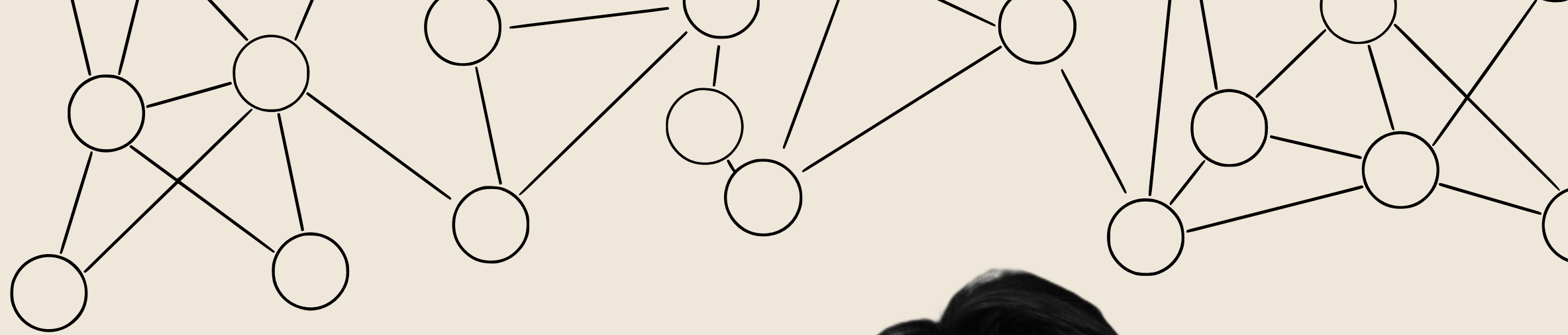
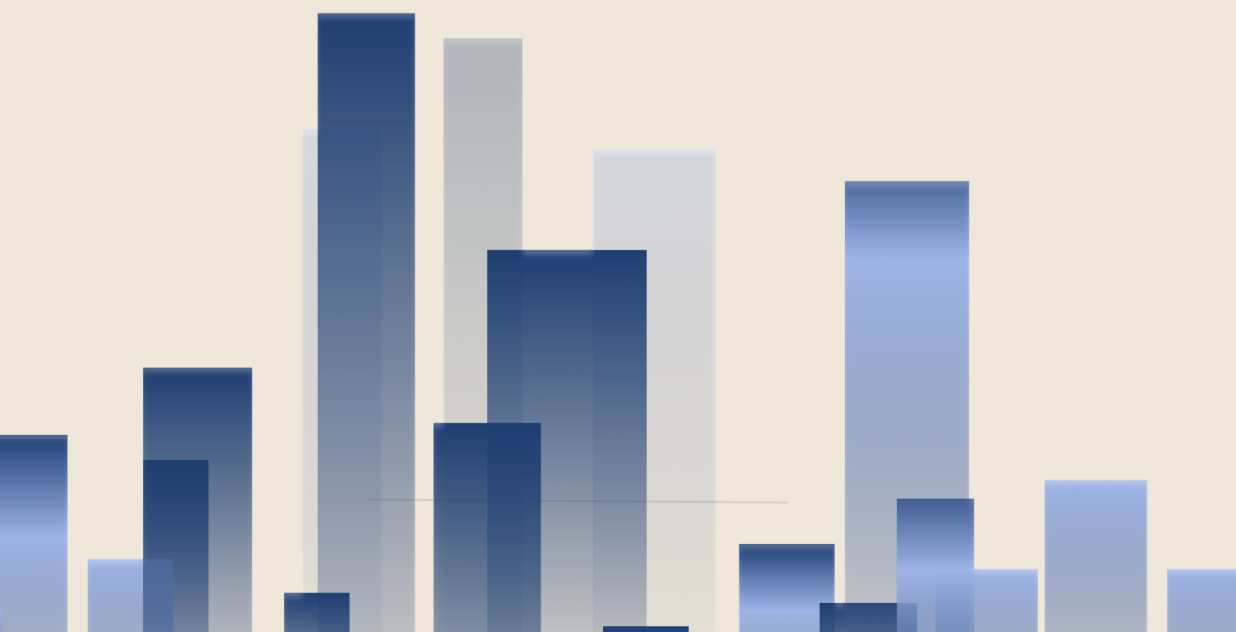




Portfolio

Statistician | Data Scientist

Afghan Al Malik Efendi

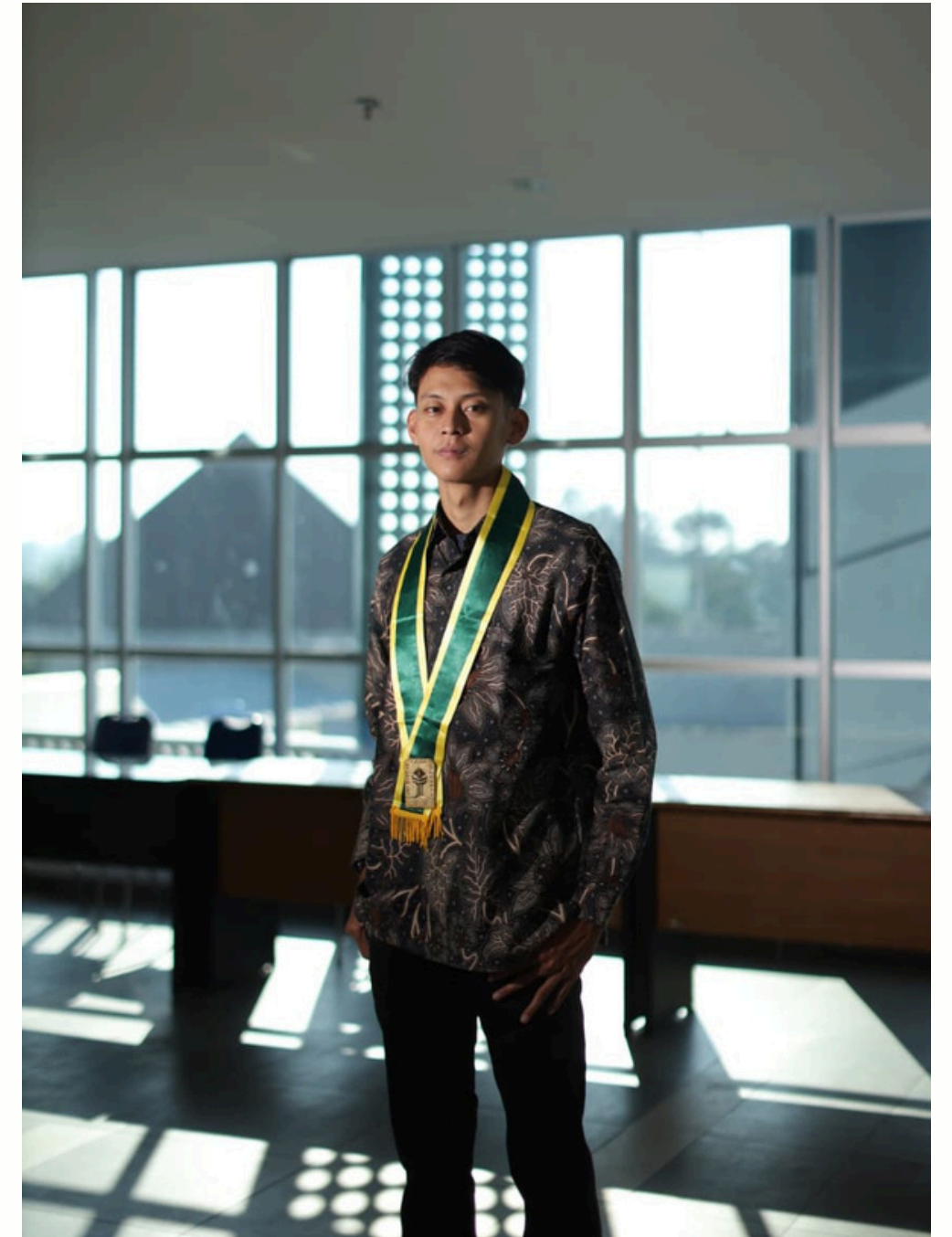


[Download CV](#)

Halo!

Saya **Afghan Al Malik Efendi**, lulusan Sarjana **Statistika** dari Universitas Islam Indonesia dengan peminatan **Data Science** dan **Machine Learning**, serta tersertifikasi **BNSP** sebagai **Data Analyst**.

Sebagai seorang **Data Enthusiast**, saya selalu mengikuti perkembangan teknologi dan terus mengembangkan kemampuan di bidang **analisis statistik, machine learning, data science**, juga **Artificial Intelligence**. Saya tertarik memahami bagaimana data dapat menjawab berbagai permasalahan dan mendukung pengambilan keputusan yang lebih baik melalui analisis yang **akurat** dan **relevan**.

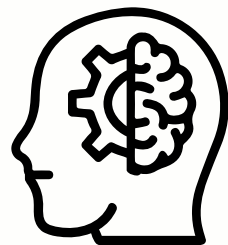


Kemampuan



Keahlian Teknis:

- Analisis Data & Statistik
- *Data Science & Analysis*
- *Data Engineering & Modelling*
- *Machine Learning*
- *Business Intelligence*
- Sistem Informasi Geografis
- Pemanfaatan *Artificial Intelligence*
- *Programming*



Soft Skills:

- *Analytical & Critical Thinking*
- *Problem Solving*
- *Attention to Detail*
- *Curiosity & Continuous Learning*

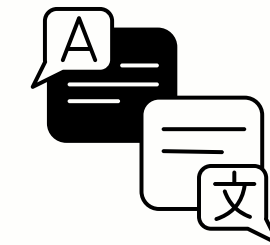
Tools & Software:

- R
- Python
- SQL
- Tableau, Looker Studio, Software visualisasi data lainnya
- Microsoft Excel, Google Sheet
- QGIS
- SPSS, SEM PLS, Software analisis statistik lainnya



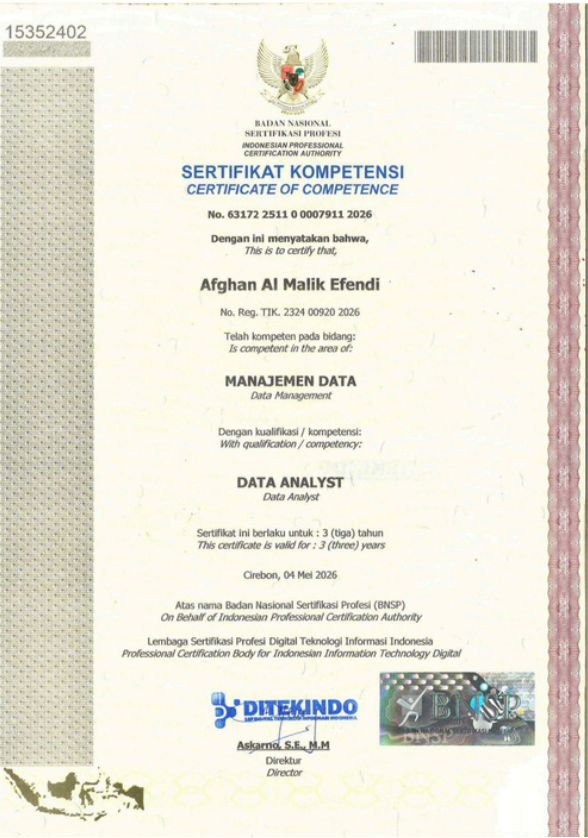
Bahasa:

- Indonesia (Native)
- Inggris (Proficient)



Sertifikasi

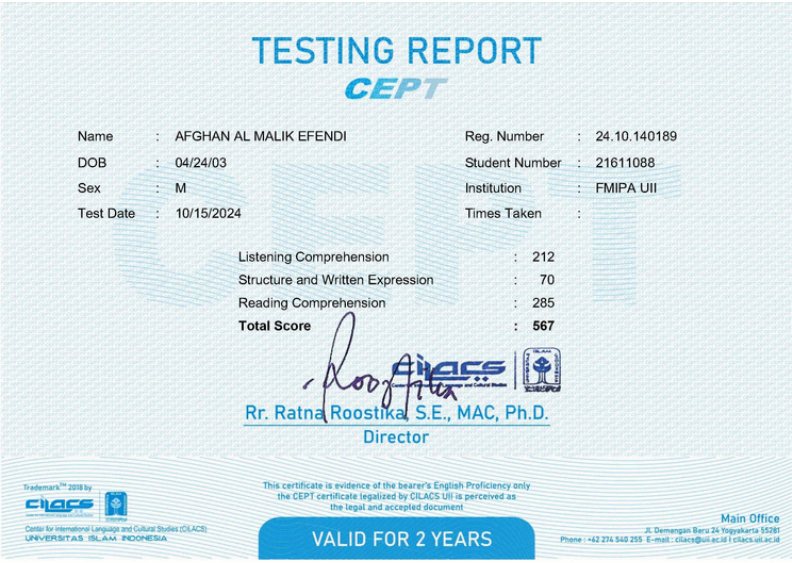
Data Analyst
Badan Nasional Sertifikasi
Profesi (BNSP)
Mei 2026 - Mei 2029



Certified International Specialist
Data Visualization (CISDV)
PASAS Institute
Februari 2024 - Februari 2027



Certificate of English Proficiency
Test
Cilacs UII
Oktober 2024 - Oktober 2027



Pengalaman



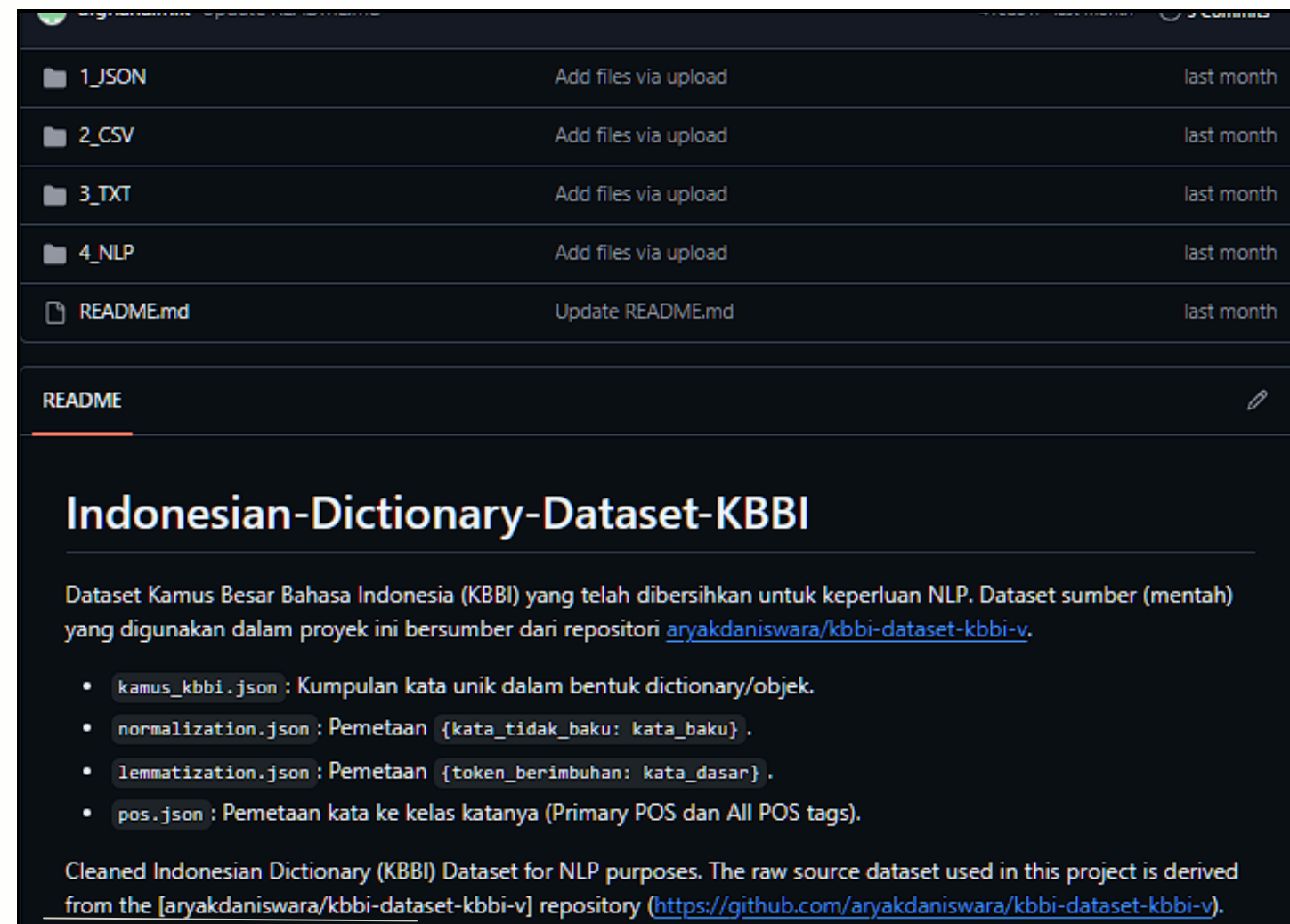
Magang - Divisi Statistik Produksi Badan Pusat Statistik Prov. Jawa Barat Januari 2024 – Februari 2024

Selama menjalani program magang di Badan Pusat Statistik (BPS) Provinsi Jawa Barat, terlibat dalam **pengembangan aplikasi menggunakan software R untuk menghitung Concentration Ratio** sebagai indikator tingkat konsentrasi pangsa pasar. Selain itu, berkontribusi dalam **pengolahan, analisis, dan visualisasi data industri untuk publikasi Indikator Industri Manufaktur Indonesia 2021**, serta mengolah data sensus untuk mendukung publikasi Sensus Pertanian 2023.



Cleaned Indonesian Dictionary Dataset for NLP

Menyusun dataset KBBI yang telah dibersihkan dan distrukturkan untuk mendukung normalisasi, lemmatisasi, part-of-speech tagging, dan pengembangan aplikasi NLP bahasa Indonesia.



Problem

Dataset kamus mentah memerlukan pembersihan dan restrukturisasi sebelum dapat digunakan dalam pipeline NLP. Kosakata, bentuk tidak baku, kata berimbuhan, dan kelas kata harus disajikan dalam format konsisten agar mudah digunakan oleh peneliti maupun developer.

Action

Membersihkan dataset sumber, mengorganisasi kosakata unik, serta membangun mapping untuk kata tidak baku, kata dasar, dan kelas kata. Hasilnya disediakan dalam beberapa format dan didokumentasikan melalui repository GitHub agar mudah digunakan kembali.

Result

Project menghasilkan resource KBBI siap pakai yang mencakup kamus kata unik, normalisasi kata tidak baku, lemmatisasi token berimbuhan, dan part-of-speech. Struktur multi-format meningkatkan kemudahan integrasi untuk preprocessing dan pengembangan NLP bahasa Indonesia.

Tools

- *Python*

Skills

- text normalization
- lemmatization

<https://github.com/afghanalmlk/Indonesian-Dictionary-Dataset-KBBI>

Tuberculosis Risk Factor Analysis

Menganalisis faktor risiko tuberkulosis di Jawa Barat menggunakan data Survei Kesehatan Indonesia 2023, penanganan missing data, dan regresi logistik berbobot desain survei kompleks.

ANALISIS FAKTOR RISIKO TUBERKULOSIS
MENGUNAKAN REGRESI LOGISTIK BERBASIS
DESAIN SURVEI KOMPLEKS DENGAN
PENANGANAN DATA HILANG
(Studi Kasus: Data SKI 2023 di Provinsi Jawa Barat)

TUGAS AKHIR
Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program
Studi Statistika



Disusun Oleh:
Afghan Al Malik Efendi
21611088

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2026

ABSTRACT

*Risk Factor Analysis of Tuberculosis Using Complex Survey Design Logistic Regression with Missing Data Handling
(Case Study: 2023 Indonesian Health Survey Data in West Java Province)*

Afghan Al Malik Efendi
Department of Statistics, Faculty of Mathematics and Natural Sciences
Universitas Islam Indonesia

Tuberculosis (TB) is a serious global health problem, with Indonesia ranking second highest in the world and a steadily increasing incidence trend, particularly in West Java Province. This study aims to analyze the risk factors of TB in West Java by applying complex survey design logistic regression with missing data handling, using data from the 2023 Indonesian Health Survey (SKI 2023). The main methodological challenges include missing data and the complex survey structure, including stratification, clustering, and weighting. The initial stage of the study involved evaluating three imputation methods: Simple Imputation, Bayesian Ridge, and Random Forest. Validation results showed that Random Forest performed best (MAPE = 9.2%, average F1-Score = 0.45) and was selected to handle missing data. Subsequently, survey-weighted logistic regression modeling showed that the complex survey design logistic regression model had good calibration (Archer-Lemeshow test, $p = 0.927$) and adequate discrimination ability ($AUC = 0.736$). Significant risk factors included a history of household contact with TB patients ($OR = 15.22$), comorbid diabetes mellitus ($OR = 4.85$), preference for hospital care ($OR = 5.90$), low education level ($OR = 3.28$), underweight nutritional status ($OR = 1.89$), and lack of physical activity ($OR = 1.63$). Conversely, informal employment or unemployment and younger age (<15 years) were protective factors. Using complex survey design logistic regression resulted in shifts in coefficient values due to sample weighting and survey structure, as well as increased standard errors caused by intra-cluster correlation, ensuring estimates better reflect the population. Ignoring survey design and weighting in complex survey data may lead to inaccurate estimates and inferences. These findings can provide a valid basis for designing targeted interventions to reduce TB cases in West Java Province.

Keywords: Tuberculosis, Indonesian Health Survey (SKI 2023), Missing data handling, Logistic regression, Complex survey design, Risk factors.

Problem

Analisis faktor risiko tuberkulosis menghadapi dua tantangan utama, yaitu missing data dan struktur survei kompleks yang mencakup stratifikasi, clustering, serta pembobotan. Model konvensional berisiko menghasilkan estimasi dan inferensi yang tidak merepresentasikan populasi Jawa Barat secara tepat.

Action

Membandingkan Simple Imputation, Bayesian Ridge, dan Random Forest untuk menangani missing data. Metode terbaik kemudian digunakan sebelum membangun regresi logistik berbobot survei, mengevaluasi kalibrasi dan diskriminasi model, serta menginterpretasikan odds ratio faktor risiko.

Result

Random Forest dipilih sebagai metode imputasi terbaik dengan MAPE 9,2%. Model survei menunjukkan kalibrasi yang baik dengan p-value 0,927 dan AUC 0,736. Faktor risiko terkuat adalah kontak serumah dengan pasien TB, dengan odds ratio 15,22.

Tools

- R
- Python

Skills

- Statistical Analysis
- Logistic Regression
- Machine Learning
- Imputation

SiDaGi — Web App Manajemen Dapur

Mengembangkan web application untuk menata operasional dapur, mengintegrasikan data, approval, pengadaan, anggaran, dan laporan dalam satu alur kerja berbasis peran.



Problem

Operasional dapur masih menggunakan data yang tidak terstruktur dan komunikasi melalui WhatsApp. Kondisi tersebut menyulitkan koordinasi antar-role, memperlambat proses review dan approval, serta meningkatkan risiko miskomunikasi, kesalahan pencatatan, dan ketidaksesuaian antara kebutuhan bahan dengan anggaran yang tersedia.

Action

Mengembangkan SiDaGi sebagai platform terpusat untuk mengelola data dapur dan menggantikan proses manual yang tersebar. Sistem dilengkapi workflow berbasis role, proses review dan approval, penghitungan kebutuhan serta budget otomatis, dan pengelolaan dokumen dari tahap perencanaan menu hingga laporan pertanggungjawaban.

Result

SiDaGi menghasilkan workflow operasional yang lebih terstruktur dengan data, approval, pengadaan, dan keuangan dalam satu platform. Sistem membantu mengurangi ketergantungan pada komunikasi WhatsApp, mempercepat koordinasi antar-role, mengurangi potensi kesalahan budget, dan meningkatkan keterlacakan transaksi serta laporan dapur.

Tools

- *Javascript*
- *Node*
- *React*
- *PostgreSQL*

Skills

- Full-stack Web Development
- Database Management
- Workflow Automation

<https://sidagi.com/>

Sentiment Analysis of Prambanan Temple Reviews

Menganalisis 3.400 ulasan TripAdvisor untuk mengukur sentimen pengunjung terhadap arsitektur, aktivitas, kebersihan, aksesibilitas, dan harga di Candi Prambanan.



Problem

Ulasan pengunjung mengandung opini yang tidak terstruktur dan sulit dianalisis secara manual. Diperlukan metode untuk mengelompokkan sentimen berdasarkan lima aspek utama yang ingin diketahui

Action

mengumpulkan ulasan dari TripAdvisor, membersihkan teks, melakukan tokenisasi, dan mengategorikan kalimat berdasarkan lima aspek. Sentimen dianalisis menggunakan Bing Lexicon, kemudian divisualisasikan melalui word cloud, grouped bar chart, dan bigram network.

Result

Analisis menunjukkan bahwa ulasan pengunjung secara keseluruhan cenderung positif. Aktivitas memperoleh respons paling positif, sedangkan harga dan arsitektur mencatat sentimen positif terendah. Temuan ini dapat mendukung evaluasi harga tiket, aksesibilitas, dan layanan pengunjung.

Tools

- *R*
- *dplyr*
- *ggplot2*

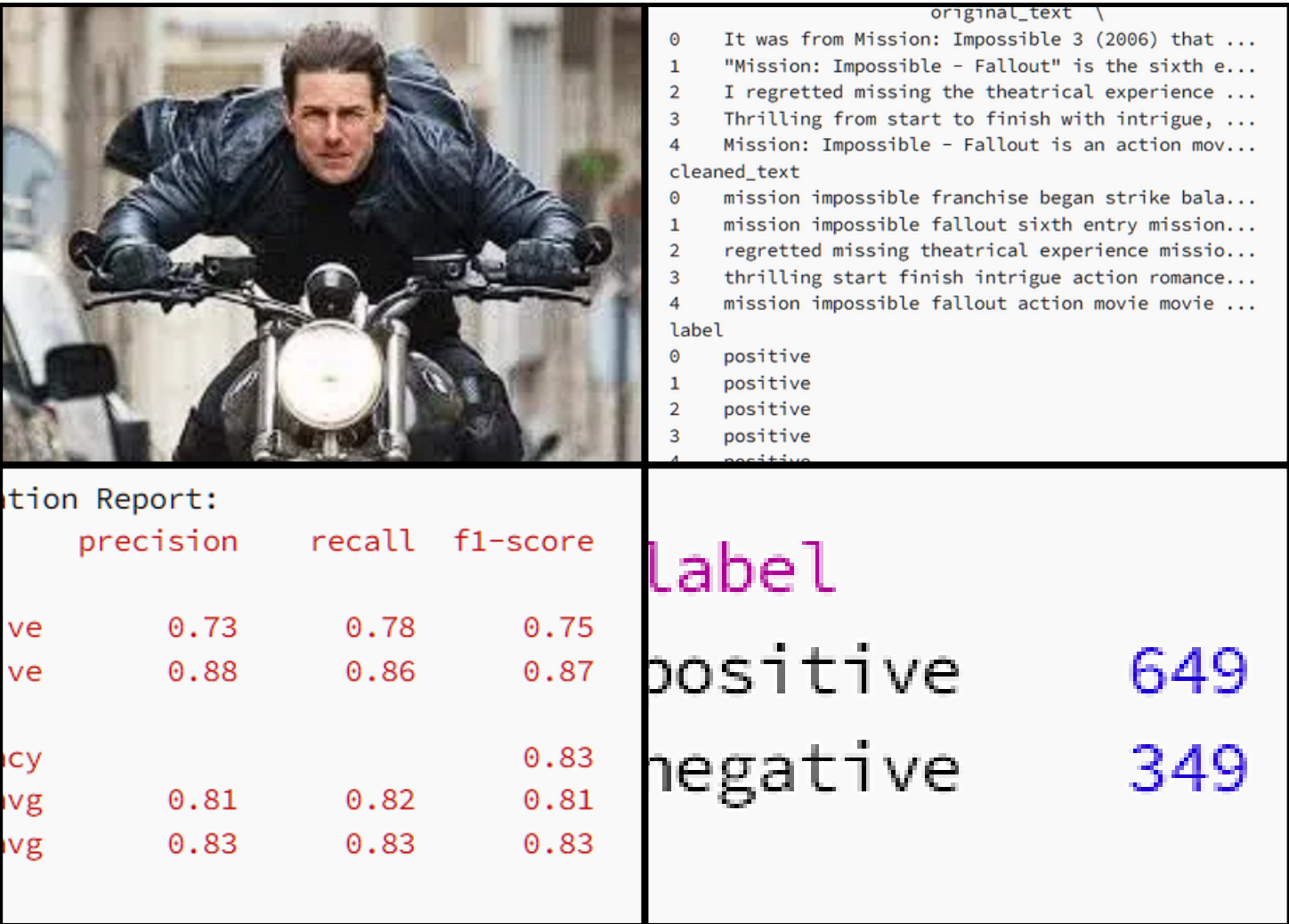
Skills

- NLP
- *Text Mining*
- *Sentiment Analysis*

<https://medium.com/@afghanalmalik2404/analyzing-sentiments-of-prambanan-temple-reviews-a774f423938a>

BERT Sentiment Analysis on IMDb Reviews

Membangun pipeline untuk scraping, preprocessing, pelabelan, fine-tuning BERT, dan evaluasi sentimen ulasan film *Mission: Impossible – Fallout*.



Problem

Ulasan IMDb merupakan data tidak terstruktur yang harus dikumpulkan dan dibersihkan sebelum digunakan. Selain belum memiliki label sentimen, opini pengguna juga memiliki konteks dan nuansa bahasa yang membutuhkan model transformer agar dapat diklasifikasikan dengan lebih akurat.

Action

membuat scraper berbasis Requests dan BeautifulSoup, kemudian melakukan tokenisasi, stopwords removal, dan lemmatization. Ulasan dilabeli dengan VADER, dibagi menjadi training dan testing set, lalu digunakan untuk fine-tuning BERT menggunakan PyTorch dan Hugging Face.

Result

Dataset menghasilkan 649 ulasan positif dan 349 ulasan negatif. Model BERT mencapai akurasi 83%, dengan F1-score 87% untuk sentimen positif dan 75% untuk sentimen negatif, menunjukkan performa yang baik meskipun klasifikasi negatif masih dapat ditingkatkan.

Tools

- Python
- BERT
- Hugging face

Skills

- Web Scrapping
- NLP
- transformer fine-tuning
- Sentiment Analysis

R Shiny Apps for Regression and ANOVA

Mengembangkan dua aplikasi interaktif untuk menyederhanakan Multiple Linear Regression dan ANOVA, termasuk input data, pengujian asumsi, visualisasi, prediksi, dan interpretasi hasil.



Problem

Proses Multiple Linear Regression dan ANOVA di R membutuhkan pemilihan variabel, pengujian asumsi, diagnostik, serta interpretasi output secara manual. Workflow tersebut memakan waktu dan dapat menyulitkan pengguna yang belum terbiasa dengan pemrograman statistik.

Action

Mengembangkan dua aplikasi R Shiny yang mendukung input CSV, pemilihan variabel, pengaturan train-test split, correlation matrix, diagnostic plot, dan pengujian asumsi. Aplikasi juga menyediakan prediksi regresi, ANOVA table, serta post-hoc Tukey HSD.

Result

Kedua aplikasi berhasil berjalan dan dilakukan test dengan dataset. Aplikasi MLR menghasilkan adjusted R^2 0,9873 dan skor prediksi 97,15, meskipun beberapa asumsi belum terpenuhi. Aplikasi ANOVA menemukan pengaruh sangat signifikan dari AdPlacement terhadap CTR, dengan CenterPage sebagai kelompok terbaik berdasarkan Tukey HSD.

Tools

- R
- Shiny

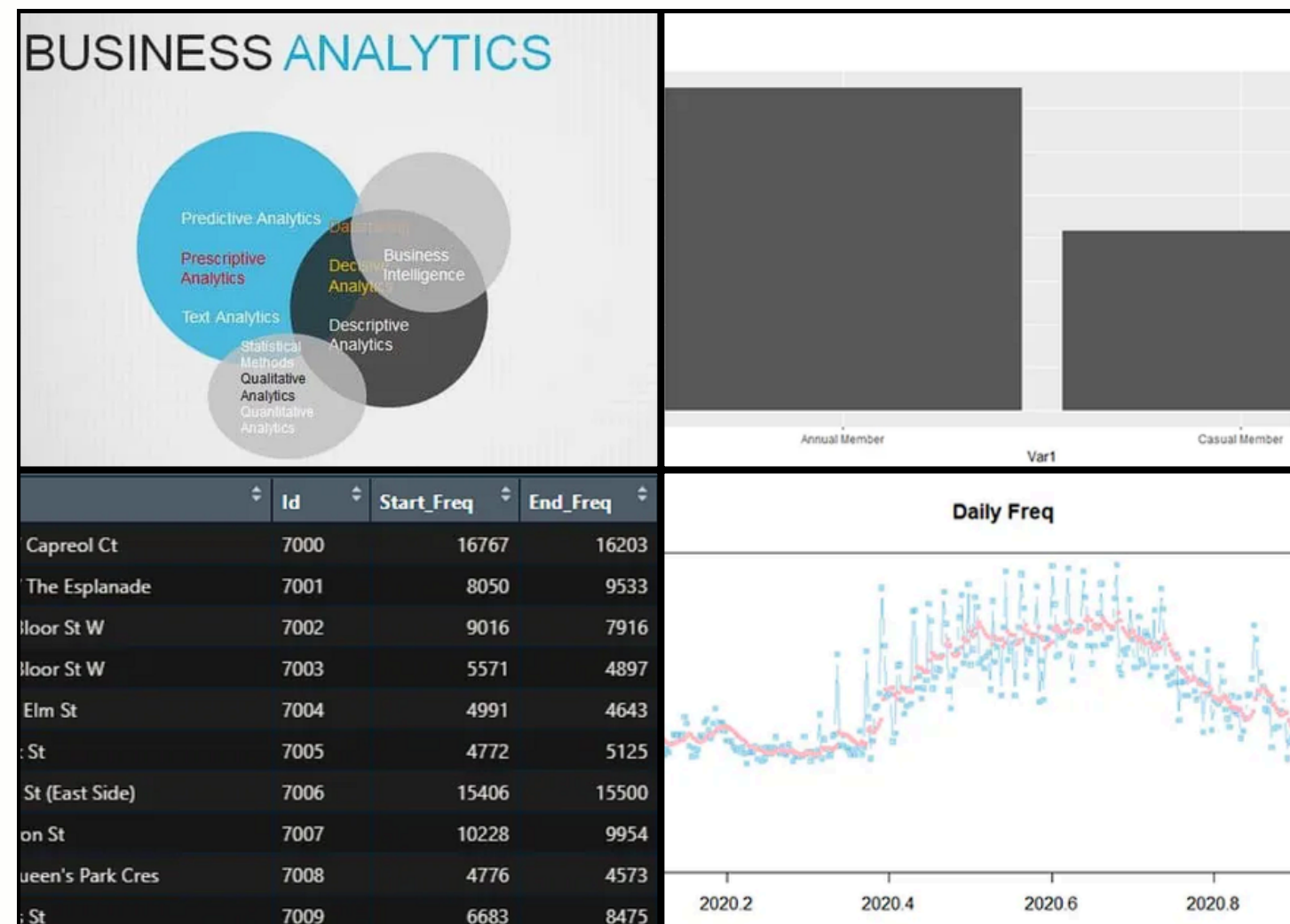
Skills

- Statistical Modelling
- *automated analytics*

<https://medium.com/@afghanalmalik2404/r-shiny-app-to-simplify-the-processes-of-multiple-linear-regression-and-anova-02b37c1ac27b>

Toronto BikeShare Business Intelligence Analysis

Menganalisis data perjalanan BikeShare Toronto tahun 2020 untuk memahami pola pengguna, sepeda, stasiun, waktu penggunaan, dan memprediksi tren 31 hari berikutnya.



Problem

Data perjalanan BikeShare Toronto tersimpan dalam dua belas dataset bulanan dan mencakup ratusan stasiun serta ribuan sepeda. Operator membutuhkan informasi tentang pola penggunaan, efektivitas armada, performa stasiun, dan perkiraan permintaan untuk mendukung keputusan operasional.

Action

Menggabungkan dataset bulanan, memvalidasi missing value, dan menyesuaikan tipe data. Analisis deskriptif dilakukan terhadap pengguna, armada, stasiun, durasi, serta pola waktu, kemudian dilanjutkan dengan Double Exponential Smoothing untuk memprediksi tren 31 hari.

Result

Analisis menemukan bahwa penggunaan tertinggi terjadi pada Sabtu, Agustus, dan sekitar pukul 17.00. Model Holt menghasilkan MAPE 38,95%, sehingga cukup memadai untuk melihat arah tren dan mendukung pengaturan armada, pemeliharaan, serta pengembangan stasiun.

Tools

- R
- dplyr
- ggplot2

Skills

- Business Intelligence
- Descriptive Analytics
- Time Series Analysis

<https://medium.com/@afghanalmalik2404/descriptive-statistics-analysis-the-key-to-business-intelligence-010b6a270510>

Household Object Image Classification Using CNN

Mengembangkan Convolutional Neural Network untuk mengklasifikasikan gambar sapu, tangga, ember, dan penggorengan menggunakan dataset yang tersedia pada yolov8.



Problem

Model harus mengklasifikasikan empat jenis peralatan rumah tangga dengan hanya sepuluh gambar per kategori. Keterbatasan data meningkatkan risiko overfitting dan menyulitkan model untuk mempelajari variasi bentuk, sudut pengambilan, serta latar belakang objek.

Action

melakukan augmentasi dan membagi gambar menjadi training serta validation set. Model dibangun menggunakan tiga convolutional layer, max pooling, dense layer, dan dropout, kemudian dilatih dengan Adam optimizer dan early stopping untuk mengurangi overfitting.

Result

Model memperoleh akurasi validasi 75%, hasil yang cukup baik untuk dataset berukuran sangat kecil. Pada pengujian gambar eksternal, tiga dari empat objek berhasil diklasifikasikan dengan benar, sedangkan gambar penggorengan masih salah dikenali sebagai tangga.

Tools

- Python
- Tensorflow
- yolov8

Skills

- Deep Learning
- Convolutional Neural Network (CNN)
- Image Recognition

<https://medium.com/@afghanalmalik2404/convolutional-neural-networks-for-image-classifying-using-python-3968d7b1f17b>

Terima Kasih

Ayo terhubung dengan saya!

✉ afghanalmalik2404@gmail.com

☎ +6287749833016

in <https://www.linkedin.com/in/afghan-al-malik-efendi/>

🐙 <https://github.com/afghanalmlk>

M <https://medium.com/@afghanalmalik2404>

