Presentation by **Qonitah Sarah**

Final Project Data Science

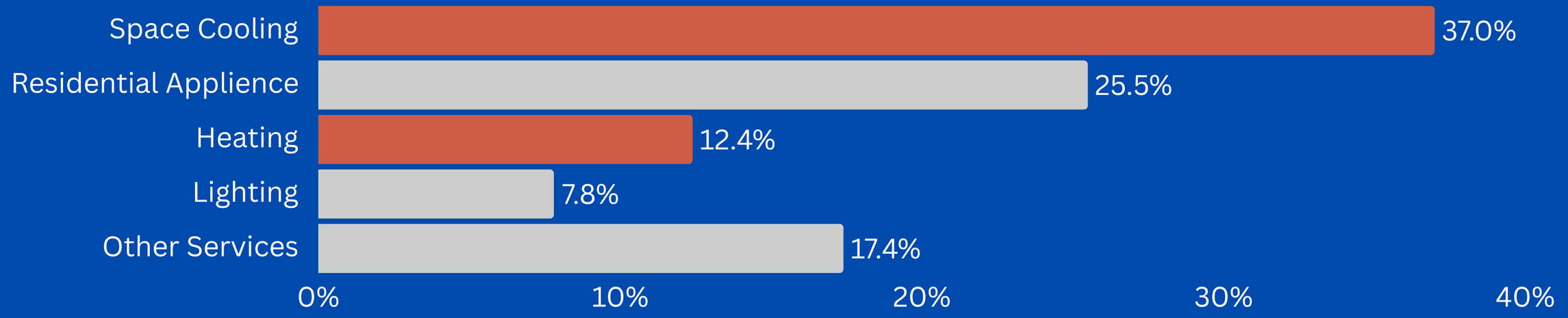
Keep Preferring Silent Waste — or Move to Smart Prediction?

Membangun Supervised Learning
Untuk Memprediksi Tingkat
Okupansi Ruangan

Bagaimana memprediksi secara otomatis jumlah penghuni yang berada di suatu ruang sebagai input untuk sistem kontrol HVAC?

Bangunan dalam perencanaannya sangat mempertimbangkan berbagai hal terkait dengan kenyamanan, dan salah satu bagiannya adalah terkait ***Thermal Comfort*** yang seringnya kita jumpai diwujudkan dengan pemasangan alat-alat pendingin, pemanas, ventilasi atau yang disebut dengan HVAC. Kenyamanan di satu sisi menjadi kebutuhan dan di sisi lain menjadi hal yang bisa memicu berbagai dampak buruk terutama kepada lingkungan.

Oleh karena itu, salah satu cara yang dapat ditempuh sebagai jalan tengah adalah mengoptimalkan kinerja pada berbagai peralatan penghawaan yang berbasis permintaan.



Dibuat ulang dari sumber asli: Air Conditioning Biggest Factor in Growing Electricity Demand

Perjalanan menuju hasil dengan KDD framework

Selection

Preprocessing / Cleaning

Transformation

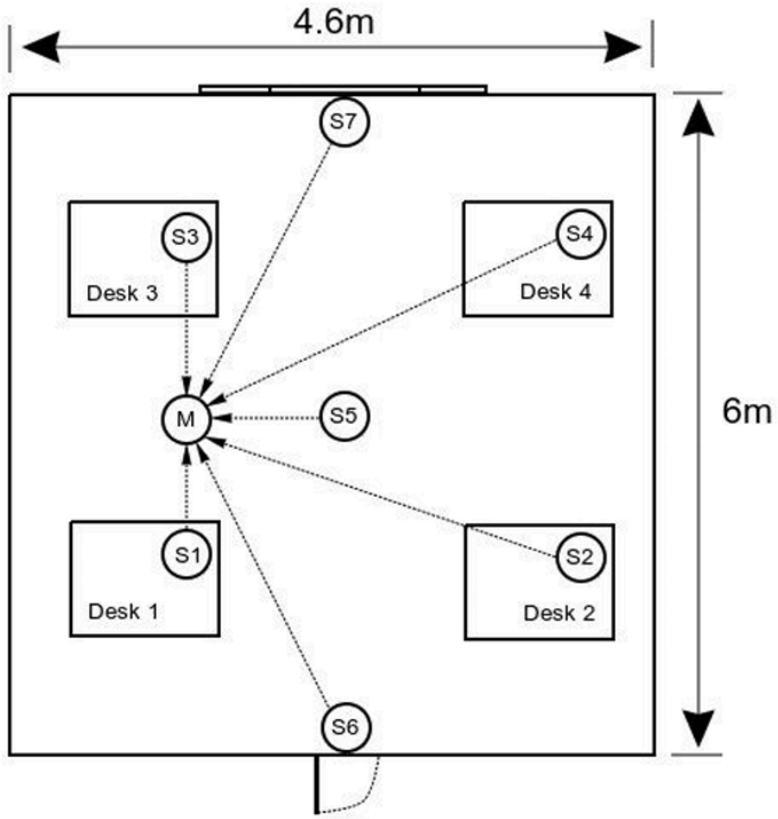
Data Mining

Interpretation / Evaluation

Berkenalan dengan data yang akan model pelajari: Room Occupancy Estimation

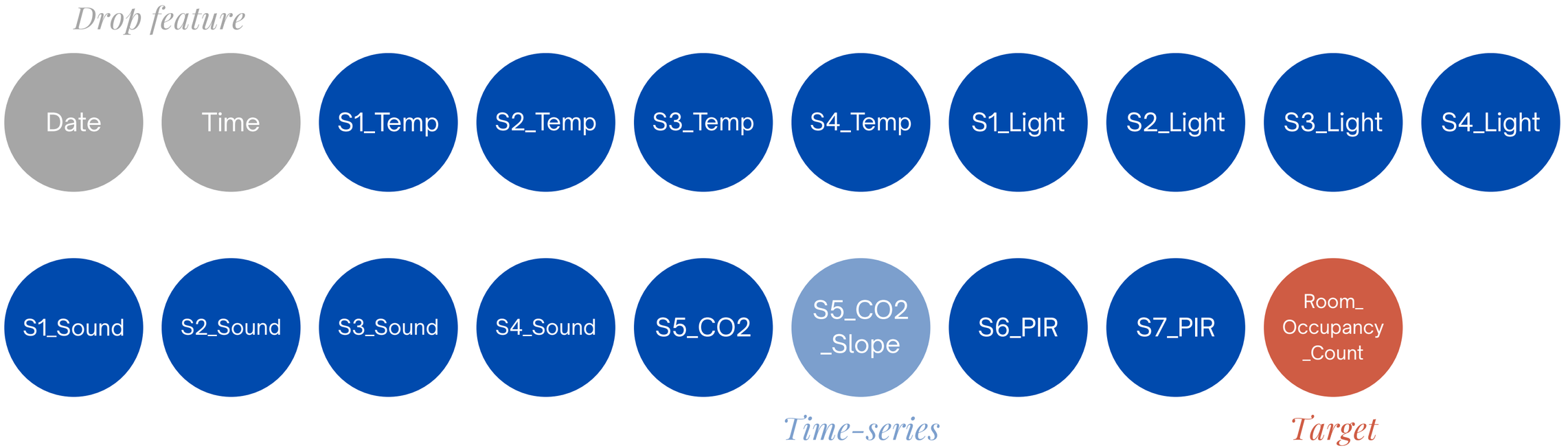
Link dataset dapat diakses di sini!

Dataset merupakan hasil eksperimen di sebuah ruangan (test room) berukuran **4,6 x 6 m** yang dilaksanakan dalam **7 hari**. Pendekatan yang dilakukan untuk mendeteksi jumlah penghuni adalah dengan ***non instrusive enviromental sensors***, dimana jumlah dapat diidentifikasi tanpa mengorbankan privasi. 5 sensor yang digunakan pada penelitian di antaranya adalah:



CO2 Temperature Light Motion Sound

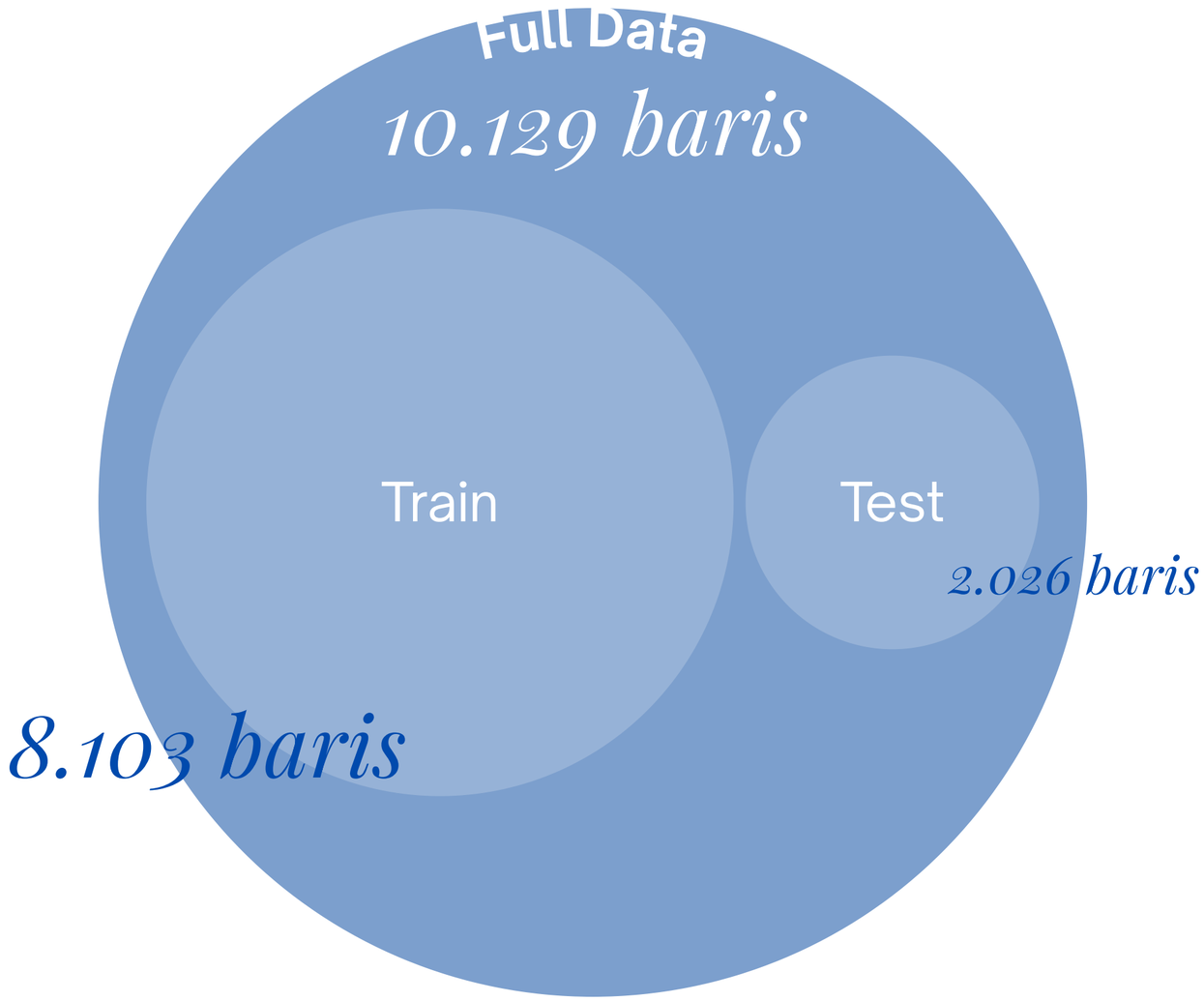
Sumber gambar denah: Machine Learning-based Occupancy Estimation Using Multivariate Sensor Nodes Journal



Membangun model klasifikasi supervised learning berdasarkan data hasil sensor

Data merupakan output dari hasil berbagai sensor. Masing-masing-masing sensor beserta lokasi penempatannya berdiri sendiri sebagai masing-masing kolom. **Salah satu fiturnya adalah CO2 slope yang memiliki baris dependen antar waktu ke waktu sehingga data memiliki tipe time-series**.

Label (target) yaitu **Room_Occupancy_Count adalah ground truth yang diperoleh secara manual melalui pengisian absen** yang memiliki 4 nilai unik, setiap angka mengindikasikan jumlah penghuni yaitu 0, 1, 2, dan 3 orang.



Data dibagi dengan perbandingan 80:20 dan displit dengan tetap memperhatikan urutan waktu

Persentase data lebih dominan untuk pelatihan model (data train) agar model dapat belajar lebih banyak data sehingga dapat menghadapi kondisi real yang akan sangat bervariasi.

Pembagian tetap mempertahankan urutan waktu dengan baris 1 - 8.103 adalah data palatihan sementara baris selanjutnya yaitu urutan ke-8.104 - 10.129 menjadi data test. Pembagian berdasarkan rentang waktu adalah untuk **menghindari data leakage pada fitur S5_co2_slope**.

Duplicate Value

Tidak ditemukan nilai yang duplikat

Missing Value

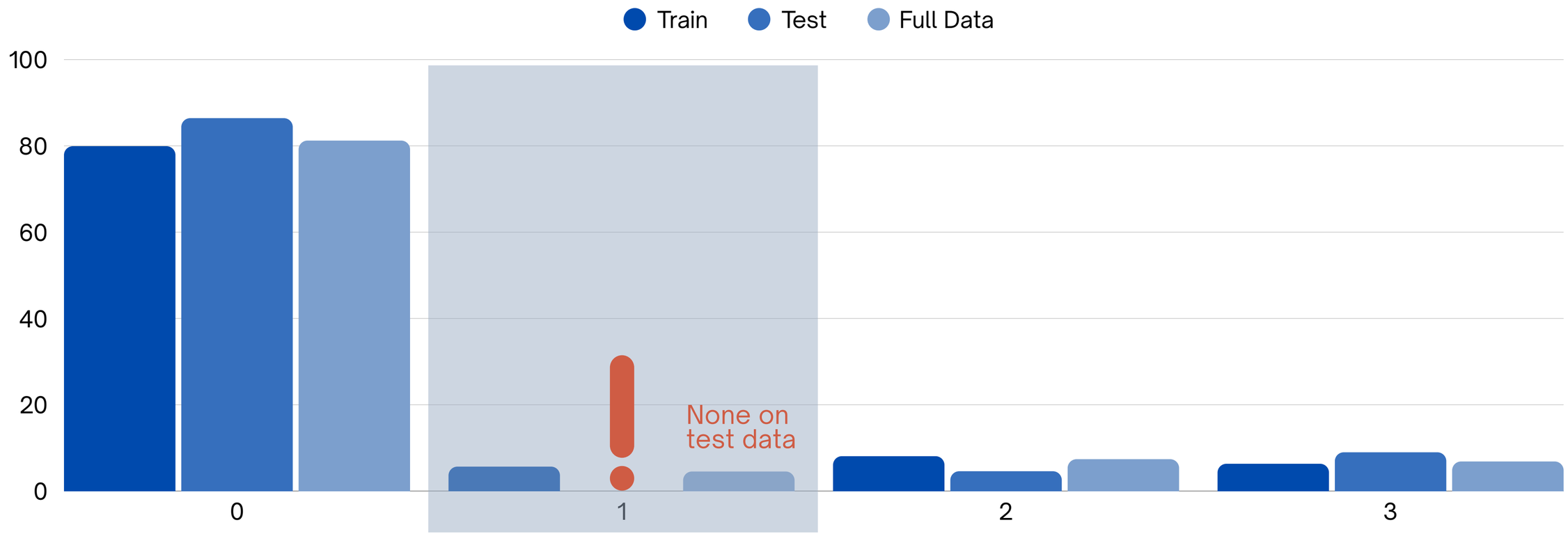
Tidak terdapat missing value yang bernilai NaN maupun value yang tidak sesuai

Outlier

Nilai outlier berada **hampir di seluruh fitur kecuali pada fitur temperatur dengan penempatan lokasi S_1, S_3, dan S_4**. Outlier merupakan **nilai yang wajar** sehingga dipertimbangkan untuk tidak dihandle

Trade off yang harus diterima: Fitur target pada data test tidak memiliki nilai 1

Untuk **mempertahankan rentang waktu** dan **terhindar dari resiko kebocoran (leakage) pada data**, maka pada **data test tidak memiliki fitur target bernilai 1** yang berimbas pada evaluasi nilai 1 yang tidak bisa diketahui dan F1 Score akan bernilai sangat buruk padahal model telah mempelajari nilai 1 sebelumnya. **Oleh karena itu metrik evaluasi utama untuk menilai performa adalah melalui cross validation**.



Fitur kategorikal telah terdistribusi dengan baik antar data namun tidak dengan fitur numerik

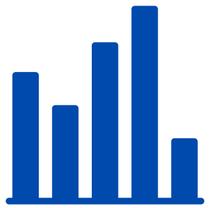
Numerikal

*Distribusi pada data numerik **tidak cukup representatif**. Ditribusi pada data train cenderung bervariasi dibandingkan data test sehingga **berpotensi overfitting**.*

Kategorikal

*Distribusi pada data kategori telah **tersebar dengan baik**. Perbedaan persentase tidak signifikan antar data.*

Bagaimana karakteristik di setiap fitur pada data pelatihan (train)?



Distribusi data

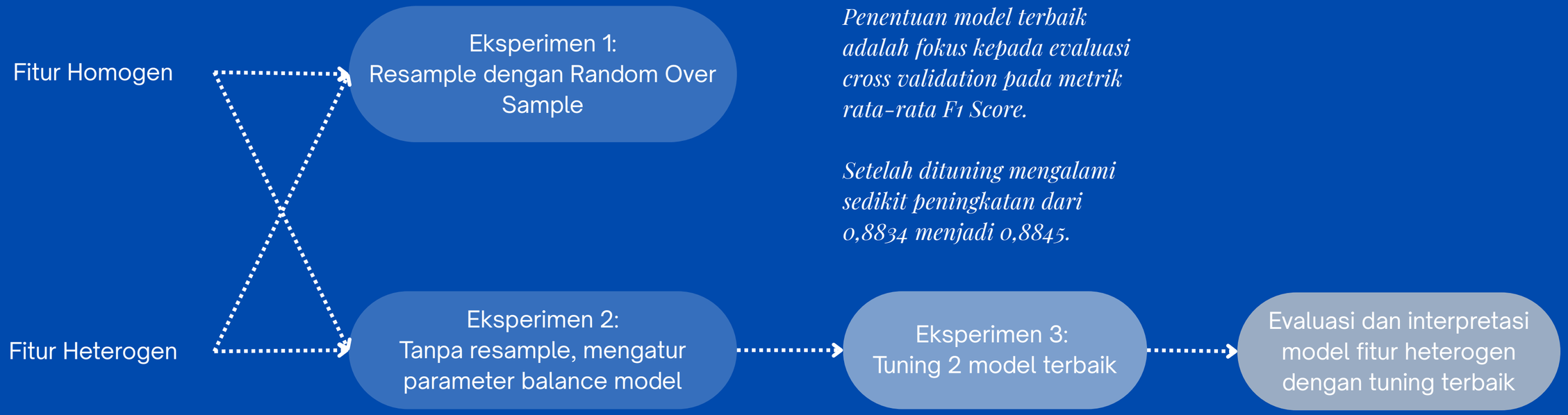
Ditribusi data pada **fitur numerik** memiliki **pola yang sangat skew** karena dipengaruhi oleh outlier. Sementara **fitur kategori** yaitu sensor gerak S6_PIR dan S_7 PIR pada masing-masing nilainya **tidak terdistribusi secara seimbang** sehingga model yang kompleks akan dominan mempelajari nilai mayoritas dan berpotensi overfitting.

Korelasi

Metode korelasi pada fitur numerik menggunakan **Pearson** dan **Spearman** untuk melihat dari 2 perspektif. **Fitur yang berkolerasi kuat pada Pearson tidak seluruhnya berkolerasi kuat pula pada Spearman**. Hal ini lantaran nilai outlier sangat berpengaruh pada Pearson dan lebih robust terhadap Spearman.

Kemampuan membedakan target

Baik pada fitur numerik maupun kategori, masing-masing nilai pada target **mampu membedakan dengan baik**. Fitur numerik divisualisasikan dengan boxplot sementara kategori melalui 100% stacked bar chart. **Semakin tinggi tingkat okupansi maka rentang distribusi cenderung ikut naik**.



Penentuan model terbaik adalah fokus kepada evaluasi cross validation pada metrik rata-rata F1 Score.

Setelah dituning mengalami sedikit peningkatan dari 0,8834 menjadi 0,8845.

Menggunakan 4 model menyesuaikan dengan karakter data. Membangun alur pipeline yang berbeda

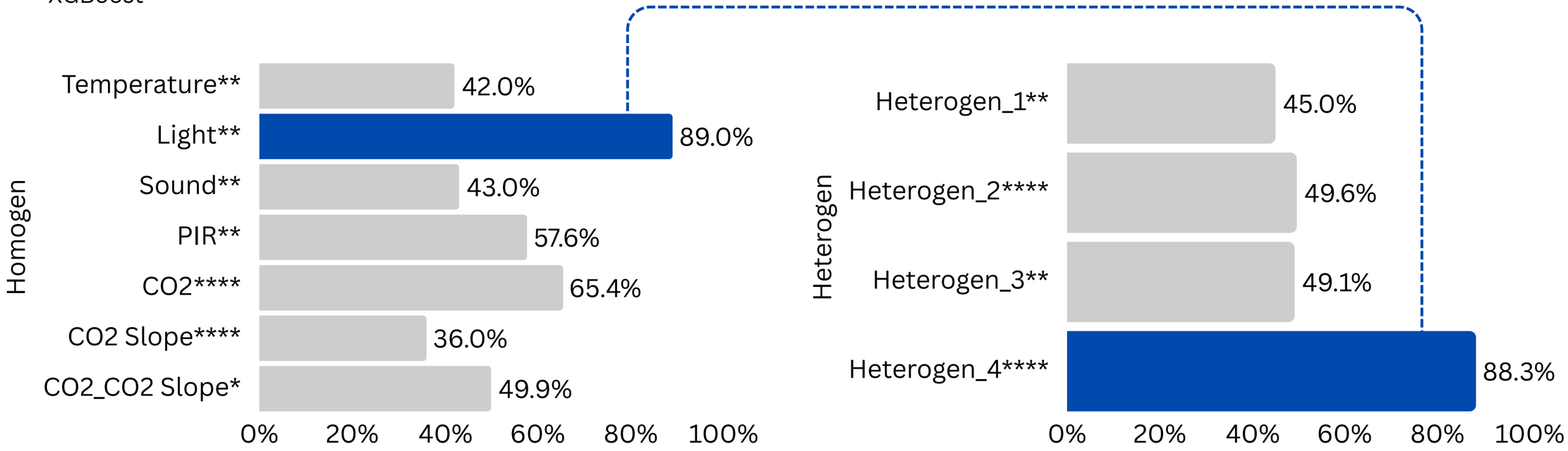
Fitur Homogen merupakan fitur-fitur yang sejenis sementara fitur heterogen adalah gabungan antar fitur yang berbeda. Homogen yang diuji adalah sebanyak **7 kelompok** dan heterogen berjumlah **4 kombinasi fitur** yang diuji secara bertahap.

Seluruh fitur dipelajari oleh 4 model berbeda yaitu **Decision Tree (baseline), SVM RBF. Random Forest, dan XGBoost**. SVM RBF merupakan model satu-satunya yang dilakukan scaling dengan **Robust Scaler karena model yang lainnya robust terhadap outlier.**

Bagaimana performa evaluasi cross validation dengan metode tanpa resample pada kelompok fitur homogen dan heterogen?

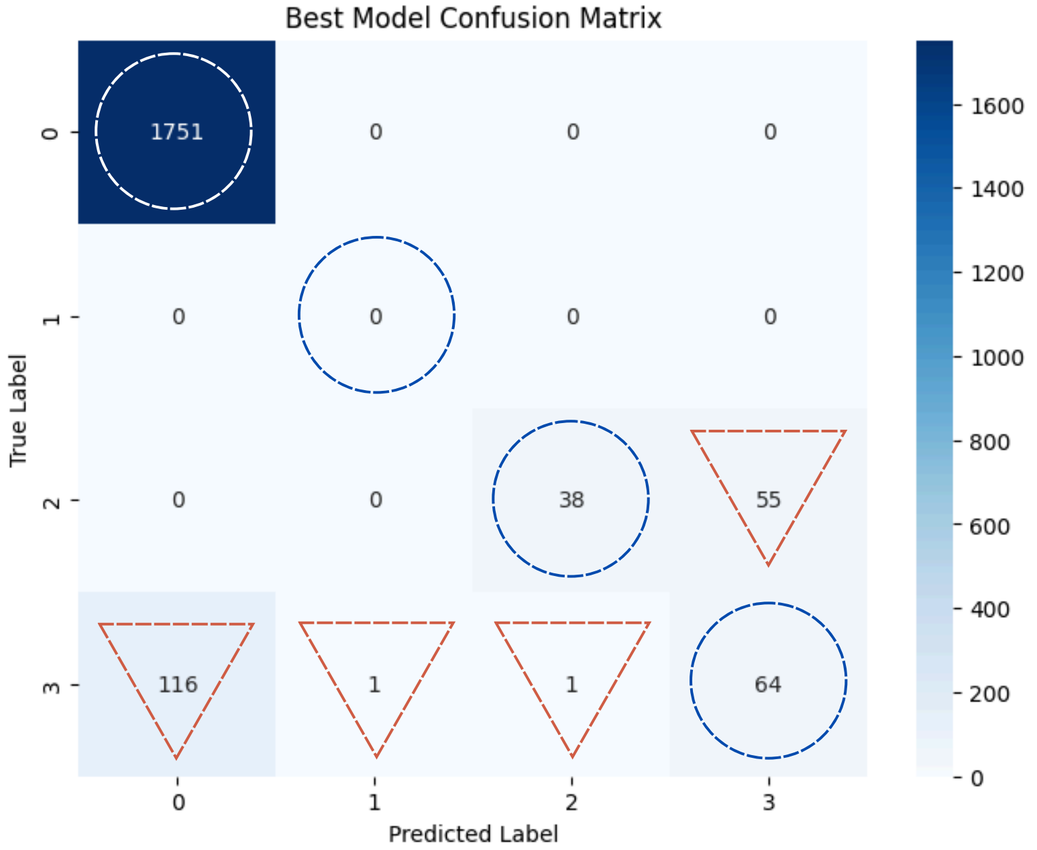
*	Decision Tree (Baseline)
**	SVM RBF
***	Random Forest
****	XGBoost

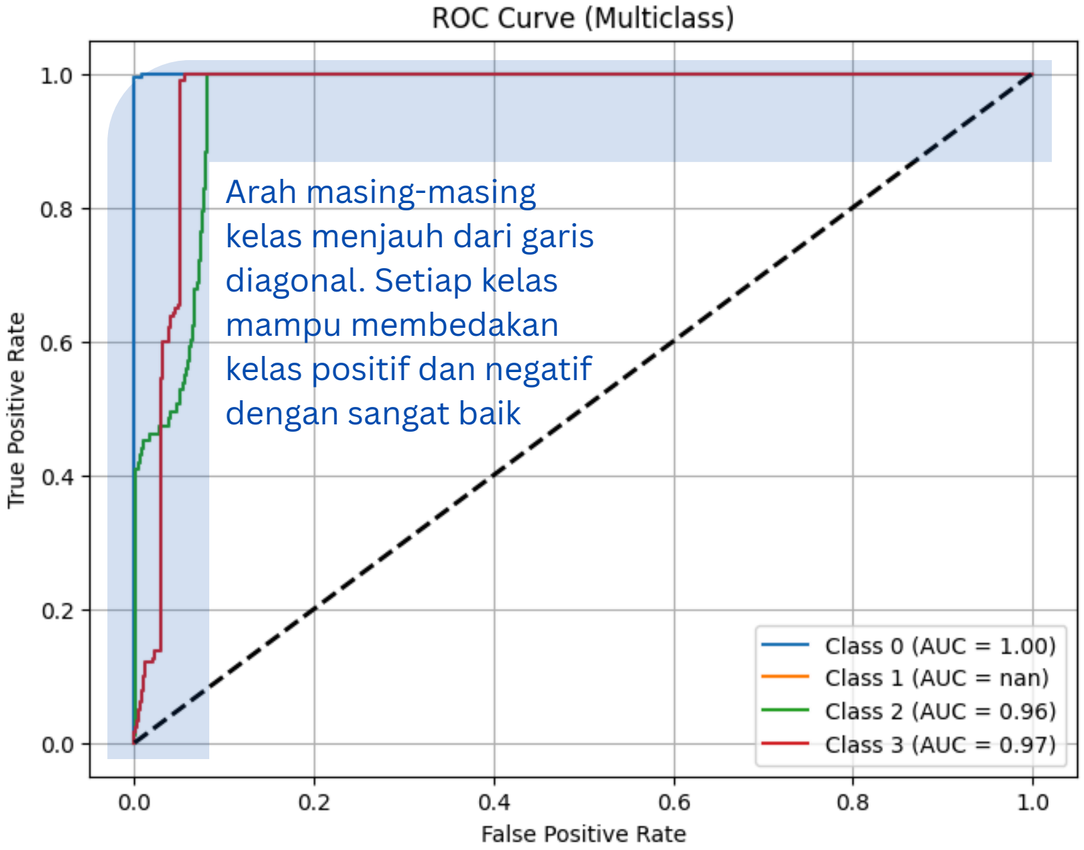
Sensor lampu bernilai sangat signifikan dibanding fitur sensor yang lain. Perfoma rata-rata F1 Score sangat meningkat saat sensor lampu dilibatkan sebagai fitur untuk membuat prediksi.



Final evaluation* dengan performa terbaik: Fitur Heterogen dengan model XGBoost

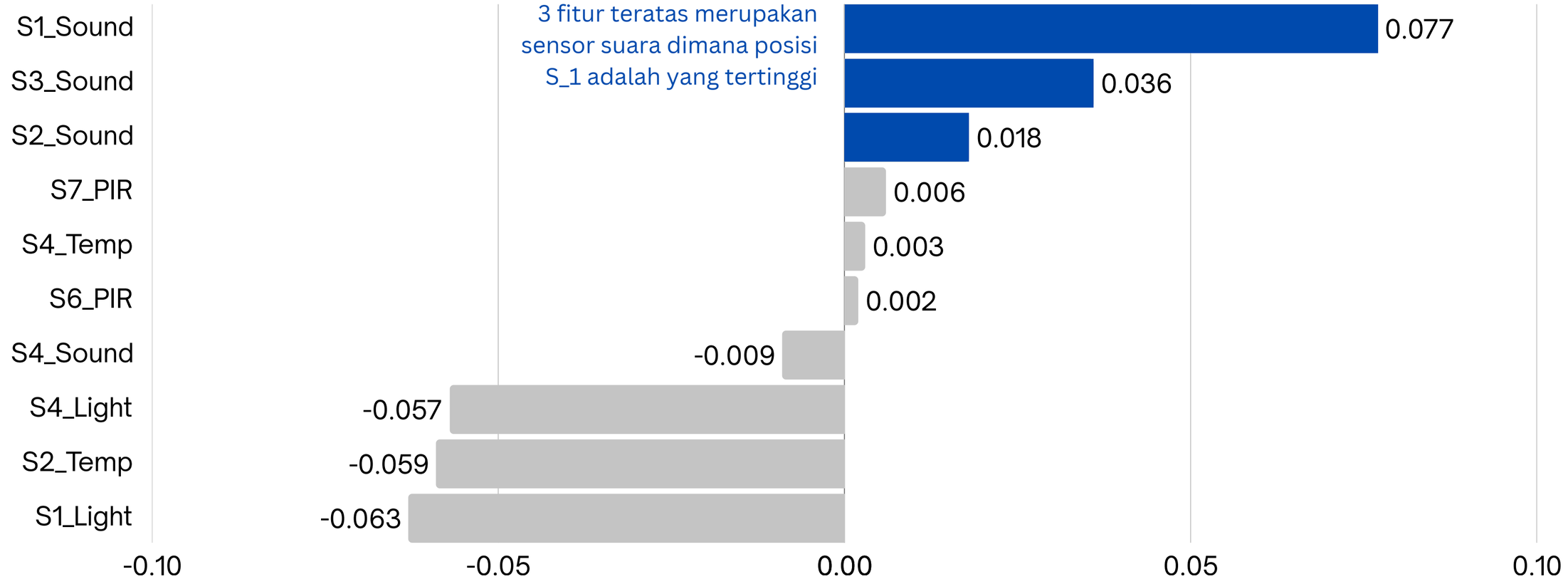
False Prediction True Prediction





*Visualisasi metrik evaluasi di atas merupakan hasil pengujian dari data test yang tidak representatif yang sebelumnya telah dijelaskan pada tahap transformation.

Fitur sensor suara dianggap sangat penting oleh model XGBoost



Tetap menggunakan Sensor

Or

Upgrade alat sensor

Jika alat sensor tetap digunakan, maka metode sama dengan yang dilakukan sebelumnya yaitu **berbasis waktu namun perlu dilakukan penambahan data lebih banyak lagi** dan dipantau secara real time agar tergenalisir dengan baik dan menghadapi resiko jika **tidak dapat dideploy secara baris per baris** karena perlu diproses beberapa waktu untuk menghasilkan co2 slope.

Jika alat sensor co2 diupgrade menjadi lebih responsif sehingga perubahan dapat secara otomatis diketahui maka pembagian data untuk pembuatan model **tidak perlu memperhatikan rentang waktu**, sehingga data walaupun jumlahnya tidak terlalu banyak (seperti jumlah yang sekarang) tetap dapat digunakan dan model tetap dapat belajar dengan baik, serta keunggulan lainnya dapat **di deploy secara per baris**.

I’m highly open to feedback, meaningful discussions, and future collaborations,

Let’s connect!

wa.me/6282187655739

qonitahsarah112@gmail.com

linkedin.com/in/qonitahsarah