (TP + FN).

Berikut adalah rumush perhitungan pada Confusion Matrix :

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Keterangan : TP = True Positive

FP = False Positive

TN = True Negative

***b. Cross Validation***

Cross Validation (Validasi Silang) adalah teknik yang digunakan dalam pemodelan statistik dan machine learning untuk mengukur kinerja dan keandalan model prediktif. Tujuan utama dari Cross Validation adalah untuk menghindari overfitting, yaitu ketika model terlalu sesuai dengan data pelatihan tetapi kurang generalisasi ke data yang tidak pernah dilihat sebelumnya. Teknik ini juga membantu dalam menilai sejauh mana model akan berhasil dalam pengujian dunia nyata. Pada validasi akan dilakukan kombinasi menggunakan adaboost, dimana adaboost digunakan sebagai boosting dalam meningkatkan performa dari akurasi prediksi menggunakan decision tree. Pada hasil validasi juga akan di bandingkan nilai performa jika menggunakan decision tree C4.5 saja dan menggunakan keduanya yaitu Decision Tree C4.5 dengan Adaboost. Pada k-fold cross validation akan menggunakan 10 fold validation dengan menggunakan bantuan tool rapid miner

## 3.7. Alat dan Bahan

Penelitian ini menggunakan perangkat keras Laptop Lenovo dengan Processor Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz 3.20 GHz, RAM 16,00 GB,

sedangkan perangkat lunak yang digunakan Microsoft Excel dan RapidMiner Studio untuk pengolahan data dan bahan yang digunakan diambil dari (https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts).