VINCENT OSBERT

Data Scientist | Machine Learning · Python · NLP · Decision Modeling

Malang, Indonesia (UTC+7) | Open to remote | +62 851-9899-8962 | vosbert151515@gmail.com
github.com/Vincent-Osbert | linkedin.com/in/vincent-osbert-6966b124b

PROFESSIONAL SUMMARY

Data Science undergraduate at Universitas Brawijaya who competes in national data science competitions: national finalist in DataQuest 5.0 (ordinal survey model, QWK 0.484), 9th place in NSC 2026 (sensor time-series forecasting), and an IndoBERT ensemble for Indonesian text in ACTION 2026 (competition macro F1 0.919). Works daily with AI coding agents to ship faster, and checks every result with cross-validation, leakage audits and tests before trusting it. Open to remote ML internships or part-time contracts.

EDUCATION

Universitas Brawijaya, Indonesia | *B.Sc. in Data Science* Aug 2024 - 2028 (expected)
GPA 3.53/4.00. Coursework: Artificial Intelligence, Data Mining, Big Data Analytics, Data Warehousing, Cloud Computing.

DATA COMPETITIONS

DataQuest 5.0, Universitas Airlangga: National Finalist | *Tree ensembles, QWK* Aug - Oct 2026

- In a 3-person team, predicted 4-level life satisfaction for 95,609 survey respondents (out-of-fold QWK 0.484); replacing argmax with expected-value thresholds added 0.06 QWK, and a learning curve over 5 data sizes flattened (log fit, R^2 0.91), pointing to missing survey questions rather than model capacity.
- Shipped a browser version of the model (440 trees, 17 features, 96.5% of full-pipeline QWK) and pitched it to the judges in the national final.

ACTION 2026 Data Mining Competition, Indonesia | *PyTorch, IndoBERT* Oct 2026 - present

- In a 3-person team, built an ensemble of IndoBERT base, IndoBERT large and IndoBERTweet to classify 23,297 Indonesian social-media comments into 9 public-service topics (multi-label); competition macro F1 (positive and negative class per topic) 0.919 in topic-grouped 5-fold CV; best public leaderboard score 0.932.

NSC 2026 (Indonesia): Air Humidity Forecasting, 9th place | *LightGBM, CatBoost* Aug - Sep 2026

- Wrote the team's full pipeline: a 4-model ensemble (LightGBM, CatBoost, ExtraTrees, XGBoost) with 8-fold chronological CV and time-based interpolation of sensor gaps; private leaderboard RMSE 3.91 on 734 unseen timestamps, notebook reruns end to end.

ARA ITS 7.0: Pothole Segmentation, 8th place | *PyTorch, SegFormer MiT-B5* 2026

- Fine-tuned a SegFormer MiT-B5 transformer on 498 labelled road images to draw pixel-level pothole masks, using Tversky loss for small potholes, a 5-fold ensemble and flip test-time augmentation: CV Dice 0.83, final private-leaderboard Dice 0.804.

Satria Data 2026: Big Data Challenge (Puspresnas) | *PyTorch, ConvNeXt, EfficientNet* 2026

- In a team, classified 26,525 waste images into recyclable, electronic and organic; moved from EfficientNet-B3 to ConvNeXt-Tiny and reached 97.3% validation accuracy vs the team's best 99%, tracing the gap mainly to training on a 3,000-image subset.

FIK FAIR 2026 Data Science Competition | *Gaussian Process, TabPFN* Sep - Oct 2026

- In a 3-person team, built the repeated-CV protocol and the final 5-seed ensemble (GP, kernel ridge and SVR hybrids + TabPFN v2): repeated-CV RMSE 7.60; kept a row-order feature (Spearman -0.98) out as leakage.

PROJECTS

Hotel Overbooking from Late-Cancellation Risk | *Python, scikit-learn, SciPy* Oct 2026

Olist Review Drop: Root-Cause Analysis | *SQL (DuckDB), Python, pandas* Oct 2026

SKILLS

Tools: Python (pandas, NumPy, scikit-learn, LightGBM, XGBoost, CatBoost, SciPy, PyTorch, Hugging Face Transformers), SQL (DuckDB: CTEs, window functions), Git/GitHub
Methods: classification, decision optimization (newsvendor), regression, NLP (multi-label text classification), computer vision (segmentation, image classification), time-based and grouped cross-validation, leakage checks, model evaluation (ROC-AUC, Brier, F1, RMSE)
AI-assisted work: Claude Code and LLM coding agents for EDA, modeling pipelines, documentation and code review; blind auditor agents to check reports; outputs verified with tests and cross-validation
Languages: Indonesian (native), English (fluent), Mandarin (basic)