BAB II
LANDASAN TEORI

2.1. Penelitian Terkait

Dalam melakukan penyusunanan penelitian ini, penulis mengumpulkan penelitian yang berkaitan dengan latar belakang penelitian dari penelitian-penelitian terdahulu sebagai referensi dan acuan dalam membuat penelitian menggunakan penerapan *adaboost* dengan *Decision Tree C4.5* dalam prediksi *fake account instagram.* Beberapa penelitian terkait yang menjadi *key paper* pada penelitian ini. Berikut adalah penelitian terkait yang menjadi dasar dalam penelitian ini.

Literature rivew yang digunakan khusus tentang fake account dari penelitian 5 tahun terakhir yaitu tahun 2019 sampai dengan 2022.

Tabel 2. 1. Penelitian Terkait

No	Penelitian	Media	Dataset	Algoritma	Hasil
1	Penerapan Machine Learning dalam mendeteksi Fake Account pada Instagram (2022) [1]	Instagram	Dataset publik dari kaggle.com dengan jumlah data sebanyak 696 data. *https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts*	Support Vector Machine, Naïve Bayes, Random Forest, Adaptive Boosting	Hasil akurasi yang didapatkan 92.5%, Random Forest sebesar 91.7%, Support Vector Machine sebesar 90.7% dan Naïve Bayes sebesar 83.6%
2	Instagram Fake Account Detection Based on Machine Learning (2022) [3]	Instagram	Dataset dalam penelitian ini adalah scraping pribadi dari penelitian berupa akun real dan akun palsu. Tidak	Logistic regression, naïve bayes, random forest, SVM	Logistic regression mendapat akurasi sebesar 90.81 %, naïve bayes dengan 94.58 %, random forest dengan 95,2 %, Support Vector

No	Penelitian	Media	Dataset	Algoritma	Hasil
			disebutkan secara rinci berapa data yang berhasil di scraping dalam jurnal tersebut		Machine dengan 80,43 % dan Neural Network dengan 94,54%
3	Deteksi Twitter Bot Menggunakan Klasifikasi Decision Tree (2020) [5]	Twitter	Dataset yang diguanakan dalam penelitian adalah dataset public. Tidak dijelaskan secara rinci sumber berasal dari mana.	Decision Tree	Hasil pengukuran menunjukkan performa accuracy model Decision Tree mencapai 88.84% dan perhitungan kurva AUC dengan nilai 0.965.
4	Fake Account Detection using Machine Learning (2022) [7]	Instagram	Dataset dalam penelitian ini dilakukan degan teknik crawling data dari twitter dataset terdiri terdiri dari berbagai atribut seperti nama, jumlah status, jumlah teman, dan jumlah pengikut.	Logistic Regression, SVM, KNN, Random forest	Hasil akurasi yang di dapat sebesar 98%.
5	Analisa Perbandingan Klasifier Decision Tree, Random Forest, Dan Adaboost Dalam Mendeteksi Serangan (2020) [8]	Analisis Pemilihan algoritma	Dalam penelitian ini adalah menganalisa manakah algoritma yang paling efesien dalam hal waktu dan performa yang digunakan.	Decision Tree, Random Forest, dan AdaBoost	Hasil yang didapatkan adalah Decision Tree menjadi klasifier yang paling efisien Dengan hasil Precision 96,41%, Recall 100%, dan Accuracy 97,05%

Dalam beberapa dataset tidak disebutkan berapa banyak data dalam dataset yang digunakan dalam penelitian tertentu dan ulasan review penelitian terdahulu yaitu pada penelitian [5] dapat dilihat bahwa algoritma *Decision Tree C4.5* merupakan algoritma yang baik dalam memprediksi maupun deteksi untuk fake akun pada media social. Kemudian dalam penelitian [1] yang juga disebutkan bahwa Algoritma adaboost mempunyai kelebihan yaitu dapat meningkatkan dan mengoptimasi agar dapat digabung dengan algoritma lain sebagai algoritma estimator sehingga menghasilkan akurasi yang terbaik dengan memiliki margin error yang kecil. Dan dalam penelitian yang lain yaitu penelitian [6] juga disebutkan adaboost dapat digunakan untuk meingkatkan performa dengan menggabungkan pada algoritma *Decision Tree C4.5* dan menghasilkan kinerja lebih baik untuk prediksi klasifikasi dalam meningkatkan performa menjadi lebih baik.

2.2. Landasan Teori

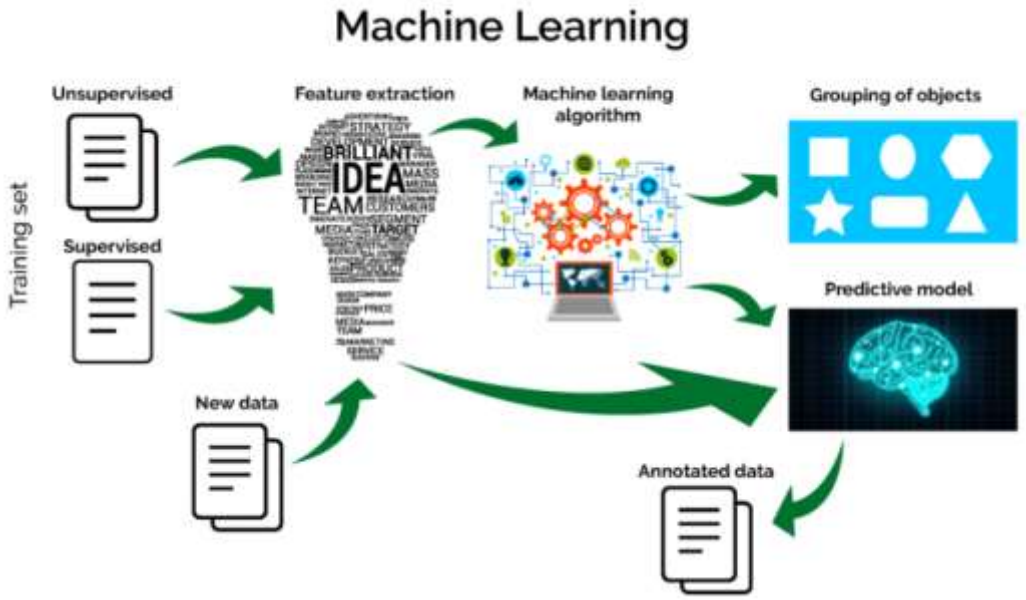
a. Instagram

Instagram adalah salah satu media sosial yang paling banyak digunakan dalam hal jumlah pengguna aktif, Terbukti dalam google play store jumlah download instagram mencapai lebih dari 1 miliar download dan lebih dari 4 juta ulasan masyarakat membuat instagram manjadi salah satu media sosial yang di gandrungi masyarakat saat ini. Instagram digunakan oleh para penggunanya untuk berbagi gambar, karya dan juga sebagai media untuk berkomunikasi [9]. Seiring dengan bertambahnya waktu, peranan media sosial Instagram juga semakin mengalami perkembangan. Selain sebagai media untuk berkomunikasi, Instagram juga digunakan sebagai sarana untuk berbisnis dan politik. Dalam beberapa tahun terakhir, banyak selebriti telah membuat akunnya di Instagram. Para selebriti menggunakan Instagram untuk mengembangkan bisnis dan penggemarnya. Selain itu, banyak selebriti lainnya menggunakan Instagram sebagai platform untuk beriklan. Ketika seseorang telah mendapatkan jumlah follower-nya lebih dari

seratus ribu bahkan jutaan, tidak mengherankan jika menggunakan akun tersebut digunakan sebagai sumber penghasilan yang menguntungkan baginya[1].

b. Machine Learning

Machine learning merupakan serangkaian teknik yang dapat membantu dalam menangani dan memprediksi data yang sangat besar dengan cara mempresentasikan data-data tersebut dengan algoritma pembelajaran.



Gambar 2. 1. Machine Learning (Putra, 2018)

Machine learning juga merupakan salah satu cabang dari artificial intelligence atau kecerdasan buatan dimana mesin dirancang agar dapat belajar sendiri tanpa arahan dari manusia terdapat beberapa metode machine learning yang digunakan data science baik untuk proses klasifikasi, klasterisasi, prediksi maupun identifikasi. Teknologi machine learning (ML) juga nerupakan mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data. Istilah machine learning pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan

dasar- dasar machine learning dan konsepnya. Sejak saat itu ML banyak yang mengembangkan. Salah satu contoh dari penerapan ML yang cukup terkenal adalah Deep Blue yang dibuat oleh IBM pada tahun 1996 [10]. Secara umum terdapat dua teknik atau metode machine learning yang digunakan dalam data science yaitu supervised learning dan unsupervised learning. Dalam penelitian [11] disebutkan bahwa kedua metode tersebut memiliki karakteristik yang dapat diterapkan untuk tujuan tujuan masing-masing penelitian. Pada penelitian ini akan menggunakan Decision tree C4.5 untuk metode Machine Learningnya.

1. Supervised Learning

Teknik supervised learning merupakan teknik yang bisa diterapkan pada pembelajaran mesin yang bisa menerima informasi yang sudah ada pada data dengan memberikan label tertentu. Diharapkan teknik ini bisa memberikan target terhadap output yang dilakukan dengan membandingkan pengalaman belajar di masa lalu. Misalkan terdapat sejumlah film yang sudah diberi label dengan kategori tertentu. Dan juga terdapat film dengan kategori komedi meliputi film 21 Jump Street dan Jumanji. Selain itu juga terdapat kategori lain misalkan kategori film horror seperti The Conjuring dan It. Ketika kita membeli film baru, maka kita akan mengidentifikasi genre dan isi dari film tersebut. Setelah film teridentifikasi barulah kita akan menyimpan film tersebut pada kategori yang sesuai[11].

2. Unsupervised Learning

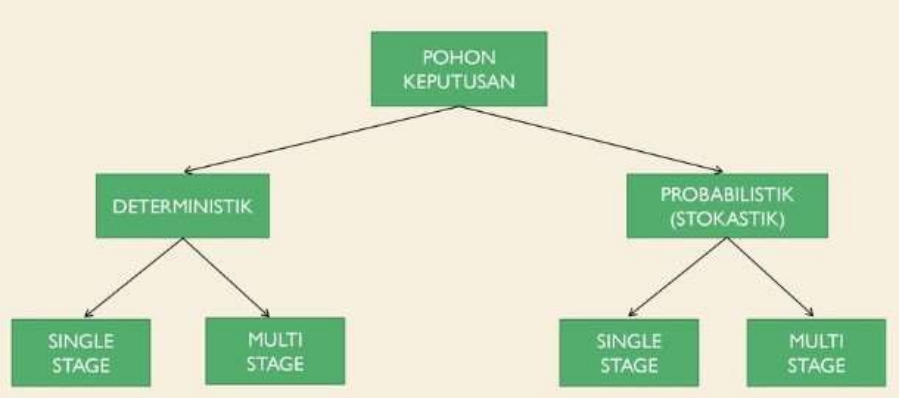
Teknik unsupervised learning merupakan teknik yang bisa diterapkan pada machine learning yang digunakan pada data yang tidak memiliki informasi yang bisa diterapkan secara langsung. Sedikit berbeda dengan supervised learning, kita tidak memiliki data apapun yang akan dijadikan acuan sebelumnya. Misalkan kita belum pernah sekalipun membeli film sama sekali, akan tetapi pada suatu waktu, kita membeli sejumlah film dan ingin membaginya ke dalam beberapa kategori agar mudah untuk ditemukan. Tentunya kita akan mengidentifikasi film-film mana saja yang mirip. Dalam hal ini misalkan kita mengidentifikasi berdasarkan dari genre film. Misalnya, kita mempunyai film the Conjuring, maka kita akan menyimpan film The Conjuring tersebut pada kategori film horror[11].

c. Preprocessing

Preprocessing (pre-processing) adalah tahapan dalam analisis data dan pengolahan data yang melibatkan serangkaian langkah atau teknik untuk membersihkan, mengubah, atau mempersiapkan data mentah sebelum data tersebut digunakan untuk analisis atau pemodelan lebih lanjut. Tujuan utama dari preprocessing adalah untuk memastikan data dalam bentuk yang tepat dan berkualitas sehingga model atau algoritma machine learning dapat bekerja dengan baik dan menghasilkan hasil yang akurat. Berikut adalah beberapa tugas umum yang dilakukan dalam preprocessing data [12]

d. Algoritma Decision Tree C4.5

Decision Tree adalah struktur flowchart berbentuk pohon, dimana simpul bagian dalam merupakan suatu tes pada atribut kemudian setiap cabang menampilkan hasil tes dan simpul daun menampilkan kelas atau penyebaran kelas. Salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan adalah algoritma C4.5. Algoritma C4.5 merupakan algoritma yang sangat populer digunakan oleh banyak peneliti di dunia. Algoritma C4.5 sebagai versi perbaikan ID3 merupakan sebuah algoritma yang diperkenalkan oleh Quinlan. Akan tetapi kelemahan hanya atribut bertipe kategorikal (nominal atau ordinal) saja yang bias di induksi oleh decision tree, sedangkan untuk menangani atribut bertipe numerik interval atau rasio tidak dapat menggunakan algoritma ID3. Sehingga dapat diketahui kelebihan algoritma C4.5 daripada algoritma ID3 antara lain, dapat menangani atribut dengan tipe numerik, memangkas pohon keputusan, dan menurunkan rule set. Penentuan fitur atau atribut sebagai pemecah simpul pada algoritma C4.5 menggunakan kriteria gain dalam pohon yang diinduksi [13].



Gambar 2. 2 Bentuk Pohon Keputusan

Hal pertama yang dilakukan dalam algoritma C4.5 adalah menghitung akar pohon. Akar akan diambil dari atribut – atribut yang akan dipilih, dan dengan menghitung nilai gain dari masing-masing atribut maka nilai gain ratio tertinggi akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, terlebih dahulu menghitung nilai entropy.

$$Entropy(D) = \sum_{i=1}^{m} -p_i \, log_2(p_i)$$

Ket : D = Himpunan kasus

m = Jumlah partisi S

pi = Proporsi Si terhadap S

$$Gain\ (A) = Entropy(D) = \sum_{i=1}^{m} \frac{|D_i|}{|D|} + Entropy\ (D_i)$$

Ket : D = Himpunan Kasus

A = Atribut

n = Jumlah partisi atribut A

$|Si|$ = Jumlah sampel pada partisi ke -i

|S| = Jumlah sampel dalam S

Untuk bisa menghitung Gain Ratio diperlukan sebuah term yang baru yaitu Split

Information. Yang dimana untuk memilih atribut test untuk simpul menggunakan nilai Gain Ratio tertinggi pada setiap atributnya. Split Information menggunakan normalisasi pada Information gain dengan persamaan.

$$SplitInfo_A\ (D) = -\sum_{i=1}^{v} \frac{|Dj|}{|D|}\ x\ log_2(\frac{|Dj|}{|D|})$$

Pemilihan pada Atribut sebagai Root Node menggunakan perhitungan Gain Ratio. Gain Ratio menggunakan persamaan.

$$GainRatio\ (A) = \frac{Gain(A)}{Splitinfo\ A^{(D)}}$$

e. Adaboost

Konsep AdaBoost muncul dari pertanyaan Kearns dan Valiant pada tahun 1988 yaitu apakah suatu pembelajaran lemah dapat ditingkatkan menjadi suatu pembelajaran yang kuat. Jawaban dari pertanyaan tersebut kemudian dijawab oleh Schapire dengan membangun suatu algoritma boosting untuk yang pertama kali. Selanjutnya algoritma ini dikembangkan lagi oleh Freund dan Schapire dengan mengajukan konsep Adaptive Boosting yang dikenal dengan nama AdaBoost. AdaBoost dan variannya telah sukses diterapkan pada beberapa bidang (domain) karena dasar teorinya kuat, prediksi yang akurat dan kesederhanaan yang besar. AdaBoost. Teknik boosting yang digunakan sebagai metode ensemble dalam machine learning disebut adaptive boosting karena bobot atau weights ditetapkan kembali ke setiap instance, dengan bobot yang lebih tinggi ditetapkan ke instance yang diklasifikasikan secara salah [14]. Algoritma adaptive boosting secara iteratif menggabungkan beberapa weak classifiers. Proses ini dimulai dengan bobot yang sama untuk semua data training. Ketika data training salah diklasifikasikan, maka bobot atau weights pada data ini di boosting, maka pengklasifikasi baru dibuat menggunakan bobot baru yang tidak sama. Proses ini diulang untuk satu set classifiers. Ini akan menjaga model training sampai kesalahan error-nya rendah.

Fokus dari metode ini adalah untuk menghasilkan serangkaian base classifiers. Training set yang digunakan untuk setiap base classifiers dipilih berdasarkan performansi dari classifiers sebelumnya. Di dalam boosting, sampel yang tidak

diprediksikan dengan benar oleh classifiers di dalam rangkaian akan dipilih lebih sering dibandingkan dengan sampel yang telah diprediksikan dengan benar. Dengan demikian, boosting mencoba menghasilkan base classifiers baru yang lebih baik untuk memprediksikan sampel yang pada base classifiers sebelumnya memiliki performansi yang buruk [15].

f. Confusion Matrix

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi, *recall, precision,* dan *error rate*. Dimana, *precision* mengevaluasi kemampuan sistem untuk menemukan peringkat yang paling relevan, dan didefinisikan sebagai persentase dokumen yang di *retrieve* dan benar-benar relevan terhadap *query*. *Recall* mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai persentase dokumen yang relevan terhadap *query*. A*ccuracy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan *error rate* merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus[16]. Berikut adalah cara pengukuran performa 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

FP = False Positive

g. K-fold Cross Validation

K-fold Cross Validation adalah teknik yang digunakan dalam machine learning untuk mengukur kinerja model dengan lebih akurat dan menghindari overfitting. K-Fold Cross Validation Bekerja dengan proses validasi model yang dilakukan dengan cara membagi dataset menjadi k bagian atau fold, dan dilakukan iterasi sebanyak k kali. Pada setiap iterasi, setiap bagian atau fold digunakan sebagai testing dataset sebanyak satu kali secara bergantian. Tahapanya adalah Pembagian Data, Iterasi, Pelatihan dan Pengujian, Evaluasi Kinerja dan Kinerja Rata-rata Bagian yang lainnya k-fold digunakan sebagai training dataset. Hal ini bertujuan untuk melakukan testing terhadap model menggunakan data yang belum pernah dilihat sebelumnya. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model machine learning [17].

h. Fake Account dan Real Account

Real account dinstagram adalah akun yang dibuat oleh pengguna yang menggunakan identitas asli mereka. Akun ini biasanya memiliki foto profil, nama lengkap, bio, dan konten yang sesuai dengan minat dan kepribadian pengguna. Real account berbeda dengan fake account, yang merupakan akun yang dibuat oleh pengguna yang menyembunyikan atau mengubah identitas mereka. Fake account sering digunakan untuk tujuan negatif, seperti menipu, mengganggu, atau mengikuti orang lain tanpa izin.