DOI: 10.22146/jnteti.Nomor DOI

Comparison of C4.5 and XGBoost Models with Hyperparameter Tuning for Mental Health Predictio

Jufourlisa Sirait[1], Kezia Hutagaol[2], Reinaldy Hutapea[3], Tennov Pakpahan[4]

Github: https://github.com/tennov1011/Project-Data-Mining.git

INTISARI — Penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost dan Decision Tree C4.5 untuk memprediksi depresi berdasarkan faktor sosial, ekonomi, dan psikologis, menggunakan dataset Exploring Mental Health Data. Proses dilakukan dengan pendekatan CRISP-DM yang mencakup pemahaman bisnis, eksplorasi data, persiapan data, pembangunan model, evaluasi, hingga deployment. Hasil evaluasi menunjukkan bahwa XGBoost dengan Grid Search Tuning memberikan akurasi terbaik sebesar 93.91%, dengan precision mencapai 94.63% dan F1-score 83.06%, yang menunjukkan kemampuan model dalam mendeteksi individu yang berisiko depresi secara lebih akurat dibandingkan dengan Decision Tree C4.5. Faktor dominan yang berkontribusi terhadap prediksi depresi meliputi Academic Pressure, Age, dan Job Satisfaction. Penelitian ini diharapkan dapat membantu dalam mendeteksi depresi lebih dini dan memberikan wawasan baru untuk pengembangan kebijakan pencegahan masalah kesehatan mental.

KATA KUNCI — Mental Health, XGBoost, Grid Search, Decision Tree C4.5, CRISP-DM

I. PENDAHULUAN

1.1 Latar Belakang

Kesehatan mental sangat penting dalam kehidupan karena berpengaruh langsung terhadap kualitas hidup, produktivitas, dan kesejahteraan sosial. Salah satu masalah kesehatan mental yang sering terjadi adalah depresi, yang semakin menjadi perhatian di seluruh dunia karena dampaknya yang besar terhadap individu dan masyarakat. Depresi bisa mengganggu berbagai aspek kehidupan, seperti hubungan sosial, kemampuan bekerja, dan bahkan menurunkan kualitas hidup secara keseluruhan. Di Indonesia, gangguan mental seperti depresi cukup banyak terjadi, namun masih banyak orang yang merasa enggan untuk membicarakannya. Berdasarkan data dari Riskesdas, angka gangguan mental di Indonesia terus meningkat setiap tahunnya, dengan kelompok usia produktif seperti remaja dan dewasa muda menjadi yang paling rentan mengalaminya[1].

Dengan berkembangnya teknologi dan semakin pesatnya kemajuan dalam bidang data science serta kecerdasan buatan (AI), berbagai pendekatan baru kini mulai diterapkan untuk mendeteksi dan menganalisis masalah kesehatan mental. Salah satunya adalah pemanfaatan teknik machine learning untuk membantu memprediksi kemungkinan seseorang mengalami depresi berdasarkan berbagai faktor yang mempengaruhinya, seperti kondisi sosial, ekonomi, dan psikologis. Dalam konteks ini, metode seperti XGBoost dan Decision Tree C4.5 Style menawarkan solusi yang menarik. XGBoost, yang dikenal karena kemampuannya dalam menangani masalah klasifikasi dengan akurasi tinggi dan kecepatan pelatihan yang baik, dapat membantu membedakan individu yang berisiko mengalami depresi dari mereka yang tidak, berdasarkan pola data yang ada. Sedangkan Decision Tree C4.5 Style, dengan kemampuannya untuk membangun model yang mudah diinterpretasikan, dapat memberikan gambaran yang jelas mengenai faktor-faktor apa saja yang paling berpengaruh dalam memprediksi depresi pada individu.

Untuk mengoptimalkan kinerja model dalam memprediksi depresi, teknik Grid Search sering digunakan untuk melakukan hyperparameter tuning. Grid Search memungkinkan pencarian kombinasi parameter terbaik untuk model XGBoost atau Decision Tree C4.5, dengan cara menguji setiap kemungkinan parameter secara sistematis dan mengukur kinerja model untuk setiap kombinasi tersebut. Penggunaan Cross Validation dalam Grid Search juga membantu memastikan bahwa model yang dipilih tidak hanya bekerja baik pada data pelatihan, tetapi juga pada data yang belum pernah dilihat sebelumnya, sehingga meningkatkan akurasi prediksi secara keseluruhan. Metode ini sangat berguna dalam memastikan model dapat mengatasi masalah overfitting dan menghasilkan prediksi yang lebih tepat dalam aplikasi nyata di bidang kesehatan mental[2].

Proyek ini akan mengadopsi dataset dari kompetisi Kaggle - Exploring Mental Health Data, yang berisi informasi tentang faktor-faktor sosial, ekonomi, dan psikologis yang dapat mempengaruhi kesehatan mental seseorang. Melalui eksplorasi data, pengembangan model prediksi, serta evaluasi model, diharapkan dapat ditemukan pola-pola yang signifikan yang dapat membantu dalam mendeteksi depresi lebih dini. Dalam proyek ini, XGBoost dan Decision Tree C4.5 Style akan digunakan untuk membangun model klasifikasi yang bertujuan untuk memprediksi apakah seseorang mengalami depresi atau tidak, serta untuk menganalisis faktor-faktor apa saja yang mempengaruhi kondisi tersebut. Proyek ini bertujuan untuk meningkatkan pemahaman tentang karakteristik dataset serta meningkatkan akurasi dalam mendeteksi depresi dengan memanfaatkan teknik-teknik machine learning yang telah terbukti efektif di berbagai bidang.

Dengan menggunakan kedua metode ini, diharapkan proyek ini tidak hanya menghasilkan model yang efektif dalam mendeteksi depresi, tetapi juga memberikan wawasan baru tentang faktor-faktor yang mempengaruhi kesehatan mental individu, yang dapat digunakan untuk mengembangkan kebijakan pencegahan yang lebih baik dalam konteks kesehatan mental masyarakat.

p-ISSN 2301–4156 | e-ISSN 2460–5719

1.2 Tujuan

Adapun tujuan dari pelaksanaan proyek ini adalah:

1. Memenuhi salah satu syarat dalam penyelesaian mata kuliah yang berkaitan dengan penerapan machine learning dalam analisis kesehatan mental.
2. Membangun model klasifikasi yang dapat memprediksi apakah seseorang mengalami depresi atau tidak menggunakan algoritma machine learning, seperti XGBoost dan Decision Tree C4.5, Grid Search, dengan tujuan untuk membantu dalam mendeteksi masalah kesehatan mental secara lebih efektif.
3. Menganalisis faktor yang berkontribusi terhadap depresi berdasarkan dataset mental health, untuk mengidentifikasi faktor sosial, ekonomi, dan psikologis yang dapat mempengaruhi kesehatan mental seseorang, sehingga dapat memberikan wawasan yang lebih dalam dalam pencegahan dan penanganan depresi.

1.3 Manfaat

Proyek ini memberikan beberapa manfaat penting seperti:

1. Proyek ini dapat membantu mendeteksi depresi lebih dini dengan menggunakan metode machine learning seperti XGBoost, Decision Tree C4.5, dan Grid Search sehingga memungkinkan intervensi yang lebih cepat dan mengurangi dampak jangka panjang dari gangguan mental tersebut.
2. Dengan menganalisis faktor sosial, ekonomi, dan psikologis dalam dataset, proyek ini memberikan pemahaman yang lebih dalam mengenai faktor-faktor yang mempengaruhi kesehatan mental dan bagaimana faktor-faktor tersebut berkontribusi pada kondisi depresi.
3. Hasil dari analisis ini dapat digunakan untuk merancang kebijakan pencegahan yang lebih tepat dalam mengatasi depresi, dengan mengandalkan informasi yang lebih akurat mengenai berbagai faktor yang mempengaruhi kesehatan mental individu.

1.4 Ruang Lingkup

Ruang lingkup dalam proyek yang dikerjakan yaitu:

1. Proyek ini akan menggunakan dataset Exploring Mental Health Data dari Kaggle untuk melakukan eksplorasi data, mempersiapkan data untuk analisis, dan memastikan bahwa dataset yang digunakan berkualitas baik sebelum diproses lebih lanjut.
2. Model prediksi depresi akan dibangun menggunakan XGBoost, Decision Tree C4.5, dan Grid Search yang bertujuan untuk memprediksi apakah seseorang berisiko mengalami depresi berdasarkan data yang telah dianalisis dan diproses.
3. Proyek ini dilaksanakan dengan mengikuti langkah-langkah CRISP-DM (Cross-Industry Standard Process for Data Mining), yang terdiri dari tahap business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Setiap tahap akan dilakukan secara berurutan untuk memastikan proses analisis dan pembuatan model berjalan dengan baik.
4. Proyek ini berfokus pada membangun model klasifikasi untuk memprediksi depresi, serta menganalisis faktor-faktor yang berkontribusi terhadap kondisi tersebut. Analisis ini bertujuan untuk menemukan pola-pola signifikan yang dapat memberikan penyebab depresi.

1.5 Istilah dan Singkatan

Singkatan	Istilah
Riskesdas	Riset Kesehatan Dasar
AI	Artificial Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
XGBoost	Extreme Gradient Boosting

II. STUDI LITERATURE

2.1 Mental Health

Kesehatan mental merupakan aspek penting dari kesejahteraan individu, yang mencakup kondisi emosional, mental, dan sosial yang mempengaruhi bagaimana seseorang berpikir, merasa, dan bertindak. Kesehatan mental pada remaja sangat dipengaruhi oleh berbagai faktor, termasuk proses belajar, jenis kelamin, tingkat pendidikan[3]. Pada masa remaja, individu seringkali menghadapi tekanan dan tantangan, yang jika tidak dikelola dengan baik, dapat menyebabkan gangguan mental seperti kecemasan dan depresi. Oleh karena itu, penting untuk memperhatikan kesehatan mental di usia remaja guna mencegah dampak negatif di kemudian hari.

Kesehatan mental di tempat kerja juga menjadi fokus penting dalam beberapa penelitian. Stres kerja merupakan faktor utama yang mempengaruhi kesehatan mental karyawan, terutama di sektor manufaktur yang seringkali menuntut kinerja tinggi dalam waktu yang singkat. Stres kerja yang tinggi, ditambah dengan kurangnya keseimbangan antara kehidupan pribadi dan pekerjaan, dapat menyebabkan gangguan mental seperti depresi dan kecemasan[4]. Oleh karena itu, penting bagi perusahaan untuk menciptakan lingkungan kerja yang mendukung kesehatan mental karyawan agar dapat meningkatkan produktivitas dan kesejahteraan mereka.

Kesehatan mental memegang peranan yang sangat penting dalam kehidupan sehari-hari. Baik di masa remaja, maupun di tempat kerja, kondisi mental seseorang sangat mempengaruhi cara berpikir, merasakan, dan bertindak. Faktor-faktor seperti, stres, serta pengaruh lingkungan dan teknologi berperan besar dalam membentuk kesehatan mental. Oleh karena itu, penting bagi kita untuk memberikan perhatian lebih terhadap kesehatan mental, menciptakan lingkungan yang mendukung, serta menerapkan langkah-langkah yang dapat membantu menjaga kesejahteraan psikologis. dengan begitu, kita dapat hidup lebih produktif, lebih bahagia, dan memiliki hubungan sosial yang lebih baik.

2.2 XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning yang digunakan untuk klasifikasi dan prediksi dengan menggabungkan beberapa model prediksi untuk meningkatkan akurasi. Algoritma ini memanfaatkan teknik

gradient boosting, di mana model baru memperbaiki kesalahan model sebelumnya[5]. XGBoost sangat efisien dalam menangani data besar dan kompleks berkat penggunaan multi-threading dan pruning, serta dilengkapi dengan fitur regularization untuk mencegah overfitting.

XGBoost efektif dalam mengolah data yang tidak terstruktur dan fitur yang saling berhubungan, menjadikannya populer dalam berbagai aplikasi seperti prediksi harga, deteksi anomali, dan klasifikasi gambar. Algoritma ini dapat menangani berbagai jenis data, termasuk data dengan outlier, missing values, dan distribusi yang berubah, sehingga tetap menghasilkan model yang akurat meskipun data memiliki tantangan atau kualitas rendah.

2.3 Grid Search

Grid Search adalah metode yang digunakan untuk menemukan kombinasi hyperparameter terbaik dalam machine learning. Metode ini bekerja dengan cara menguji setiap kombinasi parameter yang telah ditentukan dan memilih kombinasi yang memberikan hasil terbaik berdasarkan kinerja model. Grid Search sering digunakan dalam proses hyperparameter tuning untuk mengoptimalkan kinerja algoritma machine learning. Keunggulan dari Grid Search adalah kemampuannya untuk mengeksplorasi berbagai kombinasi hyperparameter secara sistematis, memastikan bahwa model yang dihasilkan memiliki performa yang optimal.

Menurut Nugraha dan Sasongko dalam penelitiannya, Gridsearch digunakan untuk mengoptimalkan hyperparameter pada tujuh algoritma klasifikasi machine learning, termasuk XGBoost, SVM, dan Decision Tree. Metode ini terbukti efektif dalam mencari parameter terbaik yang dapat meningkatkan akurasi model, seperti yang ditunjukkan dalam penelitian yang menggunakan dataset diabetes Pima Indian. Hasil eksperimen menunjukkan bahwa model XGBoost memperoleh skor tertinggi dengan nilai 0,772, sementara Decision Tree memperoleh nilai terendah[2]. Penggunaan Cross Validation dalam Grid Search membantu dalam mengevaluasi setiap kombinasi secara otomatis, tetapi proses ini tetap membutuhkan waktu yang lama, terutama ketika parameter yang diuji meningkat secara eksponensial[2].

2.4 Decision Tree C4.5.

Decision Tree C4.5 adalah salah satu algoritma klasifikasi dalam data mining yang dikembangkan oleh J. Ross Quinlan sebagai penyempurnaan dari algoritma ID3. Algoritma ini menggunakan pendekatan information gain ratio dalam memilih atribut terbaik untuk pemisahan data, serta mampu menangani atribut numerik, missing values, dan melakukan proses pruning untuk mengurangi kompleksitas pohon. C4.5 dikenal luas karena kemampuannya menghasilkan model klasifikasi yang mudah dipahami dan interpretatif. Algoritma C4.5 digunakan untuk mengklasifikasikan risiko penyakit diabetes berdasarkan gejala-gejala pasien. Penelitian ini menunjukkan bahwa atribut polydipsia memiliki gain ratio tertinggi dan berperan sebagai akar pohon keputusan. Hasil klasifikasi yang diperoleh menunjukkan akurasi sebesar 90,38%, membuktikan bahwa algoritma ini efektif untuk deteksi dini penyakit diabetes[6].

Penelitian sebelumnya menunjukkan bahwa C4.5 memiliki performa yang lebih baik dibandingkan dengan algoritma lain, seperti ID3, dalam hal fleksibilitas dan ketepatan hasil. Algoritma ini dapat menangani data yang kompleks, termasuk data dengan atribut numerik, data yang hilang, serta mampu menghasilkan model yang tidak terlalu rumit melalui *pruning*. C4.5 juga telah diterapkan dalam berbagai bidang seperti evaluasi akademik, prediksi penyakit, dan analisis data bisnis, karena kemampuannya menghasilkan model klasifikasi yang tidak hanya akurat tetapi juga mudah dipahami dan diinterpretasikan.

2.5 CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) adalah metode standar untuk membantu menjalankan proyek data mining agar lebih terstruktur dan efisien. CRISP-DM bisa digunakan di berbagai bidang dan teknologi, karena bersifat fleksibel. Proses CRISP-DM dibagi menjadi enam fase utama: Business Understanding (memahami tujuan bisnis), Data Understanding (memahami data yang ada), Data Preparation (mempersiapkan data untuk analisis), Modeling (membangun model analisis), Evaluation (menilai hasil model), dan Deployment (menggunakan model yang sudah dibangun). Setiap fase memiliki langkah-langkah yang jelas namun tetap bisa disesuaikan dengan kondisi dan kebutuhan proyek[7].

Metode ini sangat berguna untuk membantu dalam merencanakan, berkomunikasi, dan mendokumentasikan proses data mining. Pada fase pertama, Business Understanding, kita mulai dengan memahami tujuan bisnis dan mengubahnya menjadi masalah yang bisa dipecahkan dengan data mining. Setelah itu, pada fase Data Understanding, kita mengeksplorasi dan memahami data untuk menemukan pola atau masalah yang ada. Di fase Data Preparation, data mentah disiapkan agar siap digunakan dalam pemodelan. Pada fase Modeling, kita memilih dan menerapkan teknik analisis yang sesuai. Fase Evaluation digunakan untuk menilai apakah model yang dibuat sesuai dengan tujuan bisnis, dan akhirnya pada fase Deployment, model yang sudah jadi digunakan untuk keperluan praktis. CRISP-DM memberikan panduan yang jelas dan sistematis, baik untuk pemula maupun yang sudah berpengalaman dalam menjalankan proyek data mining.

III. METODE PENELITIAN

Proyek ini menggunakan model *Cross Industry Standard Process for Data Mining* (CRISP-DM) dalam pelaksanaannya untuk mencapai tujuan yang telah ditetapkan. CRISP-DM merupakan model yang dirancang untuk memandu dalam setiap tahap proyek data mining. Model ini mencakup beberapa fase, tugas umum, serta tugas spesifik yang memungkinkan pengguna untuk merencanakan dan mendokumentasikan proyek secara efektif. Fase utama dalam CRISP-DM meliputi pemahaman bisnis (*business understanding*), pemahaman data *(data understanding*), persiapan data (*data preparation*), pemodelan (*modelling*), evaluasi (*evaluation*), dan penerapan (*deployment*).

3.1 Business Understanding

Tujuan utama dari analisis ini adalah untuk membantu memahami dan mendeteksi depresi pada individu dengan menggunakan data kesehatan mental yang tersedia. Dengan membangun model prediksi berbasis data, proyek ini bertujuan untuk mengidentifikasi siapa saja yang berisiko mengalami depresi dan apa saja faktor yang mempengaruhinya.

Informasi yang didapat dari model ini akan sangat berguna bagi lembaga kesehatan seperti psikiater dan rumah sakit dalam merencanakan perawatan yang tepat bagi individu yang berisiko depresi. Dengan prediksi yang lebih akurat, mereka bisa

p-ISSN 2301–4156 | e-ISSN 2460–5719

mengalokasikan sumber daya seperti tenaga medis dan dukungan psikologis sesuai dengan kebutuhan di lapangan.

Beberapa manfaat yang diharapkan dari proyek ini adalah:

1. Dengan mengetahui siapa saja yang berisiko mengalami depresi, lembaga kesehatan bisa memberikan perhatian lebih awal kepada mereka yang membutuhkan bantuan, sehingga mencegah masalah menjadi lebih serius.
2. Model prediksi dapat membantu rumah sakit atau klinik dalam merencanakan tenaga medis dan layanan psikologi yang dibutuhkan berdasarkan tingkat risiko depresi yang ditemukan di masyarakat.
3. Dengan mengetahui faktor-faktor yang mempengaruhi depresi, perawatan bisa lebih fokus dan disesuaikan dengan kebutuhan individu.
4. Dengan mendeteksi depresi lebih awal dan memberikan perawatan yang lebih baik, proyek ini dapat membantu mengurangi biaya pengobatan jangka panjang dan mengurangi dampak negatif depresi pada individu dan masyarakat.

Secara keseluruhan, dengan pendekatan berbasis data ini, diharapkan bisa membantu menangani masalah kesehatan mental lebih efektif dan efisien, serta memberikan manfaat positif bagi masyarakat.

3.2 Data Understanding

Data Understanding membantu dalam mengidentifikasi karakteristik, pola, dan kualitas data yang akan diolah. Tanpa pemahaman yang baik, proses data mining bisa menghasilkan interpretasi yang salah dan keputusan yang tidak tepat. Berikut adalah beberapa hal yang dilakukan pada tahapan ini dalam proyek *Exploring Mental Health*:

a. Collecting Data

Mengumpulkan data yang diperlukan untuk analisis, termasuk informasi terkait faktor sosial, ekonomi, dan psikologis yang mempengaruhi kondisi mental individu. Dataset yang digunakan dalam proyek ini berasal dari kompetisi Kaggle *Exploring Mental Health Data*, yang terdiri dari dua bagian utama: train.csv untuk pelatihan model dan test.csv untuk pengujian model. Data ini mencakup informasi penting seperti jenis kelamin, usia, status pekerjaan, tingkat pendidikan, tekanan sosial, dan banyak lagi yang relevan dengan kesehatan mental, khususnya depresi.

b. Describe Data

Melakukan eksplorasi awal untuk memahami distribusi data, serta variasi pada fitur-fitur yang relevan dengan kondisi mental individu. Beberapa fitur yang dianalisis meliputi:

- **Data Kategorikal**: Gender, City, Degree, Working Professional or Student Profession, Sleep Duration, Dietary Habits, Have you ever had suicidal thoughts?, Family History of Mental Illness, yang dapat mempengaruhi kondisi mental seseorang.
- **Data Numerik**: Age, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Work/Study Hours, Financial Stress, Depression, yang memberikan gambaran terkait gaya hidup dan faktor-faktor yang dapat mempengaruhi stres dan kesehatan mental.

Pada tahap ini, juga dilakukan pemeriksaan awal terhadap statistik deskriptif. Berdasarkan hasil statistik deskriptif, dataset train memiliki 140.700 baris data, sementara test memiliki 93.800 baris. Kedua dataset menunjukkan mayoritas nilai yang sama pada kolom Name ("Rohan") dan Gender ("Male"). Rata-rata usia pada dataset train sekitar 40 tahun, sementara test sedikit lebih tinggi, 41 tahun, dengan rentang usia yang sama antara 18 hingga 60 tahun. Kolom City didominasi oleh "Kalyan", dan fitur Working Professional or Student Profession sebagian besar berisi kategori "Working Professional". Rata-rata Academic Pressure di train adalah 3.14 dan di test 3.16, menunjukkan tekanan akademik moderat di kedua dataset. Meskipun ada perbedaan kecil, distribusi data antara train dan test serupa, sehingga model yang dilatih di train diharapkan berfungsi baik pada tes.

Dataset ini terdiri dari 140.700 baris data dengan 17 kolom, yang mencakup berbagai fitur terkait kesehatan mental. Tipe data pada kolom seperti Gender, Profession, Sleep Duration, Degree, City, Dietary Habits, Family History of Mental Illness, dan Have you ever had suicidal thoughts? menggunakan object (data kategorikal), sementara kolom seperti Age, Work Pressure, Academic Pressure, Financial Stress, Work/Study Hours, CGPA, Study Satisfaction, dan Job Satisfaction menggunakan float64 (data numerik). Kolom Depression menggunakan int64. Semua kolom memiliki data lengkap tanpa nilai kosong, dengan total ukuran memori 18.2 MB, siap untuk dianalisis lebih lanjut.

Dalam tahapan ini, ada juga ditunjukkan hubungan dari setiap fitur-fitur yang ada dalam dataset. Terdapat beberapa hubungan penting antara fitur-fitur dalam dataset. Korelasi negatif yang kuat ditemukan antara Age dan Depression (-0.56), yang menunjukkan bahwa semakin tua usia seseorang, semakin rendah tingkat depresi yang dialami. Selain itu, Academic Pressure memiliki korelasi positif yang moderat dengan Depression (0.48), yang berarti semakin tinggi tekanan akademik, semakin tinggi kemungkinan seseorang mengalami depresi. Work Pressure dan Depression menunjukkan korelasi yang lebih lemah (0.22), sementara Job Satisfaction dan Study Satisfaction memiliki hubungan negatif dengan Depression (-0.17 dan -0.29), yang menunjukkan bahwa kepuasan dalam pekerjaan dan studi dapat mengurangi tingkat depresi. Secara keseluruhan, faktor-faktor seperti tekanan akademik dan kepuasan

kerja/studi memiliki pengaruh signifikan terhadap tingkat depresi.

c. Validation Data
Memastikan data yang dikumpulkan valid dan berkualitas agar tidak mempengaruhi akurasi model prediksi. Beberapa langkah yang dilakukan untuk memvalidasi data adalah:

- Pemeriksaan missing values, yakni memeriksa apakah ada data yang hilang pada kolom atau baris yang memerlukan penanganan, seperti pengisian, penghapusan, atau pengolahan lebih lanjut. Dari hasil, yang memiliki missing values adalah kolom Profession, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Dietary Habits, Degree, Financial Stress.
- Pemeriksaan duplikasi, yakni mengecek apakah ada data yang terduplikasi, jika ada dapat mempengaruhi hasil analisis. Namun, pada dataset ini tidak ada data yang duplikat, baik dari data train maupun test.
- Pemeriksaan outlier dilakukan untuk mencari nilai ekstrim pada kedua jenis fitur, untuk kategorikal, misalnya, pada fitur Age, meskipun distribusi usia relatif merata, terdapat beberapa puncak di usia 50-an, yang bisa menjadi indikasi adanya data yang tidak biasa. Pada Academic Pressure dan Work Pressure, distribusi data tersegmentasi dengan banyak nilai yang terfokus pada angka tertentu, seperti 3, yang bisa menunjukkan adanya nilai ekstrim atau outlier di luar kategori umum tersebut. Begitu juga dengan CGPA, meskipun nilai CGPA tersebar merata, ada beberapa individu dengan nilai CGPA sangat tinggi yang bisa dianggap sebagai outlier. Sementara itu, distribusi Study Satisfaction dan Job Satisfaction menunjukkan bahwa sebagian besar responden merasa cukup puas, namun nilai-nilai ekstrem juga perlu diperhatikan untuk memastikan tidak ada outlier yang mempengaruhi hasil. Terakhir, Work/Study Hours menunjukkan distribusi yang cukup luas, dengan beberapa jam kerja/studi yang lebih sering terjadi, yang bisa menjadi indikasi adanya outlier pada individu dengan jam kerja/studi yang sangat tinggi atau rendah. Untuk numerikal, tidak ada outlier yang dideteksi

Dengan melakukan langkah-langkah ini, diharapkan data yang digunakan untuk membangun model dapat menghasilkan analisis yang akurat dan efektif dalam memprediksi kondisi mental individu, khususnya depresi.

3.3 Data Preparation

Data Preparation adalah tahap yang sangat penting dalam memastikan bahwa data yang digunakan dalam analisis sudah siap untuk digunakan dalam proses pemodelan. Tanpa persiapan yang baik, kualitas model yang dibangun bisa terpengaruh dan menghasilkan analisis yang tidak akurat. Berikut adalah beberapa langkah yang dilakukan dalam tahapan ini dalam proyek *Exploring Mental Health*:

a. Data Selection
Pada tahap Memilih Data, dilakukan pemilihan data yang relevan dengan tujuan analisis berdasarkan uji Chi-Square dan korelasi antara fitur-fitur numerik dan kategorikal terhadap status Student atau Working Professional. Atribut yang memiliki p-value < 0.05 dianggap signifikan dan dipertahankan untuk dimasukkan ke dalam model prediksi. Berdasarkan hasil uji, atribut yang dipertahankan untuk proses modeling adalah Gender, City, Degree, Profession, Sleep Duration, dan Have you ever had suicidal thoughts?, yang semuanya menunjukkan hubungan signifikan dengan status mahasiswa atau profesional.

b. Data Cleaning
Pada tahap ini yang dilakukan adalah untuk Menghilangkan data yang tidak relevan atau bermasalah agar data yang digunakan berkualitas dan tidak mempengaruhi hasil model. Langkah-langkah yang dilakukan untuk membersihkan data adalah:

- Data yang hilang di kolom numerik diisi dengan median dan di kolom kategorikal dengan modus, yang menjamin data yang hilang diisi dengan nilai yang representatif.
- Memeriksa dan menghapus baris data yang terduplikasi menggunakan fungsi duplicated(). Hasilnya menunjukkan tidak ada data yang terduplikasi, yang memastikan bahwa dataset tidak terdistorsi.
- Memeriksa adanya nilai ekstrem (outliers) yang bisa mempengaruhi hasil analisis. Pemeriksaan dilakukan pada fitur numerik seperti Age, CGPA, dan Work/Study Hours, dan tidak ditemukan outliers yang signifikan.

c. Data Construction
Pada tahap Mengkonstruksi Data, beberapa langkah dilakukan untuk menambah fitur baru dan melakukan transformasi data agar lebih relevan dan informatif untuk model prediksi:

- **Normalisasi**
  Beberapa fitur numerik seperti Age, Work/Study Hours, Financial Stress, dan Work Pressure dinormalisasi menggunakan MinMaxScaler agar nilai fitur berada dalam rentang 0 hingga 1. Normalisasi ini penting agar semua fitur memiliki kontribusi yang setara dalam proses pelatihan model, khususnya untuk model yang sensitif terhadap perbedaan skala antar fitur.
- **Feature Engineering**
  Kolom seperti Sleep Duration dan Dietary Habits yang memiliki urutan tingkatannya diubah menjadi format numerik menggunakan mapping logis. Selain itu, dibuat fitur baru yaitu Stress_Score, yang merupakan hasil penjumlahan dari Work Pressure dan Financial Stress, untuk menggambarkan tingkat stres keseluruhan

p-ISSN 2301–4156 | e-ISSN 2460–5719

individu.

d. Labeling Data
Pada tahap Labeling Data, fitur kategorikal dalam dataset dikodekan menjadi format numerik untuk digunakan dalam model. Fitur dengan dua kategori, seperti "Have you ever had suicidal thoughts?", "Family History of Mental Illness", dan "Gender", diubah menggunakan binary encoding, dengan 'No' atau 'Male' menjadi 0 dan 'Yes' atau 'Female' menjadi 1. Untuk fitur nominal dengan lebih dari dua kategori, seperti Degree, Profession, dan City, digunakan category encoding yang mengonversi kategori menjadi angka. Selain itu, Label Encoding diterapkan pada fitur "Working Professional or Student" dengan menggunakan LabelEncoder, yang mengubah 'Student' menjadi 0 dan 'Working Professional' menjadi 1. Penting untuk memastikan bahwa proses encoding dilakukan secara konsisten antara data latih dan data uji agar tidak terjadi kesalahan dalam prediksi model.

e. Data Integration
Menggabungkan data yang diperlukan untuk analisis lebih lanjut. Namun, dalam proyek ini, proses mengintegrasikan data tidak dilakukan karena seluruh data yang dibutuhkan sudah tersedia dalam satu dataset utama, yaitu *Exploring Mental Health*. Dataset ini sudah mencakup semua informasi yang relevan

3.4 Modeling

Pada tahap Modeling, eksperimen klasifikasi dilakukan untuk memprediksi apakah seseorang mengalami depresi berdasarkan dataset yang ada. Proyek ini membandingkan tiga model klasifikasi: C4.5 Decision Tree, XGBoost (Extreme Gradient Boosting), dan XGBoost dengan Tuning GridSearch. Tujuan dari tahap ini adalah untuk mengevaluasi kinerja masing-masing model dalam mendeteksi depresi.

a. Building Testing Scenario
- **Persiapan Data**
  Data yang hilang pada kolom numerik diisi dengan median, sementara kolom kategorikal diisi dengan modus. Fitur kategorikal yang ada seperti Gender, City, dan Profession diubah menjadi format numerik menggunakan LabelEncoder atau OneHotEncoder. Meskipun XGBoost tidak terlalu sensitif terhadap skala data, normalisasi tetap dilakukan untuk membantu konvergensi model. Untuk C4.5, normalisasi tidak diwajibkan, namun tetap disarankan jika ada perbedaan skala yang besar antar fitur. Setelah itu, data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Kolom Depression digunakan sebagai target y, dan kolom Name dihapus karena tidak relevan untuk prediksi.
- **Pemilihan Model**
  C4.5 Decision Tree digunakan untuk membangun model pohon keputusan dengan menggunakan Gain Ratio. Model ini dilengkapi dengan teknik pruning untuk menghindari overfitting, namun bisa mengalami overfitting jika tidak dibatasi kedalamannya. XGBoost, di sisi lain, adalah algoritma ensemble learning berbasis gradient boosting yang menggabungkan banyak pohon keputusan lemah untuk membentuk model yang lebih kuat. XGBoost efektif dalam menangani data besar dan kompleks dan memiliki regularisasi untuk mencegah overfitting.
- **Evaluasi Model**
  Model-model yang dibangun dievaluasi menggunakan metrik akurasi, precision, recall, F1-score, dan confusion matrix untuk membandingkan seberapa baik kedua model mendeteksi depresi.
- **Perbandingan Model**
  Setelah evaluasi, C4.5 dan XGBoost dibandingkan berdasarkan akurasi, precision, recall, F1-score, dan confusion matrix untuk melihat model mana yang lebih efektif dalam mendeteksi depresi.
- **Penyempurnaan Model**
  Penyempurnaan model dilakukan menggunakan GridSearchCV untuk menyesuaikan hyperparameter pada XGBoost. Hyperparameter seperti n_estimators, learning_rate, max_depth, subsample, dan colsample_bytree disesuaikan untuk mengoptimalkan performa model dan mencegah overfitting.

b. Building Model
Pada tahap Building Model, tiga model utama dipertimbangkan untuk memprediksi depresi berdasarkan data yang telah dipersiapkan.

- **C4.5 (Decision Tree)**
  C4.5 adalah algoritma pohon keputusan yang menggunakan Gain Ratio untuk memilih pembagian terbaik dalam data. C4.5 dilengkapi dengan teknik pruning untuk menghindari overfitting, dan cukup mudah dipahami serta diinterpretasi. Model ini diterapkan dengan parameter criterion='entropy' untuk menggunakan information gain dalam membagi data, dan max_depth=8 untuk membatasi kedalaman pohon agar model tidak terlalu kompleks.
- **XGBoost (Extreme Gradient Boosting)**
  XGBoost adalah algoritma ensemble learning berbasis gradient boosting yang menggabungkan banyak pohon keputusan lemah untuk membentuk model yang kuat. XGBoost mendukung optimasi seperti regularisasi, early stopping, dan parallel processing untuk menangani overfitting. Parameter utama yang digunakan dalam model ini adalah n_estimators=1000, learning_rate=0.2, max_depth=4, subsample=0.5, dan colsample_bytree=0.8.
- **XGBoost dengan Tuning GridSearch**
  Untuk meningkatkan performa XGBoost, dilakukan GridSearchCV untuk menemukan kombinasi hyperparameter terbaik. Dalam pencarian 64 kombinasi parameter, hasil terbaik diperoleh dengan n_estimators=200, max_depth=3, min_child_weight=1, subsample=0.7, colsample_bytree=1.0, dan

learning_rate=0.2. Konfigurasi ini menunjukkan bahwa model dengan kedalaman pohon yang relatif dangkal namun memiliki banyak pohon (estimators) dapat memberikan performa yang sangat baik. Model ini mencapai akurasi validasi silang sebesar 94%.

3.5 Evaluation

Pada fase ini, dilakukan evaluasi untuk memastikan model memenuhi tujuan prediksi status kesehatan mental. Beberapa langkah yang dilakukan pada tahap ini adalah sebagai berikut:

a. Pengukuran Performa Model
Model yang telah dilatih dievaluasi menggunakan beberapa metrik kinerja, termasuk akurasi, precision, recall, dan F1-score. Metrik-metrik ini memberikan gambaran tentang kemampuan model dalam memprediksi dengan benar, terutama dalam mendeteksi kedua kelas, yaitu individu yang berisiko mengalami depresi (suicidal) dan yang tidak (non-suicidal).

b. Evaluasi pada Data Uji
Setelah pelatihan, model diuji pada data uji yang terpisah. Prediksi dilakukan menggunakan metode **predict,** yang memberikan hasil prediksi kelas langsung. Kemudian, dilakukan perhitungan accuracy score, pembuatan classification report untuk mendapatkan rincian precision, recall, F1-score, dan confusion matrix untuk menganalisis distribusi kesalahan klasifikasi.

c. Perbandingan Model
Untuk mengetahui model mana yang paling baik, dilakukan perbandingan antara beberapa model yang berbeda, seperti C4.5 (Decision Tree) dan XGBoost. Evaluasi dilakukan terhadap masing-masing model, baik yang sudah melalui tuning hyperparameter (seperti menggunakan GridSearch untuk XGBoost) maupun yang tidak. Hal ini bertujuan untuk melihat pengaruh tuning terhadap peningkatan kinerja model.

d. Analisis Fitur Dominan
Selain mengukur performa model, dilakukan analisis terhadap fitur-fitur yang paling berpengaruh dalam proses prediksi. Fitur-fitur dominan ini membantu untuk memahami faktor-faktor utama yang mempengaruhi keputusan model dalam mengklasifikasikan individu berisiko mengalami depresi.

3.6 Deployment

Pada tahap Deployment, model yang telah dibangun dan diuji perlu diterapkan pada platform yang dapat digunakan untuk prediksi secara langsung. Proses deployment ini dilakukan dengan dua cara: deployment lokal menggunakan Flask dan deployment online menggunakan Streamlit.

a. Local
Pada tahap deployment model secara lokal, langkah pertama adalah menyimpan model yang telah dilatih agar dapat digunakan kembali pada aplikasi web yang akan dijalankan. Model yang telah dibangun, baik itu XGBoost atau model lainnya, disimpan menggunakan pickle untuk mempermudah pemanggilan model yang telah dilatih di aplikasi. Begitu juga dengan scaler yang digunakan untuk normalisasi data input, seperti StandardScaler atau MinMaxScaler.
Setelah model dan scaler disimpan, langkah berikutnya adalah membuat aplikasi Flask untuk menjalankan model yang telah dilatih. Flask adalah framework web ringan untuk membuat aplikasi berbasis web dengan cepat. Di dalam aplikasi Flask, model yang telah disimpan dan scaler akan dimuat untuk melakukan prediksi berdasarkan data yang dimasukkan oleh pengguna.
Flask berguna untuk membangun antarmuka pengguna (frontend) yang menerima input melalui form. Setelah data dimasukkan, aplikasi akan memproses data tersebut, melakukan normalisasi menggunakan scaler yang telah dimuat, dan kemudian mengirim data tersebut ke model untuk memprediksi apakah individu tersebut berisiko mengalami depresi atau tidak. Hasil prediksi akan ditampilkan kepada pengguna di halaman yang sama.
Untuk menjalankan aplikasi secara lokal, perlu memastikan bahwa server Flask berjalan pada mesin lokal yang dapat diakses di localhost pada port 5001. Aplikasi ini dapat diakses oleh pengguna melalui browser dengan URL http://127.0.0.1:5001/. Setelah pengguna mengisi form dengan data yang diperlukan, seperti usia, jam kerja/studi, stres finansial, dan tekanan kerja, model akan melakukan prediksi apakah individu tersebut berisiko mengalami depresi atau tidak. Hasil prediksi tersebut akan ditampilkan pada halaman yang sama, dengan teks yang menunjukkan apakah individu tersebut terdeteksi berisiko depresi atau tidak depresi, berdasarkan input yang diberikan.

b. Online
Untuk tahapan deployment secara online digunakan Streamlit. File model scaler yang sebelumnya sudah dibuat di deployment local juga dibuat pada deployment online. Kemudian, file streamlit_app.py dibuat untuk menjalankan aplikasi di Streamlit, yang berfungsi sebagai antarmuka pengguna untuk melakukan prediksi. Selanjutnya, dibuat file requirements.txt yang berisi daftar pustaka (library) yang diperlukan agar aplikasi berjalan dengan baik pada Streamlit service. Library tersebut akan diinstal otomatis oleh Streamlit sebelum aplikasi dijalankan. Setelah itu, file streamlit_app.py dimodifikasi agar bisa berfungsi dengan layanan Streamlit, termasuk menyesuaikan input dan output model agar sesuai dengan kebutuhan pengguna. Setelah semua file siap, deployment dapat dilakukan dengan menjalankan aplikasi melalui Streamlit Cloud. Aplikasi yang telah dideploy dapat diakses secara online menggunakan URL https://project-data-mining-kel9.streamlit.app/. Hasil prediksi dari model dapat dilihat langsung oleh pengguna, yang menunjukkan bahwa proses deployment berhasil dilakukan dengan memanfaatkan Streamlit sebagai platform untuk hosting aplikasi model secara online.

p-ISSN 2301–4156 | e-ISSN 2460–5719

IV. HASIL DAN PEMBAHASAN

4.1 Hasil dan Pengujian Model

Dalam penelitian ini, kami menggunakan model XGBoost dan Decision Tree C4.5 untuk memprediksi status kesehatan mental, khususnya terkait dengan kecenderungan bunuh diri. Evaluasi dilakukan dengan menggunakan data uji dan metrik-metrik evaluasi yang relevan seperti akurasi, precision, recall, dan F1-score.

a. Evaluasi Model Decision Tree C4.5
Pada model Decision Tree C4.5, hasil evaluasi menunjukkan akurasi sebesar 92.71%. Model ini memiliki precision 92.72% untuk klasifikasi non-suicidal, namun recall untuk suicidal lebih rendah, yaitu 81.53%. Ini menunjukkan bahwa meskipun model ini cukup baik dalam mendeteksi individu yang tidak berisiko, ia lebih kesulitan untuk mendeteksi individu yang berisiko mengalami depresi atau bunuh diri, yang tercermin dalam F1-score yang mencapai 80.38%.

b. Evaluasi Model XGBoost (Tanpa Tuning)
Model XGBoost yang diuji tanpa tuning menunjukkan performa yang sangat baik, dengan akurasi mencapai 93.24%. Precision dan recall-nya berada pada 82.21% dan 80.50%, sementara F1-score adalah 81.34%. Model ini menunjukkan kestabilan yang lebih baik dalam membedakan kedua kelas, meskipun recall untuk kelas suicidal (depresi) sedikit lebih rendah dibandingkan dengan precision.

c. Evaluasi Model XGBoost dengan Grid Search Tuning
Dengan dilakukan tuning GridSearch pada XGBoost, performa model meningkat pesat. Hasil tuning menghasilkan akurasi terbaik sebesar 93.91%, dengan precision dan recall masing-masing mencapai 94.63% dan 81.66%. F1-score meningkat menjadi 83.06%, yang menunjukkan bahwa tuning dengan Grid Search berhasil mengoptimalkan hyperparameter, sehingga menghasilkan model yang lebih stabil dan lebih akurat dalam mendeteksi individu yang berisiko mengalami depresi.

d. Analisis Fitur Dominan dalam Prediksi Suicidal
Selain mengevaluasi performa model, kami juga melakukan analisis terhadap fitur-fitur yang paling berpengaruh dalam prediksi status suicidal. Analisis fitur dominan menunjukkan bahwa Working Professional or Student adalah fitur paling penting dalam memprediksi risiko bunuh diri, dengan Importance Score tertinggi (334.53). Fitur lainnya yang signifikan adalah Have you ever had suicidal thoughts? (56.85), yang menunjukkan bahwa riwayat pikiran suicidal memiliki peran besar dalam prediksi, Age (18.17), yang mencerminkan rentannya remaja dan dewasa muda terhadap tekanan psikologis, serta Academic Pressure (16.78) yang terkait dengan stres yang dialami oleh mahasiswa. Selain itu, Stress Score (15.67) menunjukkan hubungan yang kuat antara tingkat stres dan kecenderungan bunuh diri. Fitur-fitur ini memberikan wawasan bahwa tekanan akademik, pekerjaan, serta faktor psikologis seperti stres dan riwayat pikiran bunuh diri memainkan peran penting dalam memprediksi risiko kesehatan mental dan suicidal.

4.2 Interpretasi

Hasil yang diperoleh dari evaluasi ketiga model menunjukkan bahwa XGBoost dengan GridSearch tuning memberikan performa terbaik dalam memprediksi risiko depresi, khususnya risiko bunuh diri. Model ini mencapai akurasi sebesar 93.90%, dengan precision dan recall yang lebih seimbang serta F1-score tertinggi. Hal ini menunjukkan bahwa XGBoost dengan GridSearch Tuning mampu mendeteksi individu berisiko depresi dengan lebih akurat dan lebih stabil dibandingkan model lainnya. Tuning GridSearch yang dilakukan berhasil meningkatkan kemampuan model dalam mengklasifikasikan kelas suicidal, meskipun recall untuk kelas tersebut sedikit lebih rendah dibandingkan dengan precision.

Sebaliknya, meskipun Decision Tree C4.5 memberikan hasil yang cukup baik dengan akurasi 92.71%, model ini lebih kesulitan dalam mendeteksi individu yang berisiko bunuh diri, yang tercermin dalam recall yang lebih rendah pada kelas suicidal. Ini mengindikasikan bahwa meskipun model ini mampu mengklasifikasikan dengan baik untuk kelas non-suicidal, kemampuan deteksi untuk kelas suicidal perlu ditingkatkan.

XGBoost tanpa tuning sudah menunjukkan performa yang cukup baik dengan akurasi 93.24%, tetapi dengan GridSearch tuning, model ini berhasil dioptimalkan, menghasilkan precision dan recall yang lebih seimbang serta F1-score yang meningkat, yang mencerminkan peningkatan kemampuan model dalam mengklasifikasikan data secara lebih akurat.

Analisis terhadap fitur dominan juga memberikan wawasan penting, di mana Working Professional or Student dan Have you ever had suicidal thoughts? menunjukkan pengaruh terbesar terhadap prediksi risiko bunuh diri. Fitur seperti Age dan Academic Pressure juga memainkan peran penting, yang menunjukkan bahwa faktor-faktor seperti status pekerjaan atau pendidikan, usia, dan tingkat tekanan akademik atau pekerjaan sangat relevan dalam memprediksi masalah kesehatan mental. Hal ini menggarisbawahi pentingnya faktor psikologis dan sosial dalam menentukan potensi risiko depresi dan bunuh diri, yang perlu diperhatikan dalam penerapan model ini di dunia nyata.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berhasil membangun model klasifikasi menggunakan algoritma XGBoost dan Decision Tree C4.5, dengan tambahan Grid Search untuk tuning hyperparameter, yang bertujuan untuk memprediksi apakah seseorang mengalami depresi atau tidak. Model-model yang dibangun diuji dengan data kesehatan mental dan dievaluasi menggunakan metrik-metrik seperti akurasi, precision, recall, dan F1-score. Hasil evaluasi menunjukkan bahwa XGBoost dengan Grid Search Tuning memberikan performa terbaik dengan akurasi 93.91%, precision 94.63%, dan F1-score 83.06%, yang menunjukkan kemampuan model dalam mendeteksi baik individu yang berisiko mengalami depresi maupun yang tidak. Sementara itu, Decision Tree C4.5 meskipun memiliki akurasi yang tinggi, tidak seimbang dalam mendeteksi kelas minoritas (depresi), dengan F1-score yang lebih rendah.

Selain membangun model prediksi, penelitian ini juga berhasil menganalisis faktor-faktor dominan yang berkontribusi

terhadap depresi menggunakan XGBoost. Hasil analisis menunjukkan bahwa Academic Pressure, Have you ever had suicidal thoughts?, Age, Job Satisfaction, merupakan faktor sosial, ekonomi, dan psikologis yang paling berpengaruh dalam memprediksi risiko depresi. Student dan Working Professional yang menghadapi tekanan tinggi dari akademik dan pekerjaan menunjukkan kecenderungan lebih tinggi terhadap masalah kesehatan mental, terutama depresi. Faktor seperti stress score dan riwayat pikiran suicidal juga terbukti memiliki kontribusi signifikan.

Dengan demikian, hasil penelitian ini memberikan wawasan penting tentang faktor-faktor yang mempengaruhi kesehatan mental, serta pentingnya penggunaan model machine learning dalam mendeteksi dan memprediksi depresi. Penelitian ini juga dapat menjadi dasar untuk pengembangan aplikasi atau sistem yang dapat membantu deteksi dini masalah kesehatan mental, sehingga memberikan peluang bagi intervensi lebih cepat dan efektif dalam penanganan depresi.

5.2 Saran

Berdasarkan hasil yang diperoleh dalam penelitian ini, berikut beberapa saran yang dapat dipertimbangkan untuk pengembangan lebih lanjut:

1. Peningkatan Model dengan Data yang Lebih Beragam: Untuk meningkatkan akurasi model, penting untuk menggunakan data yang lebih bervariasi dan mencakup faktor-faktor lain yang mempengaruhi kesehatan mental, seperti faktor lingkungan sosial dan kebiasaan hidup.
2. Pengembangan Aplikasi Berbasis Model: Hasil penelitian ini dapat digunakan sebagai dasar untuk mengembangkan aplikasi yang dapat membantu deteksi dini depresi. Aplikasi ini akan memungkinkan intervensi lebih cepat dan efektif, dengan memberi peringatan pada individu berisiko.
3. Pengujian dengan Model Lain: Meskipun XGBoost dan C4.5 Decision Tree menunjukkan performa yang baik, pengujian dengan model lain seperti Random Forest atau SVM dapat memberikan wawasan lebih dalam mengenai akurasi model prediksi.
4. Peningkatan Keseimbangan Data: Teknik seperti oversampling atau SMOTE bisa digunakan untuk menyeimbangkan kelas dalam dataset, terutama mengingat ketidakseimbangan antara individu yang berisiko dan yang tidak.

VI. PEMBAGIAN PEKERJAAN

Anggota Kelompok	Bagian yang Dikerjakan
Jufourlisa Sirait	Business Understanding, Data Understanding, Data Preparation, Building Model, Deployment, Document, Paper, Poster
Kezia Hutagaol	Business Understanding, Data Understanding, Data Preparation, Building Model, Deployment, Document, Paper, Poster
Reinaldy Hutapea	Business Understanding, Data Understanding, Data Preparation, Building Model, Deployment, Document, Paper, Poster
Tennov Pakpahan	Business Understanding, Data Understanding, Data Preparation, Building Model, Deployment, Document, Paper, Poster

REFERENSI

[1] Riset Kesehatan Dasar (Riskesdas), “Laporan Riskesdas 2018 Nasional.pdf,” *Lembaga Penerbit Balitbangkes*. p. hal 156, 2018. [Online]. Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/3514/1/Laporan Riskesdas 2018 Nasional.pdf

[2] W. Nugraha and A. Sasongko, “Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search Hyperparameter Tuning on Classification Algorithm with Grid Search,” *Sist. J. Sist. Inf.*, vol. 11, no. 2, pp. 391–401, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[3] S. A. Melina and C. K. Herbawani, “Faktor-Faktor yang Mempengaruhi Kesehatan Mental Remaja Selama Pandemi Covid-19 : Tinjauan Literatur,” *Media Kesehat. Masy. Indones.*, vol. 21, no. 4, pp. 286–291, 2022, doi: 10.14710/mkmi.21.4.286-291.

[4] A. Nisa and U. M. Area, “Pengaruh Stres Kerja terhadap Kesehatan Mental pada Karyawan di Perusahaan Multinasional,” pp. 1–8.

[5] R. P. Tresyani, D. W. Utomo, and N. Maldini, “Deteksi Dini Gangguan Kesehatan Mental dengan Model Bert dan Algoritma Xgboost,” vol. 16, no. 01, pp. 93–98, 2025, doi: 10.35970/infotekmesin.v16i1.2535.

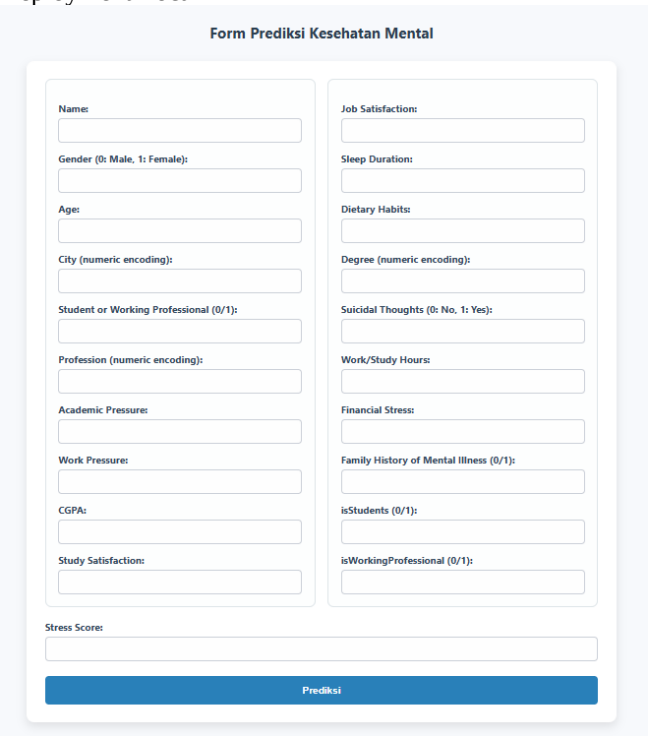
[6] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, “Classification of diabetes disease using decision tree algorithm (C4.5),” *J. Phys. Conf. Ser.*, vol. 1869, no. 1, 2021, doi: 10.1088/1742-6596/1869/1/012082.

[7] R. Wirth and J. Hipp, “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000, [Online]. Available: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

p-ISSN 2301–4156 | e-ISSN 2460–5719

LAMPIRAN

Berikut ini tampilan Form Prediksi Kesehatan Mental untuk Deployment Local



Berikut ini tampilan Form Prediksi Kesehatan Mental untuk Deployment Online