BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan mengenai perbandingan performa metode klasifikasi berbasis *Machine learning* dalam penentuan tingkat pencemaran air, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Efektivitas pra-pemrosesan data penerapan teknik *Mean Imputation* terbukti efektif dalam menangani masalah nilai yang hilang (*missing values*) pada parameter krusial seperti *pH*, *Sulfate*, dan *Trihalomethanes* tanpa mengurangi jumlah sampel secara signifikan. Selain itu, teknik normalisasi (standarisasi) berhasil menyamakan skala distribusi parameter fisikokimia air, sehingga setiap fitur memiliki kontribusi yang seimbang dalam proses pelatihan model.

2. Penanganan ketidakseimbangan data teknik *manual oversampling* (*Random Oversampling*) berhasil mengatasi ketidakseimbangan kelas (*imbalanced data*) pada dataset kualitas air. Metode ini mampu menyeimbangkan proporsi kelas "Layak Minum" dan "Tidak Layak Minum" menjadi rasio 50:50, yang mencegah model mengalami bias terhadap kelas mayoritas dan meningkatkan kemampuan model dalam mengenali kelas minoritas.

3. Berdasarkan perhitungan skor signifikansi menggunakan metode seleksi fitur *ANOVA F-value* (*SelectKBest*), diketahui bahwa parameter *Chloramines*, Solids, *Organic_carbon*, dan *Hardness* memiliki tingkat pengaruh paling signifikan terhadap prediksi potabilitas air karena memperoleh skor tertinggi dibandingkan fitur lainnya. Sebaliknya, parameter *pH*, *Turbidity*, dan *Sulfate* terbukti memiliki kontribusi yang paling rendah terhadap hasil klasifikasi kelayakan air minum karena menunjukkan nilai skor signifikansi yang paling kecil.

4. Berdasarkan pengujian komprehensif terhadap tujuh algoritma klasifikasi, algoritma *Random Forest* terbukti menjadi model yang paling reliabel dan optimal. Keunggulan ini ditunjukkan oleh pencapaian Akurasi tertinggi sebesar 85,50%, *ROC-AUC* 0,9074, PR-Score 0,9369, serta tingkat kesalahan klasifikasi paling rendah berdasarkan hasil evaluasi *Confusion Matrix*. Sebagai tindak lanjut, model *Random Forest* terbaik ini telah berhasil diimplementasikan secara praktis ke dalam antarmuka berbasis web menggunakan kerangka kerja Streamlit, yang memungkinkan pengguna mendapatkan hasil prediksi kualitas air secara *real-time*.

5.2 Saran

Berdasarkan keterbatasan dan hasil yang diperoleh dalam penelitian ini, penulis mengajukan beberapa saran untuk pengembangan penelitian selanjutnya:

1. Pengembangan teknik *resampling* pada penelitian ini menggunakan *Random Oversampling* untuk menangani ketidakseimbangan data. Penelitian selanjutnya disarankan untuk membandingkan teknik *resampling* yang lebih kompleks, seperti *Synthetic Minority Over-sampling Technique* (*SMOTE*) atau ADASYN, untuk melihat apakah variasi sintetis data dapat lebih meningkatkan generalisasi model.
2. Optimasi hiperparameter menunjukkan bahwa model seperti *XGBoost* memiliki potensi tinggi namun masih berada di bawah *Random Forest*, yang mungkin disebabkan oleh sensitivitas terhadap parameter. Disarankan untuk menerapkan metode pencarian hiperparameter otomatis seperti *Grid Search* atau *Bayesian Optimization* untuk menemukan konfigurasi parameter terbaik bagi setiap algoritma guna memaksimalkan akurasi.
3. Eksplorasi data tambahan mengingat ketergantungan model pada data sekunder publik, penelitian di masa depan dapat diperkaya dengan menggunakan dataset primer atau menggabungkan data dari berbagai sumber geografis yang berbeda. Hal ini bertujuan untuk menguji ketahanan (*robustness*) model terhadap variasi karakteristik air di lokasi nyata.