Implementasi Algoritma Naïve Bayes Untuk Analisis Klasifikasi Survei Kesehatan Mental (Studi Kasus: Open Sourcing Mental Illness)

Reza Alfaresy Chaerudin[1], Ermatita[2], Ruth Mariana Bunga Wadu[3]
[1,2,3] Program Studi Sistem Informasi, Fakultas Ilmu Komputer
[1,2,3] Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3] Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat
rezaa@upnvj.ac.id[1], ermatitaz@yahoo.com[2], ruthbungawadu@upnvj.ac.id[3]

Abstrak. Kesehatan mental telah menjadi sorotan penting dalam kehidupan masyarakat sekarang, dan tidak luput dari berbagai industri dalam dunia kerja, termasuk industri teknologi. Kesadaran akan kepentingan kesehatan mental pekerja masih sering dianggap rendah, dan hal ini juga tidak luput dalam industri teknologi, oleh karenanya Open Source Mental Illness (OSMI), sebagai lembaga yang bergerak di bidang kesehatan mental, mengadakan survei untuk mengetahui kesadaran mengenai kesehatan mental pada pekerja dalam industri teknologi. Hasil dari survei ini telah dirilis sebagai dataset, di mana dataset ini kemudian dapat dianalisis lebih lanjut menggunakan data mining dengan metode klasifikasi sebagai analisis kesadaran kesehatan mental berdasarkan data pada survei. Algoritma klasifikasi yang digunakan adalah Naïve Bayes, yang mana hasil klasifikasi ini dapat digunakan lebih dalam untuk analisis lanjut mengenai kesadaran pengaruh kesehatan mental pada pekerja industri teknologi, dalam bentuk model prediksi. Dataset yang digunakan awalnya terdiri dari 1259 record data, dimana setelah dilakukan praproses didapatkan 1254 record data. Penelitian ini dilakukan ujicoba dengan pembagian data uji sebesar 30% dan data latih sebesar 70%, dimana didapatkan hasil akurasi sebesar 72%. Analisis data mining ini kemudian dilaksanakan dalam bahasa pemrograman Python, untuk mendapatkan suatu model prediksi sederhana yang kemudian digunakan untuk sistem prediksi sederhana berbasis website.

Kata Kunci: *Data Mining*, klasifikasi, Naïve Bayes, kesehatan mental, kesadaran kesehatan mental, industri teknologi.

1 Pendahuluan

Perubahan teknologi di zaman yang modern ini telah menyebabkan perambatan aspek teknologi pada semua ranah kehidupan. Mulai dari industri, ekonomi, ilmu, hingga salah satu bidang kehidupan yang paling penting, yaitu kesehatan. Kesehatan in tidak hanya terdiri dari kesehatan fisik saja, namun juga kesehatan mental.

Dalam industri teknologi itu sendiri, terdapat defisiensi pada pengertian mendasar mengenai seberapa pentingnya masalah kesehatan mental pada lingkungan kerja industry teknologi tersebut. Tidak sedikit perusahaan yang kurang memperhatikan efek yang timbul pada pekerja mereka mengenai kondisi kesehatan mental yang mereka miliki. Ini dibuktikan oleh riset yang dilakukan pada mahasiswa jurusan *STEM* (*Science, Technology, Engineering, Math*) yang belajar di *University of California, Berkeley*, lebih dari 50% mahasiswa mengidap masalah kesehatan mental. [1] Hal ini disadari oleh *Open Source Mental Illness* (OSMI), suatu perusahaan *non-profit* yang bergerak dalam bidang kesehatan, khususnya kesehatan mental, pada industri dan komunitas teknologi. Mereka pun menjalankan survei mengenai kesadaran kesehatan mental pada pekerja dalam industry teknologi pada tahun 2014. Hasil dari survei masih merupakan data-data, yang kemudian dapat dilakukan analisis lebih lanjut.

Salah satu analisis yang dapat dilakukan pada hasil data survei ini adalah analisis klasifikasi. Analisis klasifikasi adalah suatu teknik yang ditujukan untuk menganalisis keterhubungan antara variabel-variabel sebagai prediktor untuk suatu kelas atau kategori variabel respon. Metode pengklasifikasian ini akan melakukan prediksi pada setiap peluang dari kategori dalam variabel kualitatif sehingga menjadi dasar untuk pembuatan klasifikasi itu sendiri. [2]

Proses klasifikasi pada hasil data survei ini akan dijadikan sebagai dasar untuk sebuah analisis dan sistem prediksi kesadaran mengenai pengaruh kesehatan mental pada pekerja dalam industri teknologi; analisis ini dapat digunakan sebagai indikator untuk mengetahui apakah pekerja dalam industri teknologi sendiri sudah sadar betul dengan pengaruh kesehatan mental secara mendasar. WHO telah memprediksi bahwa pada dua dekade yang akan datang, lebih dari 300 penduduk dunia akan mengalami gangguan kesehatan mental jika tidak dilakukan aksi preventif sebelumnya [3], sehingga perlu adanya pengetahuan dasar mengenai sudahkah

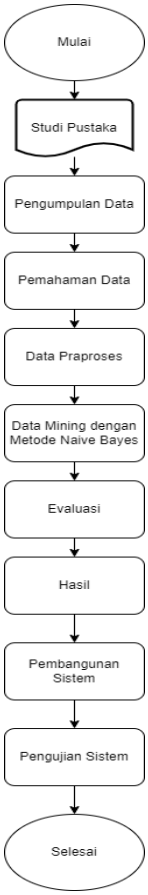
ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

pekerja industri teknologi mengenal betul dengan dampak kesehatan mental agar dapat menghindari kemungkinan buruk ini.

Algoritma yang digunakan pada analisis adalah *Naïve Bayes.* Algoritma *Naïve Bayes* merupakan salah satu metode klasifikasi dimana dilakukan prediksi peluang yang akan terjadi di masa depan berdasarkan data yang telah ada sebelumnya. Algoritma ini berdasarkan teorema Bayes yang didasari oleh penghitungan peluang untuk suatu hipotesis kasus [4].

Alasan mengapa algoritma ini dipilih untuk analisis klasifikasi pada studi kasus ini adalah karena tingkat akurasi *Naïve Bayes* yang cukup tinggi dan mudah diaplikasikan pada studi kasus yang diteliti. Akurasi dari algoritma *Naïve Bayes* ini juga nanti akan ditampilkan untuk mengetahui seberapa besar persentase akurasi dari klasifikasi dengan algoritma *Naïve Bayes* untuk data kasus ini. Keakurasian dan ketepatan algoritma telah dibuktikan oleh beberapa penelitian sebelumnya yang membandingkan ketelitian antara algoritma-algoritma klasifikasi [5] Salah satu contohnya adalah penelitian yang melakukan perbandingan *Naïve Bayes* dan *Support Vector Machine (SVM),* dimana ditemukan perbedaan hingga 15% antara kedua algoritma [6] Perbedaan ini bisa berubah tergantung pada jenis dan jumlah data yang digunakan, sehingga perlu dilakukan beberapa percobaan dengan perhitungan yang dilakukan pada algoritma *Naïve Bayes* untuk menentukan persentase terbesar terdapat pada perhitungan tertentu.

2 Metodologi Penelitian



Gambar. 1. Alur Penelitian

2.1 Studi Pustaka

Pada tahap ini, penulisan proposal tugas akhir ini diawali dengan studi pustaka. Studi pustaka ini ditujukan untuk memperdalam ilmu teori yang digunakan sebagai dasar dari penelitian ini, dengan fokus ada pada studi *data mining*, khususnya teori algoritma klasifikasi *Naïve Bayes*. Pada studi pustaka, penulis mencari teori-teori mendalam mengenai *data mining*, algoritma klasifikasi, dan kepentingan kesehatan mental itu melalui jurnal dan buku yang didapatkan secara *online*.

2.2 Pengumpulan Data

Pada tahap pengumpulan data, data yang akan dipakai untuk penelitian ini didapatkan dari sumber tertentu. Pengumpulan data ini juga dapat dibilang bersifat studi dokumen, karena data yang didapat berdasarkan dokumen yang telah dipublikasikan secara umum. Setelah melakukan sedikit riset mengenai topik penelitian, dataset penelitian yang berupa hasil survei kesehatan mental pekerja industri teknologi ini didapatkan langsung dari *website official Open Source Mental Illness,* yang menyediakan langsung dataset survey. *Dataset* ini juga tersedia lewat *website Kaggle*. Data yang diterima langsung dari *website* OSMI telah berbentuk *.csv*, sehingga bisa langsung dipakai pada *Jupyter Notebook* tanpa harus merubah bentuk format data.

2.3 Pemahaman Data

Sebelum data dapat di proses lebih lanjut, diperlukan pemahaman dasar pada dataset yang akan dipakai, seperti variabel apa yang akan dianalisis (atau disebut variabel fitur), berapa banyak data yang dipakai, atau variabel prediksi apa yang ingin dihasilkan pada analisis klasifikasi data mining penelitian ini (atau disebut variabel label).

2.4 Data Praproses

Tahapan data praproses ditujukan untuk membersihkan data yang akan dipakai pada analisis. Pembersihan ini dilakukan dengan mengecek apakah ada data yang kosong, hilang atau tidak lengkap untuk menghindari ketidaktepatan prediksi, dan reduksi variabel dalam dataset yang akan digunakan untuk analisis untuk memastikan data memang relevan pada tujuan analisis. Pada penelitian kasus ini, data praproses dilaksanakan dengan mengurangi variabel data yang akan dipakai pada analisis. Data kemudian akan diperiksa untuk memastikan apakah ada kekosongan pada data yang perlu dibersihkan sebelum bisa dilanjutkan proses. Variabel fitur-fitur ini akan digunakan untuk mencari target variabel label yang sudah ditentukan, yaitu obs_consequence.

2.5 Data Mining dengan Metode Naïve Bayes

Metode klasifikasi yang digunakan salah satunya adalah Naïve Bayes Classifier. Data diproses melalui bahasa pemrograman Python dengan platform Jupyter Notebook. Algoritma klasifikasi yang dipilih berfungsi untuk memproses dataset, membentuk pola dengan melakukan data training, dan mengeluarkan output nilai yang digunakan sebagai basis klasifikasi untuk dataset hasil survei kesehatan mental OSMI ini, di mana hasil output prediksi berupa Yes dan No untuk kesadaran pekerja industri teknologi mengenai pengaruh kesehatan mental.

Algoritma Naïve Bayes Classifier memiliki beberapa tahapan yang akan dilakukan ketika melakukan klasifikasi, yaitu [7]:

1) Mempersiapkan data training. Data training ini berasal dari dataset survei yang dianalisis. Kita perlu menentukan seberapa banyak data training dan data testing yang akan dipakai, misalnya 30% data testing dan 70% data training. Pembagian data ini akan mempengaruhi hasil klasifikasi yang dilakukan.
2) Menghitung jumlah data dan probabilitas yang ada. Perhitungan ini menggunakan rumus perhitungan algoritma Naïve Bayes. Perhitungan ini termasuk perhitungan posterior, contohnya perhitungan probabilitas pada kelas/variabel label.
3) Mencari nilai probabilitas yang muncul dari penghitungan jumlah data sesuai dengan kategori yang sama, yang kemudian dibagi jumlah data kategori tersebut. Baru setelah semua probabilitas telah dihitung dapat dilakukan proses klasifikasi, dimana data testing akan diuji dan diprediksi kelas/variabel label berdasarkan model prediksi dengan probabilitas yang telah dihitung.
4) Membuat dan menampilkan hasil evaluasi. Setelah melakukan semua perhitungan probabilitas dan klasifikasi, kita dapat mencoba membuat confusion matrix dari hasil yang dilakukan. Setelah dijabarkan confusion matrix yang didapatkan, maka dapat ditentukan nilai precision, recall, specificity dan akurasi dari perhitungan sebelumnya.

2.6 Evaluasi

Evaluasi dilakukan setelah data telah diproses melalui klasifikasi dengan algoritma Naïve Bayes. Informasi yang dievaluasi berupa tingkat akurasi dari prediksi yang dilakukan oleh proses klasifikasi, yang disampaikan melalui tabel confusion matrix.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

2.7 Hasil

Setelah dilakukan evaluasi, maka hasil klasifikasi dari algoritma Naïve Bayes akan ditampilkan dalam bentuk classification report dalam Python. Akurasi yang didapatkan akan kemudian juga ditampilkan dalam bentuk persentase berdasarkan rumus yang telah ditentukan.

2.8 Pembangunan Sistem

Setelah memastikan model prediksi yang dibuat dengan Python telah berjalan dengan baik, selanjutnya adalah perancangan sistem model prediksi kesadaran kesehatan mental dengan algoritma Naïve Bayes berbasis website. Kemudian dilakukan pembangunan sistem. Sistem yang dibuat menggunakan framework Flask, dimana framework ini dipilih karena dapat mengimplementasikan Python pada website yang akan dibuat. File model prediksi yang dibuat pada Python akan diekspor, lalu kemudian diimpor pada sistem website yang dibangun.

2.9 Pengujian Sistem

Tahap terakhir adalah implementasi dan pengujian sistem. Untuk tahap ini, dilakukan proses untuk memastikan apakah sistem yang dibangun telah berjalan dan diimplementasikan dengan baik. Pengujian yang dilakukan menggunakan Blackbox Testing. Blackbox Testing digunakan untuk menguji hasil respon dari sistem, apakah sudah sesuai dengan hasil yang diharapkan pada rancangan sistem website.

3 Hasil dan Penelitian

3.1 Pengujian SistemPraproses Data

Pada temuan awal dataset, dapat diamati bahwa mayoritas responden survei menyatakan tidak sadar mengenai pengaruh kesehatan mental pada lingkungan kerja dalam industri teknologi (kolom obs_consequence), dengan 1075 data No dibandingkan 184 data Yes. Dikarenakan adanya kesenjangan besar antara kedua kelas data, maka dilakukan teknik resampling yang ditujukan untuk meningkatkan ketepatan model dalam melakukan klasifikasi untuk setiap kelas. Teknik resampling yang dipilih adalah SMOTE (Synthetic Minority Oversampling Technique). SMOTE merupakan suatu varian dari teknik oversampling, yaitu membuat data sintesis baru dalam kelas minoritas untuk penyeimbangan data kelas [8]. Teknik SMOTE ini akan diaplikasikan pada tahap pembuatan data latih, karena data sintesis harus dihindarkan pada data uji agar akurasi model yang dibuat memang sesuai dengan data yang nyata.

3.2 Implementasi Algoritma Naïve Bayes

3.2.1 *Data Cleaning*

Proses Data Cleaning, atau Pembersihan Data, adalah proses yang merujuk pada tindakan untuk "membersihkan" data yang akan dipakai dari kekurangan yang dapat menggangu proses analisis yang akan dijalankan. Dalam proses pembersihan data ini, yang pertama dilakukan adalah pengecekan pada dataset survei, apakah ada kolom variabel data yang masih memiliki missing value atau redundansi data. Dari semua variabel, ditemukan bahwa ada 4 variabel yang memiliki missing value yaitu state dengan 515 data kosong, self_employed dengan 18 data kosong, work_interfere dengan 264 data kosong, dan comment dengan 1095 data kosong. Dari empat variabel ini, ada dua variabel yang memiliki tingkat kekosongan yang melebihi 50% dari jumlah data, yaitu state (mencapai lebih dari 50% dari data) dan comment (yang mencapai lebih dari 85% kekosongan data). Oleh karena itu, kedua kolom ini akan kemudian dihapus dan tidak digunakan pada proses analisis, karena variabel-variabel ini tidak dapat membantu saat analisis dengan besarnya kekosongan data yang ada.

Sementara itu, untuk kolom self-employed dan work_interfere, akan dilakukan pengisian kekosongan data dengan cara mengisi entry yang kosong dengan rata-rata data untuk kolom tersebut menggunakan Python. Untuk self_employed, rerata isi data untuk kolom tersebut adalah No (1077 data dari total 1259 data), maka missing value tersebut akan diisi dengan data No. Sementara itu, untuk work_interfere, rerata isi data untuk kolom ini adalah "Sometimes" (202 data), maka entry yang masih kosong akan diisi dengan data "Sometimes".

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

Ada juga kejanggalan yang ditemukan pada kolom Age, dimana kolom Age memiliki umur dibawah 0 tahun dan 100 tahun. Karena kolom ini tidak logis, maka perlu dihapus terlebih dahulu agar tidak menganggu analisis yang berjalan. Hasilnya menjadi 1254 record data yang sudah dibersihkan. Selain itu, ditemukan bahwa dalam 1254 record data, semua data adalah distinct atau unik, sehingga bisa dipastikan bahwa tidak ada redundansi data yang terjadi.

3.2.2 *Data Selection*

Proses selanjutnya adalah Data Selection, atau Pemilihan Data. Kolom yang akan digunakan adalah kolom yang lebih berhubungan langsung dengan tujuan penelitian, yaitu kesadaran mengenai pengaruh kesehatan mental pada pekerja industri teknologi. Kolom yang dipilih juga merupakan kolom yang memang langsung mengenai responden itu sendiri, bukan perusahaan yang responden berasal. Pemilihan kolom ini berdasarkan studi pustaka yang telah peneliti lakukan sebelumnya, misalnya dipilih kolom Age untuk digunakan sebagai variabel fitur karena telah dilakukan studi mengenai hubungan rentang umur dengan kondisi kesehatan mental [9], ataupun Gender untuk masuk sebagai variabel fitur karena telah ditemukan korelasi antara jenis kelamin dengan kesadaran kesehatan mental [10].

Hasilnya, ada 10 variabel yang akan digunakan untuk analisis klasifikasi data mining ini sebagai fitur, yaitu Age, Gender, family_history, treatment, work_interfere, mental_health_consequence, coworkers, supervisor, mental_health_interview, dan mental_vs_physical. Variabel label dari analisis ini adalah obs_consequences.

3.2.3 *Data Transformation*

Proses berikutnya adalah Data Transformation, atau Transformasi Data. Pada tahap ini, ada dua variabel yang ini ditransformasikan terlebih dahulu, yaitu kolom Age dan Gender. Kolom Age masih berupa data integer yang numerikal (dengan rentang nilai yang cukup besar, sehingga sulit untuk dijadikan acuan penelitian secara langsung), sehingga untuk menyamakan data ini dengan data variabel ini yang utamanya adalah data kategorikal, perubahan yang dilakukan adalah:

Tabel 1. *Data Transformation Age*

Kategori *Age*	Rentang *Age*
Young Adult	<30
Adult	30-50
Elder	>50

Transformasi yang dilakukan selanjutnya adalah untuk kolom Gender. Pada kolom Gender, terlihat bahwa pilihan responden tidaklah seragam, walaupun mereka mengarah pada jawaban yang sama atau mirip, sehingga rentang nilai untuk kolom Gender sulit untuk dijadikan acuan secara langsung. Oleh karena itu, dilakukan juga penyamaan data dalam kolom Gender menjadi data kategorikal yang seragam, dengan perubahan seperti berikut:

Tabel 2. *Data Transformation Gender*

Kategori *Gender*	Rentang *Gender*
Male	*Male, male, M, m, Male, Cis Male, Man, cis male, Mail, Male-ish, Male (CIS), Cis Man, msle, Malr, Mal, maile, Make*
Female	*Female, female, F, f, Woman, Female, femail, Cis Female, cis-female/femme, Femake, Female (cis), woman*
Other	*Female (trans), queer/she/they, non-binary, fluid, queer, Androgyne, Trans-female, male leaning androgynous, Agender, A little about you, Nah, All, ostensibly male, unsure what that really means, Genderqueer, Enby, p, Neuter, something kinda male?, Guy (-ish) ^_^, Trans woman*

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

3.3 Praktik *Data Mining Python* dengan Naïve Bayes

3.3.1 Persiapan Data Uji dan Data Latih

Langkah pertama dalam melakukan perhitungan Naïve Bayes adalah menentukan dulu data uji dan data latih yang akan dipakai dalam analisis. Untuk pengujian ini, ditentukan bahwa data uji yang dipakai adalah 30% dari dataset, sementara data latih yang dipakai adalah 70% dari dataset. Dari 1254 data, maka didapatkan sebanyak 376 data uji dan 783 data latih. Selain itu, ada juga 1072 data pada kelas No, dan 182 data pada kelas Yes. Setelah dibagi dalam data uji dan data latih, didapatkan bahwa pada data uji, ada 327 data kelas No, dan 50 data kelas Yes. Sementara itu, pada data latih, didapatkan 745 data kelas No, dan 132 data kelas Yes.

Namun, sebelum data latih dapat digunakan, perlu diingat bahwa perlu dilakukan teknik resampling, yaitu SMOTE, pada data latih terlebih dahulu sebelum dimulai perhitungan. Hal ini diperlukan untuk meningkatkan ketepatan model untuk mengklasifikasikan kelas minoritas, karena model akan kesulitan melakukan klasifikasi dengan tepat pada kelas minoritas yang memiliki jumlah data yang terlalu kecil. Teknik SMOTE akan membuat suatu data sintetis baru berdasarkan data kelas minoritas yang telah ada, sehingga data tersebut akan memenuhi kelas minoritas dalam proses oversampling, dimana untuk pembangunan model ini, ditentukan bahwa resampling SMOTE untuk kelas minoritas yang dilakukan adalah sebesar 70% dari data kelas mayoritas. Penggunaan teknik SMOTE ini akan dilakukan dengan Python. Berikut adalah tabel jumlah data setelah dilakukan resampling pada data uji:

Tabel 3. Tabel Jumlah Data

Jumlah Data (1254)	Data Latih (1266)	Data Uji (377)
No (1072)	745	327
Yes (182)	521	50

3.3.2 Perhitungan Probabilitas Dari Jumlah Data

Langkah selanjutnya adalah menghitung probabilitas dari jumlah data yang didapatkan. Perhitungan dilakukan dengan data latih yang telah ditentukan.

$$P(Yes|\quad) = \frac{Yes}{Total\ Data} \tag{1}$$

$$P(Yes|\quad) = \frac{521}{1266} = 0{,}411532385466034755134281200631 91$$

$$P(No|\quad) = \frac{No}{Total\ Data} \tag{2}$$

$$P(No|\quad) = \frac{745}{1266} = 0{,}588467614533965244865718799368 09$$

3.3.3 Perhitungan *Prior Probability*

Langkah berikutnya adalah menghitung prior probability. Langkah ini dilakukan dengan menghitung setiap probabilitas dari setiap variabel untuk setiap kelas. Ada 10 variabel yang digunakan sebagai fitur dalam analisis. Hasil dari setiap perhitungan probabilitas untuk setiap variabel akan disajikan dalam bentuk tabel conditional probability. Contoh perhitungan yang ditampilkan adalah untuk variabel treatment:

$$P(treatment\ Yes|Yes) = \frac{treatment\ Yes.P(Yes)}{Total\ Data\ Yes} \tag{3}$$

$$P(treatment\ Yes|Yes) = \frac{294}{521} = 0.564299424184261036468330134 3570$$

$$P(treatmentYes|No) = \frac{treatment\ Yes.P(No)}{P(No)} \tag{4}$$

$$P(treatment\ Yes|No) = \frac{422}{745} = 0.566442953020134228187919463 08725$$

$$P(treatment\ No|Yes) = \frac{treatment\ No.P(Yes)}{P(Yes)} \tag{5}$$

$$P(treatment\ No|Yes) = \frac{218}{521} = 0.418426103646833013435700575815 74$$

$$P(No|No) = \frac{treatment\ No.P(No)}{P(No)} \quad (6)$$

$$P(treatment\ No|No) = \frac{314}{745} = 0,421476510067114093959731543624 16$$

Tabel 4. Tabel *Conditional Probability treatment*

Treatment	obs_consequence	
	Yes	***No***
Yes	0.564	0.566
No	0.418	0.422

Setelah melakukan semua perhitungan probabilitas untuk setiap variabel, maka kita dapat mencoba menguji perhitungan probabilitas dengan sebuah data baru, dimana data baru tersebut memiliki data sebagai berikut: Age Adult, Gender Female, family_history Yes, Treatment No, work_interfere Often, mental_health_consequence Maybe, coworkers No, supervisor No, mental_health_interview No, mental_vs_physical Yes.

Perhitungan yang dilakukan adalah Joint Probability Distribution, yaitu perhitungan probabilitas berdasarkan perhitungan prior probablity, dimana perhitungan ini dilakukan dengan menghitung probabilitas untuk data baru dengan parameter yang ditentukan adalah hasil prior probability sebelumnya. Maka, perhitungan yang dilakukan adalah sebagai berikut:

Probability (*Yes*):

$$P(Yes) * P(Age\ Adult|\ Yes) * P(Gender\ Female|\ Yes) * P(family\_history\ Yes\ |\ Yes) * P(treatment\ No|\ Yes) * P(work\_interfere\ Often\ |\ Yes) * P(mental\_health\_consequence\ No\ |\ Yes) * P(coworkers\ No|\ Yes) * P(supervisor\ No|\ Yes) * P(mental\_health\_interview\ No\ |Yes) * P(mental\_vs\_physical\ No\ |\ Yes)$$

$$=0.41 * 0.474 * 0.250 * 0.396 * 0.418 * 0.140 * 0.319 * 0.204 * 0.340 * 0.816 * 0.382$$

$$= 7.76524209 \times 10^{-6} = 39.8\%$$

Probability (*No*):

$$P(No) * P(Age\ Adult|\ No) * P(Gender\ Female|\ No) * P(family\_history\ Yes\ |\ No) * P(treatment\ No|\ No) * P(work\_interfere\ Often\ |\ No) * P(mental\_health\_consequence\ No\ |\ No) * P(coworkers\ No|\ No) * P(supervisor\ No|\ No) * P(mental\_health\_interview\ No\ |No) * P(mental\_vs\_physical\ No\ |\ No)$$

$$= 0.59 * 0.478 * 0.250 * 0.397 * 0.422 * 0.140 * 0.319 * 0.204 * 0.344 * 0.825 * 0.384$$

$$= 1.17277683 \times 10^{-5} = 60.2\%$$

3.3.4 Pembuatan Hasil Evaluasi

Langkah terakhir adalah membuat hasil evaluasi, yaitu dalam bentuk confusion matrix. Confusion matrix ini dilakukan setelah melakukan ujicoba perhitungan data dengan model prediksi probabilitas seperti yang dibuat sebelumnya. Pada pengujian ini, dengan ukuran 30% data uji, didapatkan 239 data TP (True Positive), 31 TN (True Negative), 19 FP (False Positive), dan 88 FN (False Negative). Totalnya, ada 270 data yang benar diklasifikasi, dan 116 data yang salah diklasifikasi.

Tabel 5. Tabel *Confusion Matrix*

Confusion Matrix		***Predicted Class***	
		Positive	***Negative***
Actual Class	*Positive*	239	19
	Negative	88	31

Dari tabel yang telah ditampilkan, maka dapat dihitung precision, recall dan specificity sebagai berikut:

$$Precision = \frac{239}{239+88} \quad (7)$$
$$Precision = 0.73$$
$$Recall = \frac{239}{239+19} \quad (8)$$
$$Recall = 0.92$$
$$Specificity = \frac{31}{31+88} \quad (9)$$
$$Specificity = 0.26$$

Sehingga, akurasi yang didapatkan dari pengujian analisis ini adalah sebagai berikut:

$$Accuracy = \frac{\sum Correct}{n} \times 100\% \quad (10)$$
$$Accuracy = \frac{270}{376} \times 100\%$$
$$Accuracy = 71.6\% = 72\%$$

Maka, akurasi yang didapatkan dari perhitungan metode Naïve Bayes adalah 72%

Untuk melakukan evaluasi lebih lanjut, maka dilakukan juga perbandingan akurasi dengan perhitungan pada pembagian data uji dan data latih yang berbeda. Dalam percobaan ini, selain menggunakan perhitungan 70% data latih dan 30% data uji, ada juga perhitungan untuk pembagian 60% data latih dan 40% data uji, serta 80% data latih dan data uji. Percobaan perhitungan tersebut menghasilkan evaluasi sebagai berikut:

Tabel 6. Evaluasi Hasil Akurasi Percobaan

No	*Data Latih*	*Data Uji*	*Precission*	*Recall*	*Specificity*	*Akurasi*
1	70%	30%	0.73	0.92	0.26	72%
2	60%	40%	0.74	0.90	0.24	71%
3	80%	20%	0.70	0.90	0.27	69%

Selain itu, perlu diingat bahwa perhitungan dilakukan dengan data latih yang telah di-resampling. Tujuan awal dari resampling adalah untuk meningkatkan akurasi untuk kelas minoritas dari data yang diteliti, karena adanya kesenjangan besar antara data kelas mayoritas dan kelas minoritas. Oleh karena itu, dilakukan juga perhitungan untuk data sebelum dilakukan resampling, dengan evaluasi sebagai berikut:

Tabel 7. Tabel Perbandingan Evaluasi untuk Data Kelas Minoritas Sebelum dan Sesudah Resempling

Resampling Kelas Minoritas	*Data Latih*	*Data Uji*	*Precision*	*Recall*	*F1-Score*
70%	30%	0.73	0.92	0.26	72%
60%	40%	0.74	0.90	0.24	71%
80%	20%	0.70	0.90	0.27	69%

Dari evaluasi kelas minoritas yang telah dilakukan pada data sebelum dan sesudah di-resampling, maka dapat dilihat bahwa walaupun model prediksi dengan data yang telah resampling memiliki akurasi lebih rendah daripada data yang belum resampling, akurasi dari model prediksi pada data yang telah resampling dapat lebih baik melakukan prediksi untuk data kelas minoritas, dalam kasus ini kelas Yes, sehingga dapat lebih handal digunakan sebagai prediktor untuk data kelas No dan Yes.

3.4 Praktik *Data Mining Python* dengan Naïve Bayes

Setelah melakukan perhitungan analisis dengan algoritma Naïve Bayes, maka langkah terakhir adalah melakukan praktik langsung data mining menggunakan bahasa pemrograman Python. Praktik ini dilakukan untuk membandingkan dengan perhitungan dari algoritma Naïve Bayes yang telah dilakukan, serta membuat model prediksi dari klasifikasi yang telah dilakukan.

3.5 Perancangan dan Pembangunan Sistem

Setelah model prediksi telah dibuat dan dibuktikan berhasil dipakai, tahap selanjutnya adalah perancangan dan pembangunan sistem sederhana untuk prediksi kesadaran mengenai kesehatan mental pada pekerja teknologi industri. Sistem yang dibangun ini adalah berbentuk basis website, dimana website itu sendiri dirancang dengan framework Flask, dan diimplementasikan algoritma data mining Naïve Bayes dengan Python.

3.6 Implementasi dan Pengujian Sistem

Setelah sistem telah dirancang dan dibangun, langkah terakhir adalah implementasi dan pengujian sistem. Untuk pengujian sistem ini, dilakukan blackbox testing, yaitu pengujian yang digunakan untuk melihat bagaimana proses input-output pada sistem, memastikan output yang dihasilkan sudah sesuai.

Pengujian sistem yang dilakukan adalah sebagai berikut:

Tabel 8. Ujicoba Blackbox Testing pada Website

Modul	*Input*	*Proses*	*Output*	*Status*
Halaman Utama	Akses ke URL Halaman Utama	Mengarahkan ke sistem website dan Halaman Utama	Menampilkan halaman Utama	*Success*
	Klik bar About Us	Mengarahkan ke halaman About Us	Menampilkan halaman About Us	*Success*
	Klik button Start Prediction	Mengarahkan ke halaman Pengisian Data	Menampilkan halaman Pengisian Data	*Success*
Halaman Pengisian Data	Klik button Submit	Mengarahkan ke halaman Prediction	Menampilkan halaman Prediction	*Success*
	Klik button Reset	Menghapus semua input data pada halaman	Semua data di halaman telah kosong kembali	*Success*
Halaman *Prediction*	Klik button Retry Prediction	Mengarahkan kembali ke halaman Pengisian Data	Menampilkan kembali halaman Pengisian Data	*Success*

4 Kesimpulan dan Saran

Dari penelitian yang telah dilakukan, kesimpulan yang dapat ditarik adalah hasil klasifikasi yang didapatkan dari prediksi data survei kesehatan pada pekerja dalam industri teknologi merupakan suatu model prediksi yang dapat digunakan untuk memprediksi apakah responden yang bekerja dalam industri teknologi bisa dikatakan sebagai sudah sadar atau belum mengenai pengaruh kesehatan mental, selain itu hasil akurasi yang didapatkan untuk analisis klasifikasi yang dilakukan dengan algoritma Naïve Bayes adalah 72% pada ukuran data uji 30% dan data latih 70%, yang merupakan hasil akurasi tertinggi yang didapatkan setelah dilakukan percobaan dengan ukuran pembagian data lain pada saat percobaan.

Saran yang dapat diberikan oleh penulis untuk mengembangkan topik penelitian lebih lanjut adalah seperti misalnya akurasi dari penelitian ini dapat ditingkatkan lebih baik, karena masih berada di cakupan akurasi 70-80%. Langkah yang dapat dilakukan untuk meningkatkan akurasi prediksi misalnya mengganti ukuran data uji

dan data latih, mengganti metode atau jumlah resampling yang digunakan, dan seterusnya sehingga didapatkan nilai akurasi yang lebih tinggi. Selain itu juga, analisis dapat dilakukan dengan algoritma-algoritma klasifikasi lain, seperti Decision Tree, K-Nearest Neighbor dan Apriori. Perbedaan algoritma ini juga dapat dibandingkan hasilnya, manakah algoritma yang lebih akurat dalam melakukan klasifikasi pada analisis data mining hasil survei.

Referensi

[1] C. Murphy and J. Akullina, "We're All in This Together: CS Students, the Tech Industry, and Mental Health (Abstract Only)," in *SIGCSE '18: Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018, pp. 1071, doi: 10.1145/3159450.3162189.

[2] F. Sohil, M. U. Sohali, and J. Shabbir, "An introduction to statistical learning with applications in R," Stat. Theory Relat. Fields, 2021, doi: 10.1080/24754269.2021.1980261.

[3] WHO, "WHO Mental Health Atlas 2017," *World Heal. Organ.*, vol. 2016, 2018.

[4] M. Sudha and B. Poorva, "Predictive tool for dermatology disease diagnosis using machine learning techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9, pp. 355–360, 2019, doi: 10.35940/ijitee.g5376.078919.

[5] K. Pal and B. V Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," *in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 83–87, doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[6] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," *in 2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 2018, pp. 1–6, doi: 10.1109/CITSM.2018.8674352.

[7] Bustami, "Penerapan Algoritma Naive Bayes Untuk Nasabah Asuransi," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2018.

[8] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[9] E. L. Yearwood and V. P. Hines-Martin, "Editorial: Impact of social determinants of health on mental health," *Arch. Psychiatr. Nurs.*, vol. 35, no. 1, pp. A1–A2, 2021, doi: 10.1016/j.apnu.2020.12.001.

[10] S. Yu, "Uncovering the hidden impacts of inequality on mental health: a global study," *Transl. Psychiatry*, 2018, doi: 10.1038/s41398-018-0148-0.