**LAPORAN PROYEK MATA KULIAH  
12S4054 – DATA MINING**

**Exploring Mental Health Data – XGBoost Approach**

****

**Disusun Oleh:**

|  |  |
| --- | --- |
| 12S21012 | Walker Valentinus Simanjuntak |
| 12S21013 | Markus Pardianto Hutagalung |
| 12S21014 | Fritz Kevin Manurung |

**Tautan GitHub** : <https://github.com/walkervalentinuss/Project-Data-Mining>

|  |  |  |
| --- | --- | --- |
| **PROGRAM STUDI SARJANA SISTEM INFORMASI**  **FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**  **INSTITUT TEKNOLOGI DEL**  **2024** | | |
| Nama Dokumen: LP-DAMI-24-04 | Tanggal: 30 November 2024 | Jumlah Halaman: 19 |

**DAFTAR ISI**

BAB I PENDAHULUAN 5

1.1 Latar Belakang 5

1.2 Tujuan 6

1.3 Manfaat 6

1.4 Ruang Lingkup 6

1.5 Istilah dan Singkatan 7

BAB II STUDI LITERATURE 8

2.1 *Classification* 8

2.1.1 Random Forest 8

2.1.2 XGBoost 8

2.1.3 Neural Network 9

2.2 *Mental Health* 9

BAB III METODE 10

3.1 CRISP-DM 10

3.1.1 Business Understanding 10

3.1.2 Data Understanding 11

3.1.3 Data Preparation 19

3.1.4 Modelling 22

3.1.5 Evaluation 29

3.1.6 Deployment 29

3.2 Timeline 30

BAB IV HASIL PENGUJIAN 31

BAB V 32

ANALISIS 32

BAB VI KESIMPULAN 33

BAB VII PEMBAGIAN PEKERJAAN 34

REFERENSI 35

LAMPIRAN 37

**DAFTAR GAMBAR**

[Gambar 1. CRISP-DM Step 10](#_Toc183893996)

[Gambar 2. Correlations 16](#_Toc183893997)

[Gambar 3. Missing Values 17](#_Toc183893998)

[Gambar 4. Persentase Missing Value 17](#_Toc183893999)

[Gambar 5. Sample Data 18](#_Toc183894000)

[*Gambar 6. Imbalance Data* 18](#_Toc183894001)

[Gambar 7. Cross-Validation Students 23](#_Toc183894002)

[Gambar 8. Cross-Validation Working Proffesionals 23](#_Toc183894003)

[Gambar 9. Evaluasi Model Random Forest data Students 24](#_Toc183894004)

[Gambar 10. Evaluasi Model Random Forest Working Proffesional 24](#_Toc183894005)

[Gambar 11. Best Estimators for Students 26](#_Toc183894006)

[Gambar 12. Best Estimators for Working Proffesionals 26](#_Toc183894007)

[Gambar 13. Evaluasi Model XGBoost data Students 26](#_Toc183894008)

[Gambar 14. Evaluasi Model XGBoost data Working Proffesionals 27](#_Toc183894009)

[Gambar 15. Evaluasi Model Neural Network data Students 29](#_Toc183894010)

[Gambar 16. Evaluasi Model Neural Network data Working Proffesionals 29](#_Toc183894011)

[Gambar 17. Timeline 30](#_Toc183894012)

**DAFTAR TABEL**

[Tabel 1. Istilah dan Singkatan 7](#_Toc183894445)

[Tabel 2. Variables 11](#_Toc183894446)

[Tabel 3. Data Preparation 19](#_Toc183894447)

# PENDAHULUAN

## Latar Belakang

Pada tahun 2008, The Mental Health Foundation menyatakan bahwa kesehatan mental didefinisikan oleh cara seseorang berpikir dan cara seseorang merasakan sesuatu tentang diri dan kehidupan mereka, dan juga hal – hal ini memengaruhi cara seseorang dalam melewati masa – masa yang sulit. Kesehatan mental dipandang dapat memberikan dampak yang berbeda- beda kepada seseorang, sehingga dapat mempengaruhi bagaiman cara seseorang mengambil keputusan, hal ini juga akan mempengaruhi kemampuan seseorang untuk memaksimalkan kapasitas yang ada pada dirinya dalam memaksimalkan peluan yang tersedia. Dampak ini juga dapat mempengaruhi cara seseorang berkomunikasi dengan keluarga, tempat kerja, komunitas, dan lingkungan pertemanan. Terdapat hubungan yang erat antara kesehatan fisik dan mental, karena kedua hal ini memengaruhi satu sama lain secara langsung maupun tidak langsung[1].

AI (Artificial Intelligence) muncul sebagai salah satu pendekatan digital dalam dunia kesehatan mental. Pendekatan ini khususnya machine learning, sudah digunakan untuk mengembangkan solusi aplikasi prediksi, deteksi, dan perawatan kesehatan mental dalam bentuk saran berdasarkan data historis pengguna[2]. Salah satu penelitian berjudul “Depression and Suicide Risk Detection on Social Media using fastText Embedding and XGBoost Classifier” membahas upaya mendeteksi konten yang berisiko tinggi terkait depresi dan potensi bunuh diri pada media sosial dengan menggunakan pendekatan NLP (Natural Language Processing) dan pembelajaran mesin. Dalam eksperimen, framework yang diusulkan menggunakan fastText embedding untuk analisis kontekstual, TF-IDF vector untuk relevansi istilah, dan XGBoost sebagai classifier untuk klasifikasi yang akurat. Hasil eksperimen menunjukkan bahwa pendekatan ini mencapai nilai AUC sebesar 0,78 dan weighted F1-score sebesar 0,71 pada dataset Reddit, serta meningkatkan akurasi dan weighted F1-score dibandingkan dengan model baseline seperti SVM, Random Classifier, Logistic Regression, MLP, LSTM, dan RNN. Analisis dan eksperimen yang dilakukan menunjukkan kinerja yang kuat terhadap berbagai model embedding dan classifier dalam menghadapi masalah yang menantang ini[3].

Berdasarkan penelitian sebelumnya, peneliti akan menggunakan algoritma XGBoost untuk mendeteksi apakah seseorang mengalami depresi atau tidak. Dataset yang digunakan dalam proyek ini tidak diambil dari dataset dunia nyata, melainkan dari kompetisi Kaggle Season 4 Episode 11[4]. Deteksi ini akan menjadi langkah awal bagi individu untuk mengidentifikasi apakah mereka mengalami gejala depresi. Hasil deteksi ini diharapkan dapat menjadi dasar bagi individu untuk memperbaiki pola hidup dan menjaga kesehatan mental mereka. Siklus pengerjaan dan pembangunan model pada proyek ini akan menggunakan siklus CRISP-DM, agar sesuai dengan siklus yang ada di dunia industri[5].

## Tujuan

Adapun tujuan dari pembuatan proyek ini adalah:

1. Salah satu syarat penyelesaian mata kuliah Data Mining T.A. 2024/2025
2. Memberikan pemahaman tentang penggunaan CRISP-DM dalam proyek data mining
3. Mendeteksi gejala depresi menggunakan algoritma XGBoost

## Manfaat

Adapun tujuan laporan proyek ini adalah sebagai berikut:

1. Mengembangkan keterampilan dalam machine learning untuk mendeteksi kesehatan mental seseorang
2. Meningkatkan kesadaran mahasiswa tentang pentingnya kesehatan mental
3. Menggunakan hasil prediksi sebagai langkah awal identifikasi gejala depresi

## Ruang Lingkup

Ruang lingkup dari proyek yang kami kerjakan adalah:

1. Siklus pengerjaan mengikuti CRISP-DM yang terdiri dari tahap business understanding, data understanding, data preparation, modeling, evaluation, dan deployment.
2. Dataset yang digunakan untuk melatih model adalah Dataset Kaggle Playground Competition yang terdiri dari 2 file yaitu train.csv dan test.csv dengan total 141.000 baris data latih.

## Istilah dan Singkatan

**Tabel 1. Istilah dan Singkatan**

| **Singkatan** | **Definisi** |
| --- | --- |
| AI |  |
| CRISP-DM |  |
| XGBoost |  |
| SVM |  |
| MLP |  |
| AUC |  |
| F1-Score |  |
| LSTM |  |
| RNN |  |

# STUDI LITERATURE

## *Classification*

Dalam ilmu data dan pembelajaran mesin, klasifikasi adalah proses untuk mengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan fitur atau atributnya. Dalam proses klasifikasi, data diberi label sesuai dengan kelasnya, dan model dilatih untuk mengidentifikasi pola yang membedakan satu kelas dari kelas lainnya. Proses klasifikasi dilakukan melalui algoritma yang mengidentifikasi hubungan antara atribut atau fitur (variabel input) dan label atau kelas (variabel keluaran). Setelah dilatih, model klasifikasi dapat memprediksi atau mengklasifikasikan data baru yang belum pernah dilihat dengan menggunakan pola-pola yang telah dipelajari dari data pelatihan. Dalam klasifikasi, beberapa algoritma populer yang sering digunakan meliputi logistic regression, decision tree, random forest, support vector machine (SVM), dan XGBoost. Pemilihan algoritma yang tepat tergantung pada jenis data, jumlah fitur, dan kompleksitas pola yang ingin dikenali oleh model [6].

### Random Forest

Algoritma Random Forest adalah salah satu metode klasifikasi yang sangat efektif untuk digunakan dalam mendeteksi kesalahan. Algoritma ini mampu bekerja dengan cepat bahkan pada kumpulan data yang besar. Namun, algoritma ini memiliki kelemahan, yaitu cenderung mengalami overfitting ketika digunakan pada data yang tidak seimbang[7].

### XGBoost

XGBoost *(Extreme Gradient Boosting)* adalah algoritma machine learning yang sangat efisien dan populer, terutama dalam kompetisi data science, yang mengimplementasikan teknik boosting untuk menggabungkan beberapa model lemah menjadi satu model yang kuat. Algoritma ini dirancang untuk kecepatan dan efisiensi memori, serta dilengkapi dengan fitur regularisasi yang membantu mengurangi overfitting. XGBoost dapat digunakan untuk berbagai jenis masalah, termasuk klasifikasi dan regresi, dan mampu menangani data yang hilang dengan baik. Keunggulannya dalam memberikan akurasi tinggi dan kemudahan interpretasi menjadikannya pilihan utama bagi banyak praktisi dalam analisis data, termasuk dalam aplikasi kesehatan mental untuk mendiagnosis depresi berdasarkan biomarker [8].

### Neural Network

Neural network merupakan model matematis yang terinspirasi dari cara kerja dan struktur otak manusia. Proses pembelajarannya melibatkan pelatihan model menggunakan dataset tertentu, di mana bobot koneksi antara neuron dioptimalkan untuk meningkatkan akurasi prediksi. Dalam beberapa tahun terakhir, neural network telah memberikan kemajuan besar di berbagai bidang, seperti pengenalan gambar, analisis bahasa alami, dan peramalan data[9].

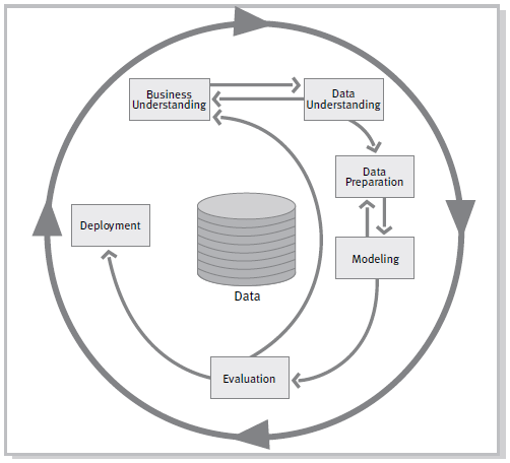
## *Mental Health*

Mental health merupakan keadaan kesehatan psikologis dan emosional seseorang. Ini mencakup bagaimana seseorang berpikir, merasa, dan berperilaku, serta kemampuan untuk mengatasi stres, berinteraksi dengan orang lain, dan membuat keputusan. Kesehatan mental yang baik memungkinkan individu untuk mengenali kemampuan mereka, mengelola stres kehidupan sehari-hari, bekerja secara produktif, dan berkontribusi positif kepada komunitas mereka. Masalah kesehatan mental dapat mencakup berbagai kondisi, seperti depresi, kecemasan, gangguan perilaku, dan lainnya [10].

# METODE

## CRISP-DM

CRISP-DM adalah salah satu siklus pengerjaan proyek dalam Data Mining. Model ini mencakup tahap pengerjaan, apa yang dilakukan pada setiap tahap, dan hasil dari setiap tahap tersebut. Siklus ini dapat dilihat pada Gambar 1.



**Gambar 1. CRISP-DM Step**

Urutan yang ditunjukkan oleh tanda panah pada gambar tidak kaku, tanda panah hanya menunjukkan keterikatan antar tahapan[11]. Metode ini akan kami gunakan dalam pengerjaan proyek Data Mining kami, setiap tahapan pada siklus akan dijelaskan pada sub-bab berikutnya.

### Business Understanding

Tujuan utama proyek ini adalah untuk mengembangkan sebuah aplikasi dunia nyata, khususnya dalam konteks kesehatan mental. Beberapa masalah yang ingin diselesaikan adalah,

1. Depresi adalah masalah kesehatan mental yang signifikan di masyarakat, dan deteksi dini dapat menjadi langkah penting untuk memberikan perawatan yang tepat waktu.
2. Mengingat stigma yang masih ada terkait dengan kesehatan mental, alat deteksi yang efektif dan efisien dapat memberikan dampak positif dalam meningkatkan kesadaran individu agar lebih peka dengan kesehatan mental.
3. Dengan menggunakan algoritma XGBoost untuk memproses data, proyek ini berfokus pada penggunaan data besar (big data) dan kecerdasan buatan (AI) untuk memberikan solusi yang dapat diterapkan di dunia nyata, yaitu dalam bidang kesehatan mental.

### Data Understanding

Tahap Data Understanding berfokus pada eksplorasi, pemeriksaan, dan pengumpulan wawasan awal dari data yang akan digunakan untuk menyelesaikan masalah yang telah didefinisikan. Tujuannya adalah untuk memahami karakteristik data secara menyeluruh dan memastikan bahwa data tersebut sesuai untuk analisis lebih lanjut.

#### **Variables**

Variable yang dimiliki pada dataset ini berjumlah 20 kolom, statistik masing-masing kolom dapat dilihat pada Tabel 2.

**Tabel 2. Variables**

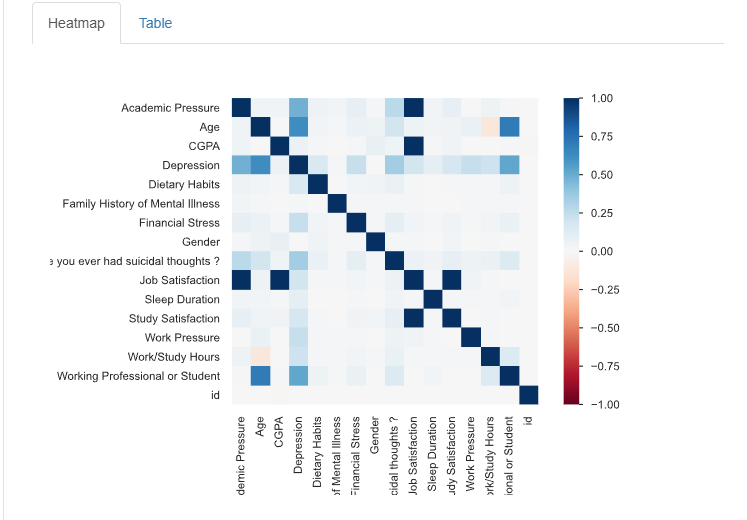
|  |  |  |
| --- | --- | --- |
| **Column** | **Types** | **Statistic** |
| id | Real number (R) |  |
| Name | Text |  |
| Gender | Categorical |  |
| Age | Real number (R) |  |
| City | Text |  |
| Working Professional or Student | Categorical |  |
| Profession | Text |  |
| Academic Pressure | Categorical |  |
| Work Pressure | Categorical |  |
| CGPA | Real number (R) |  |
| Study Satisfaction | Categorical |  |
| Job Satisfaction | Categorical |  |
| Sleep Duration | Categorical |  |
| Dietary Habits | Categorical |  |
| Degree | Text |  |
| Have you ever had suicidal thoughts? | Boolean |  |
| Work/Study Hours | Real number (R) |  |
| Financial Stress | Categorical |  |
| Family History of Mental Illness | Boolean |  |
| Depression | Categorical |  |

Dari keterangan tabel di atas, dapat kita lihat bahwa variabel pada dataset terdiri dari beberapa kolom kategorikal dan kolom numerical. Dari tabel tersebut juga terdapat 2 grup utama yang dapat dianalisa lebih lanjut, yaitu Students dan Working Professional. Data masukan untuk grup Students mungkin saja berbeda dengan data masukan untuk grup Working Professional, sehingga pemisahan dataset untuk kedua grup ini dapat dilakukan. Untuk kolom pada pelatihan model nantinya, kami tidak akan menggunakan variabel seperti Id, Name, City, Profession, dan Degree.

#### **Correlations**

Korelasi adalah ukuran statistik yang menggambarkan sejauh mana dua variabel memiliki hubungan linier satu sama lain. Heatmap pada Gambar memberikan nilai korelasi antar variabel dalam dataset. Korelasi diukur dalam skala -1 hingga 1, dengan interpretasi sebagai berikut:

* Korelasi positif (mendekati 1): Ketika satu variabel meningkat, variabel lain cenderung meningkat.
* Korelasi negatif (mendekati -1): Ketika satu variabel meningkat, variabel lain cenderung menurun.
* Korelasi mendekati 0: Tidak ada hubungan linier yang signifikan antara dua variabel.



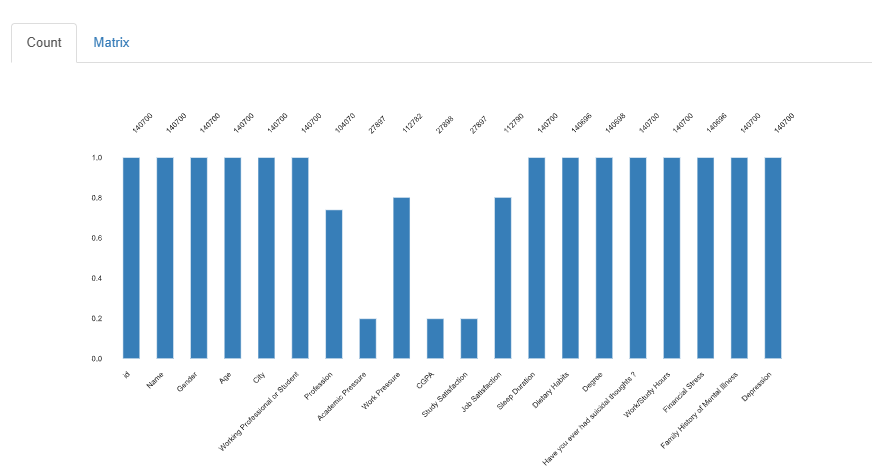
**Gambar 2. Correlations**

#### **Missing Values**

Missing values (nilai yang hilang) adalah data yang tidak tercatat atau kosong dalam dataset, baik karena tidak tersedia, tidak dicatat dengan benar, atau terjadi kesalahan selama pengumpulan atau pengolahan data. Diagram pada Gambar 3. adalah visualisasi missing values pada dataset, di mana setiap kolom dataset direpresentasikan oleh batang vertikal.

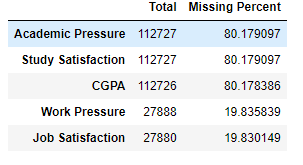
Interpretasi Tinggi Batang:

* Kolom dengan missing values sedikit atau tidak ada: Kolom seperti id, Gender, dan beberapa kolom lain memiliki tinggi batang mendekati 1.0, menunjukkan hampir semua data lengkap.
* Kolom dengan missing values banyak: Kolom seperti Profession, Work Pressure, CCPA, Study Satisfaction, dan Job Satisfaction memiliki batang yang lebih pendek, menunjukkan adanya data yang hilang dalam jumlah signifikan.



**Gambar 3. Missing Values**

Dari visualisasi missing value pada gambar, dapat kita lihat bahwa terdapat banyak data kosong pada variabel Academic Pressure, Work Pressure, CGPA, Study Satisfaction, dan Job Satisfaction. Setelah dihitung, ada beberapa kolom yang memiliki nilai missing value lebih dari 80%.



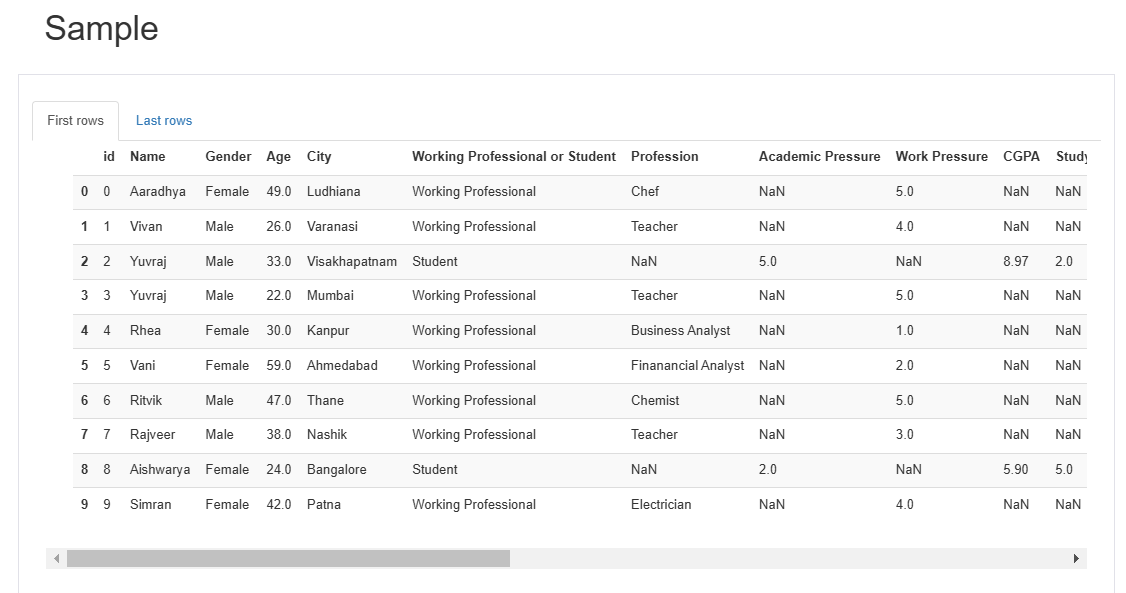
**Gambar 4. Persentase Missing Value**

Tabel di atas menunjukkan bahwa ada pola missing value pada dataset ini. Polanya terdapat pada:

* Setiap baris data dengan nilai Working\_Professional or Student == Student, akan memiliki missing value pada kolom Work Pressure dan Job Satisfaction, hal ini karena setiap mahasiswa pasti belum merasakan apa itu tekanan bekerja dan kepuasan dalam bekerja
* Setiap baris data dengan nilai Working\_Professional or Student == Working Professional, akan memiliki missing value pada kolom Academic Pressure, Study Satisfaction, dan CGPA, hal ini karena setiap pekerja tidak lagi merasakan apa itu tekanan akademis, kepuasan belajar dan nilai indeks prestasi seperti mahasiswa.

#### **Sample Data**

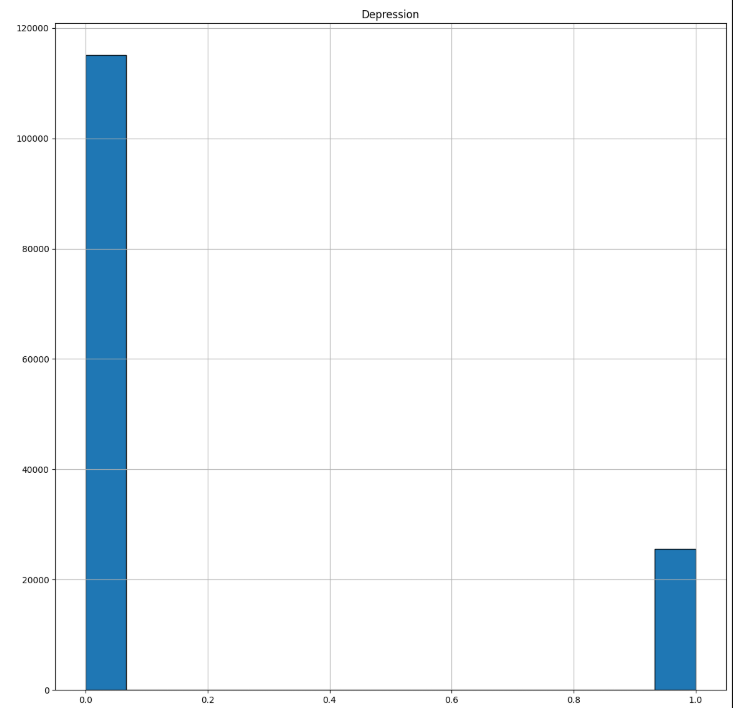
Berikut ini adalah contoh sample data yang ada pada dataset Exploring Mental Health Data:



**Gambar 5. Sample Data**

#### **Imbalance Data**

Terdapat Imbalance Data pada Dataset yang akan digunakan. Dapat kita lihat pada gambar 5, lebih banyak data dengan label 0 (Tidak Depresi) dibandingkan dengan label 1 (Depresi), sehingga diperlukan teknik oversampling atau undersampling agar data yang diolah seimbang.



***Gambar 6. Imbalance Data***

### Data Preparation

Tahap data preparation adalah proses awal dalam analisis data yang meliputi serangkaian langkah untuk mengolah, membersihkan, dan mempersiapkan data agar siap digunakan dalam model analisis, termasuk Machine Learning. Proses ini penting dalam memastikan bahwa data yang digunakan berkualitas tinggi dan relevan untuk tujuan analisis. Berikut ini tahapan Data Preparation yang kami lakukan dalam mengolah dataset:

**Tabel 3. Data Preparation**

| **No.** | **Kode** | **Deskripsi** |
| --- | --- | --- |
| 1. |  | Kode ini bertujuan untuk mengidentifikasi dan menyimpan kolom-kolom kategorikal yang memiliki lebih dari 30 kemunculan untuk dianalisis lebih lanjut. Outputnya akan menghasilkan daftar kolom relevan yang dipilih, berisi nama kolom seperti "Gender", "Sleep Duration", dan "Dietary Habits" yang memiliki cukup data untuk dianalisis. |
| 2. |  | Pada tahap ini, dilakukan perhitungan total dan persentase nilai yang hilang dalam setiap kolom pada dataset. Ini dilakukan karena nilai yang hilang dapat mempengaruhi hasil analisis dan model yang akan digunakan. Dengan memahami sejauh mana nilai hilang dalam dataset, dapat membantu membuat keputusan yang tepat mengenai cara menangani kolom-kolom yang memiliki banyak nilai hilang, apakah melalui pengisian, penghapusan, atau strategi lain. Dari output yang dihasilkan tabel yang menunjukkan total dan persentase nilai hilang di setiap kolom, memungkinkan kita untuk mengidentifikasi kolom yang berpotensi bermasalah, misalnya "Academic Pressure" dengan persentase tinggi nilai hilang. |
| 3. |  | Di tahap ini, dataset dipisahkan menjadi dua subset yaitu "Mahasiswa" dan "Profesional". Ini memungkinkan untuk membandingkan langsung antara dua kelompok yang dipisahkan. |
| 4. |  | Pada langkah ini, kolom-kolom tidak relevan atau yang terlalu banyak nilai hilang akan dihapus dari dataset “Student”. Untuk menangani nilai hilang dengan fokus pada kolom yang masih ingin dianalisis. |
| 5. |  | Dalam tahap ini, diterapkan teknik imputasi untuk mengisi nilai yang hilang pada kolom-kolom numerik dengan menggunakan “SimpleImputer”. Metode ini dapat menggantikan nilai hilang dengan rata-rata, median, atau modus dari data yang ada. Ini dilakukan untuk menjaga integritas data dan memastikan bahwa setiap observasi memiliki semua fitur yang diperlukan untuk analisis, agar menghindari penghapusan data yang mungkin masih digunakan. Output akhir yang dihasilkan adalah dataset mahasiswa yang lengkap, dengan nilai hilang pada kolom numerik telah diisi dengan rata-rata, sehingga semua variabel siap untuk digunakan dalam model analisis. |
| 6. |  | Dalam tahap ini, dilakukan perhitungan total jumlah nilai yang hilang dan persentasenya dalam dataset "Working Professional". Dapat membantu mengidentifikasi kolom-kolom yang memiliki nilai hilang signifikan, seperti "Work Pressure" dan "Job Satisfaction". Dengan pemahaman ini, kita dapat membuat keputusan tentang bagaimana menangani nilai hilang tersebut. Output akhirnya akan menampilkan tabel dengan kolom jumlah total nilai hilang dan persentasenya. |
| 7. |  | Pada kode ini dilakukan pemisahan dataset menjadi dua bagian: fitur (X) yang terdiri dari variabel independen yang akan digunakan untuk memprediksi, dan label (y) yang merupakan hasil atau variabel target yang ingin dianalisis. |
| 8. |  | Pada langkah ini, akan mengubah variabel kategorikal menjadi format numerik. Contohnya seperti variabel "Job Title" dan "Gender" akan dikodekan menjadi angka, yang memudahkan model dalam memproses data. |
| 9. |  | Di tahap ini, akan dilakukan standardisasi fitur-fitur dalam dataset. Standardisasi membantu dalam mengoptimalkan model dengan menyesuaikan fitur agar memiliki mean 0 dan deviasi standar 1, yang seringkali meningkatkan kinerja model secara keseluruhan. |
| 10. |  | Dalam langkah ini, akan menangani masalah ketidakseimbangan kelas dengan menggunakan SMOTE (Synthetic Minority Over-sampling Technique). Dengan metode ini dapat menduplikasi atau menciptakan contoh baru dari kelas minoritas (misalnya, kondisi depresi) yang membantu meningkatkan kinerja model di kelas-kelas yang kurang terwakili. Pada output pertama oleh “y\_students” menunjukkan ada 16,319 siswa yang mengalami depresi dan 11,552 yang tidak. Ketidakseimbangan terlihat di sini, dengan kelas depresif lebih dominan. Dan setelah menghitung jumlah kategori dalam “y\_working” menunjukkan ketidakseimbangan yang lebih ekstrem, dengan 9,227 profesional yang mengalami depresi dibandingkan 103,496 yang tidak. Setelah menerapkan SMOTE, baik untuk dataset mahasiswa maupun profesional, hasil akhir akan berupa dua dataset yang lebih seimbang. |
| 11. |  | Selanjutnya akan dibagi dataset menjadi set train dan set test dengan rasio 80:20. Ini bertujuan untuk mengevaluasi kinerja model secara objektif setelah dilakukan training model, sehingga dapat mengetahui seberapa baik model dapat menggeneralisasi pada data baru. |

### Modelling

Pada tahap modelling kami menggunakan 3 model dalam mengerjakan proyek ini, yaitu 2 model sebagai pembanding dan 1 sebagai model utama. Dengan Random Forest dan Neural Network sebagai model pembanding dan XGBoost sebagai model utama.

#### **Random Forest**

Pada proyek ini, algoritma *Random Forest* digunakan untuk melakukan klasifikasi pada dua kelompok dataset: *students* dan *working professionals*. Algoritma ini dipilih karena kemampuannya yang baik dalam menangani data beragam dan meminimalkan risiko *overfitting*.

Langkah-langkah implementasi meliputi:

1. **Inisialisasi Model**

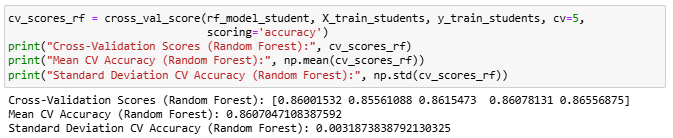
Model diinisialisasi dengan 100 trees untuk memastikan keseimbangan antara akurasi dan waktu pemrosesan. Parameter random\_state diatur untuk menjaga hasil yang konsisten.

|  |
| --- |
| **For Students:**  rf\_model\_student = RandomForestClassifier(n\_estimators=100, random\_state=42)  **For Working Proffesionals:**  rf\_model\_working = RandomForestClassifier(n\_estimators=100, random\_state=42) |

1. **Evaluasi Cross-Validation**

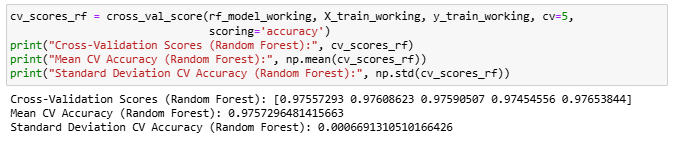
Model dievaluasi menggunakan metode *k-fold cross-validation* dengan k=5. Proses ini memberikan gambaran yang lebih general terhadap performa model pada data yang belum pernah dilihat. Hasil rata-rata akurasi CV untuk dataset *students* dan *working professionals* masing-masing adalah:

* + Dataset *Students*: Dengan mean **[0.8607047108387592]** dan standar deviasi **[0.0031873838792130325]**.



**Gambar 7. Cross-Validation Students**

* + Dataset *Working Professionals*: Dengan mean **[0.9757296481415663]** dan standar deviasi**[0.0006691310510166426]**.



**Gambar 8. Cross-Validation Working Proffesionals**

1. **Training Model**

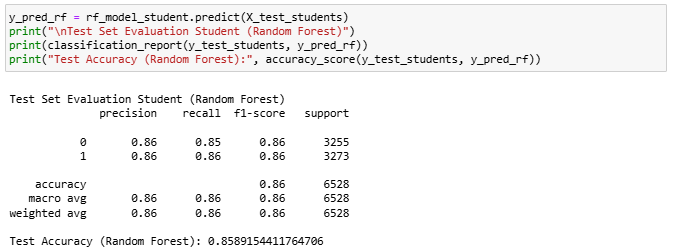
Sebelum melakukan evaluasi data nya akan dilatih terlebih dahulu.

|  |
| --- |
| **For Students:**  rf\_model\_student.fit(X\_train\_students, y\_train\_students)  **For Working Profesionals:**  rf\_model\_working.fit(X\_train\_working, y\_train\_working) |

1. **Evaluasi Model**

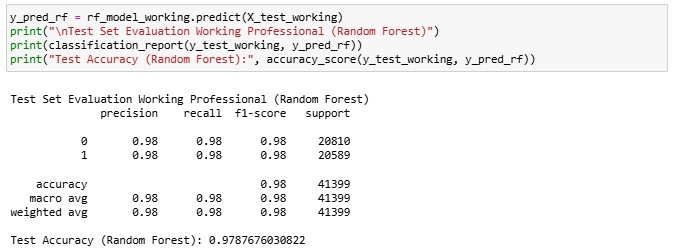
Setelah model dilatih performanya akan dievaluasi pada data uji menggunakan metrik akurasi, *precision*, *recall*, dan *f1-score*. Akurasi pada data uji adalah sebagai berikut:

* + Dataset *Students*: **[0.8589154411764706]**

****

**Gambar 9. Evaluasi Model Random Forest data Students**

* + Dataset *Working Professionals*: **[0.9787676030822]**

****

**Gambar 10. Evaluasi Model Random Forest Working Proffesional**

#### **XGBoost**

Pada proyek ini, algoritma XGBoost digunakan dipilih karena performanya yang unggul pada data yang memiliki korelasi fitur yang kompleks.

Langkah-langkah implementasi meliputi:

1. **Optimasi Parameter dengan Grid Search**

Ruang parameter meliputi jumlah pohon keputusan (n\_estimators), kedalaman maksimum pohon (max\_depth), kecepatan pembelajaran (learning\_rate), dan parameter lainnya. Proses Grid Search menggunakan 3-lipatan cross-validation untuk mengidentifikasi parameter terbaik berdasarkan akurasi.

|  |
| --- |
| **## Definisi Ruang Parameter untuk *Grid Search***  param\_grid = {  'n\_estimators': [50, 100, 200],  'max\_depth': [3, 5, 7],  'learning\_rate': [0.01, 0.1, 0.2],  'subsample': [0.8, 1.0],  'colsample\_bytree': [0.8, 1.0]  } |

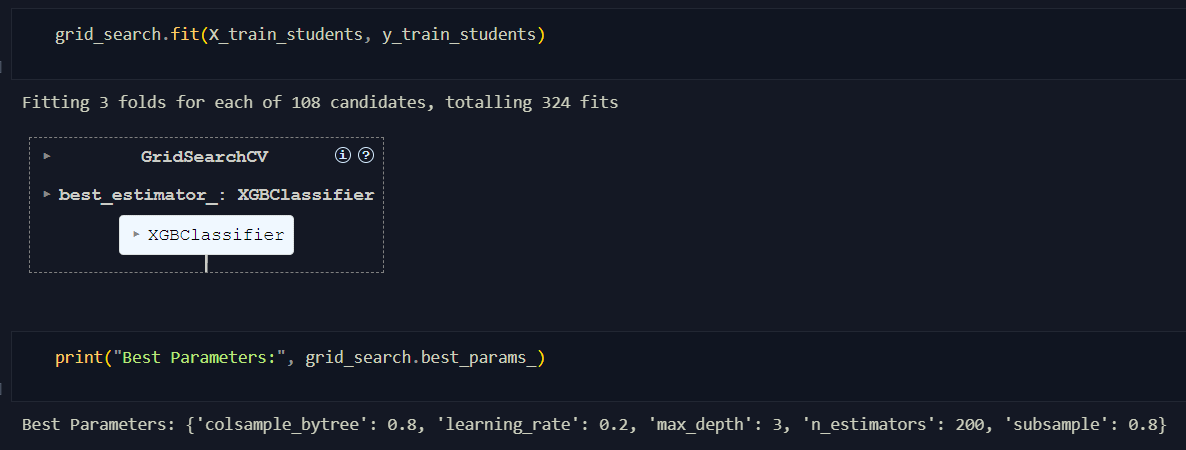
|  |
| --- |
| **## Inisialisasi Model XGBoost**  xgb\_model\_student = XGBClassifier(random\_state=42, eval\_metric='logloss') |

|  |
| --- |
| ## **Proses *Grid Search Cross-Validation***  grid\_search = GridSearchCV(estimator=xgb\_model\_student,  param\_grid=param\_grid,  cv=3, # 3-Fold Cross Validation  scoring='accuracy',  verbose=1,  n\_jobs=-1) |

1. **Hasil Optimasi**

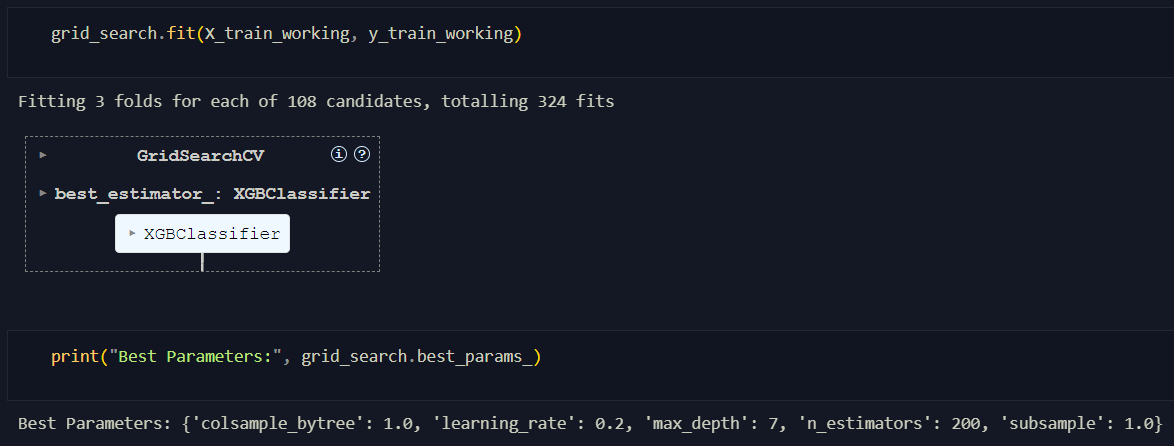
Kombinasi parameter terbaik untuk masing-masing dataset adalah:

* Dataset Students: **Best Parameters: {'colsample\_bytree': 1.0, 'learning\_rate': 0.2, 'max\_depth': 3, 'n\_estimators': 200, 'subsample': 1.0}**



**Gambar 11. Best Estimators for Students**

* Dataset Working Professionals: **Best Parameters: {'colsample\_bytree': 0.8, 'learning\_rate': 0.2, 'max\_depth': 5, 'n\_estimators': 200, 'subsample': 0.8}**

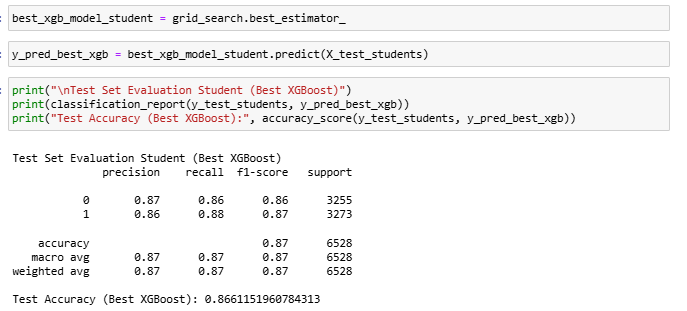


**Gambar 12. Best Estimators for Working Proffesionals**

1. **Evaluasi Model pada Data Uji**

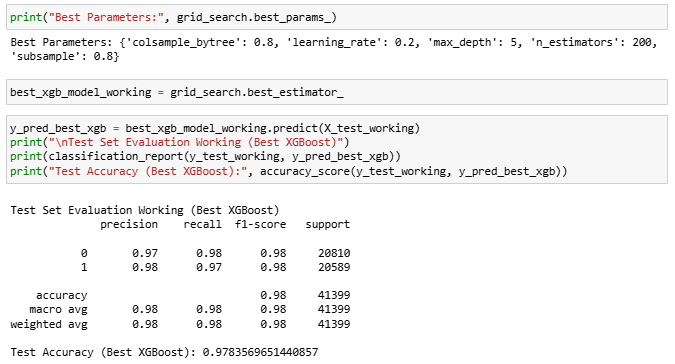
Model terbaik diuji pada data uji menggunakan metrik akurasi, precision, recall, dan f1-score. Hasil evaluasi adalah sebagai berikut:

* Dataset Students: Akurasi **[0.8661151960784313]**



**Gambar 13. Evaluasi Model XGBoost data Students**

* Dataset Working Professionals: Akurasi **[0.9783569651440857]**

****

**Gambar 14. Evaluasi Model XGBoost data Working Proffesionals**

#### **Neural Network**

Pada penelitian ini, Neural Network (NN) digunakan untuk klasifikasi biner pada dua kelompok dataset, yaitu *students* dan *working professionals*. Model NN dirancang untuk mengoptimalkan performa melalui lapisan tersembunyi yang efektif dan teknik regulasi.

Langkah-langkah implementasi meliputi:

1. **Arsitektur Model**

Model terdiri dari dua lapisan tersembunyi dengan jumlah neuron berturut-turut 64 dan 32, menggunakan fungsi aktivasi ReLU. *Dropout* dengan nilai 0.3 diterapkan pada kedua lapisan untuk mencegah *overfitting*. Lapisan keluaran menggunakan fungsi aktivasi sigmoid untuk menghasilkan probabilitas klasifikasi.

|  |
| --- |
| ## **Arsitektur Neural Network**  **--Students--**  model\_student = Sequential([  Dense(64, activation='relu', input\_shape=(X\_train\_students.shape[1],)),  Dropout(0.3),  Dense(32, activation='relu'),  Dropout(0.3),  Dense(1, activation='sigmoid') # Output layer for binary classification  ])  **--Working Proffesionals—**  model\_working = Sequential([  Dense(64, activation='relu', input\_shape=(X\_train\_working.shape[1],)),  Dropout(0.3),  Dense(32, activation='relu'),  Dropout(0.3),  Dense(1, activation='sigmoid') # Output layer for binary classification  ]) |

1. **Kompilasi dan Pelatihan Model**

Model dikompilasi menggunakan optimizer Adam, fungsi *loss* Binary Cross-Entropy, dan metrik akurasi. Proses pelatihan menggunakan data validasi sebesar 20% dari data latih dengan *Early Stopping* untuk menghentikan pelatihan ketika tidak ada perbaikan pada *validation loss* selama 5 epoch.

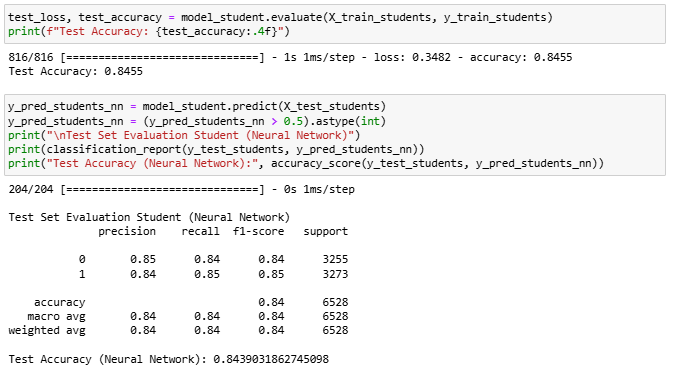
|  |
| --- |
| ## **For Students**  model\_student.compile(optimizer='adam', loss='binary\_crossentropy', metrics=['accuracy'])  early\_stop = EarlyStopping(monitor='val\_loss', patience=5, restore\_best\_weights=True)  history\_student = model\_student.fit(  X\_train\_students, y\_train\_students,  epochs=100,  batch\_size=32,  validation\_split=0.2,  callbacks=[early\_stop],  verbose=1  ) |

|  |
| --- |
| ## **For Working Proffesionals**  model\_working.compile(optimizer='adam', loss='binary\_crossentropy', metrics=['accuracy'])  early\_stop = EarlyStopping(monitor='val\_loss', patience=5, restore\_best\_weights=True)  history\_working = model\_working.fit(  X\_train\_working, y\_train\_working,  epochs=100,  batch\_size=32,  validation\_split=0.2,  callbacks=[early\_stop],  verbose=1  ) |

1. **Hasil Evaluasi**

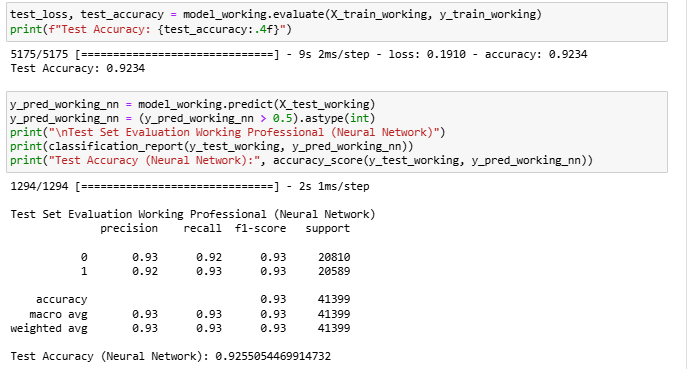
Performa model pada data uji dievaluasi menggunakan akurasi dan metrik lainnya. Hasil akurasi untuk masing-masing dataset adalah:

* + Dataset *Students*: **[0.8439031862745098]**



**Gambar 15. Evaluasi Model Neural Network data Students**

* + Dataset *Working Professionals*: **[0.9255054469914732]**



**Gambar 16. Evaluasi Model Neural Network data Working Proffesionals**

### Evaluation

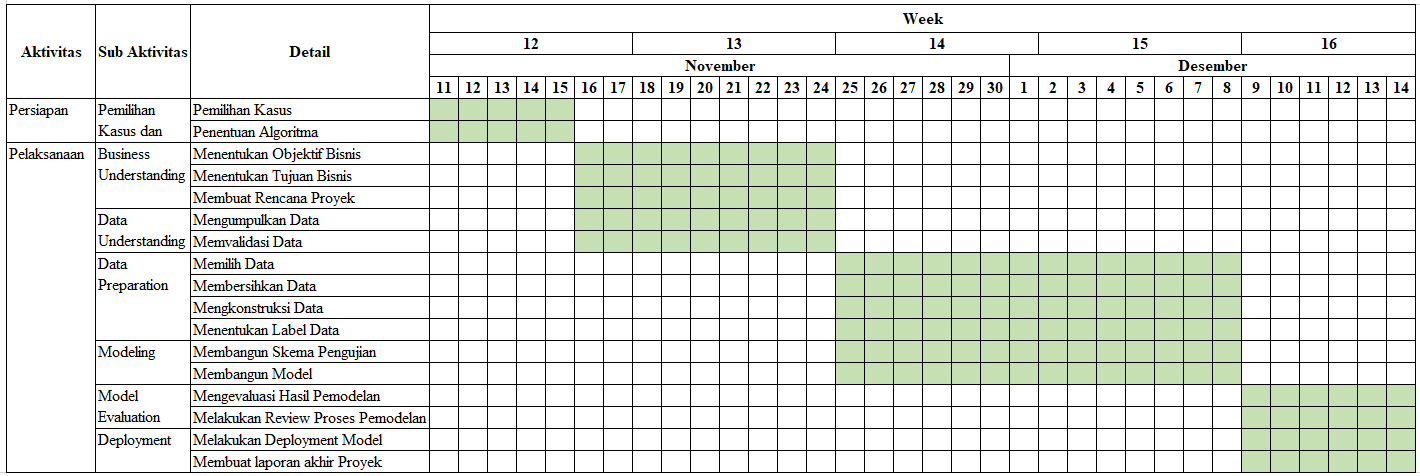
eval

### Deployment

Deploy

## Timeline

Berikut ini adalah Timeline dari pengerjaan proyek yang akan kami lakukan sampai dengan selesai:



**Gambar 17. Timeline**

# HASIL PENGUJIAN

# 

# ANALISIS

# KESIMPULAN

# PEMBAGIAN PEKERJAAN

|  |  |  |
| --- | --- | --- |
| 1 | Walker Valentinus Simanjuntak |  |
| 2 | Markus Pardianto Hutagalung |  |
| 3 | Fritz Kevin Manurung |  |

# REFERENSI

[1] D. Bhugra, A. Till, and N. Sartorius, “What is mental health?,” *Int. J. Soc. Psychiatry*, vol. 59, no. 1, pp. 3–4, 2013, doi: 10.1177/0020764012463315.

[2] S. D’Alfonso, “AI in mental health,” *Curr. Opin. Psychol.*, vol. 36, pp. 112–117, 2020, doi: 10.1016/j.copsyc.2020.04.005.

[3] S. Ghosal and A. Jain, “Depression and Suicide Risk Detection on Social Media using fastText Embedding and XGBoost Classifier,” *Procedia Comput. Sci.*, vol. 218, pp. 1631–1639, 2022, doi: 10.1016/j.procs.2023.01.141.

[4] Kaggle, “Exploring Mental Health Data,” 2024. [Online]. Available: https://www.kaggle.com/competitions/playground-series-s4e11/data?select=train.csv.

[5] N. Hotz, “What is CRISP DM?,” *Data Science Process Alliance*, 2024. https://www.datascience-pm.com/crisp-dm-2/.

[6] R. A. M. Pratama, K. I. Rachmadiansyah, and S. Sidharta, “Technique of Mental Health Issues Classification based on Machine Learning: Systematic Literature Review,” *Procedia Comput. Sci.*, vol. 227, pp. 137–146, 2023, doi: 10.1016/j.procs.2023.10.511.

[7] E. R. B. Sebayang, Y. H. Chrisnanto, and Melina, “Klasifikasi Data Kesehatan Mental di Industri Teknologi Menggunakan Algoritma Random Forest,” *IJESPG J.*, vol. 1, no. 3, pp. 237–253, 2023.

[8] A. Sharma and W. J. M. I. Verbeke, “Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081),” *Front. Big Data*, vol. 3, no. April, pp. 1–11, 2020, doi: 10.3389/fdata.2020.00015.

[9] Cahya Alkahfi, “Neural Network untuk Klasifikasi Multikelas dengan Keras,” *sainsdata.id*, 2024. https://sainsdata.id/machine-learning/2751/neural-network-untuk-klasifikasi-multikelas-dengan-keras/.

[10] M. Rezapour and L. Hansen, “A machine learning analysis of COVID-19 mental health data,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-19314-1.

[11] 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.

# LAMPIRAN