BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai evaluasi efektivitas empat model klasifikasi, yaitu XGBoost, Random Forest, Logistic Regression, dan K-Nearest Neighbors (KNN) dalam memprediksi gaji di sektor data science menggunakan dataset "Data Science Salaries 2023" dapat disimpulkan bahwa:

1. XGBoost merupakan model dengan kinerja terbaik, mencapai akurasi sebesar 98,93% dan log loss sebesar 0,038. Model ini menunjukkan kemampuan superior dalam menangani dataset yang kompleks dan menghasilkan prediksi yang sangat akurat.
2. Random Forest juga memberikan performa yang kompetitif dengan akurasi 96,67%, meskipun dengan log loss yang lebih tinggi dibandingkan XGBoost.
3. Logistic Regression menunjukkan hasil yang cukup baik, tetapi tidak seoptimal algoritma berbasis pohon dalam menangani hubungan antar fitur yang kompleks. Logistic Regression memberikan hasil yang stabil namun kurang efektif pada data dengan distribusi yang kompleks.
4. K-Nearest Neighbors (KNN) menunjukkan performa yang lebih rendah dengan akurasi 85,35% dan log loss yang relatif tinggi. Selain itu, KNN sensitif terhadap skala fitur, kesulitan menangani dataset yang tidak seimbang, serta menghadapi tantangan dalam pemilihan parameter jarak dan ketergantungan pada distribusi data.

Penelitian ini menekankan pentingnya tahapan pra-pemrosesan data, seperti pembersihan duplikasi, rekayasa fitur, dan penyesuaian variabel, yang berkontribusi signifikan terhadap peningkatan kinerja model. Selain itu, algoritma berbasis ensemble seperti XGBoost direkomendasikan untuk analisis prediktif yang memerlukan akurasi tinggi dan kemampuan menangani kompleksitas data.

5.2 Saran

Penelitian ini masih memiliki ruang untuk pengembangan, terutama dalam pengayaan dataset. Penelitian mendatang disarankan menggunakan data lintas tahun atau waktu nyata untuk menghasilkan prediksi yang lebih dinamis. Selain itu, eksplorasi algoritma lain seperti deep learning dapat meningkatkan akurasi, sementara teknik penanganan ketidakseimbangan data seperti oversampling atau undersampling dapat memperbaiki representasi prediksi.

Penyertaan variabel tambahan, seperti tingkat pendidikan dan pengalaman internasional, juga dapat memberikan analisis yang lebih komprehensif. Kendala waktu dan keterbatasan infrastruktur komputasi yang dihadapi dapat diatasi dengan memanfaatkan layanan cloud untuk efisiensi. Dokumentasi yang lebih sistematis pada setiap tahap, seperti pra-pemrosesan data dan rekayasa fitur, sangat penting untuk memudahkan reproduksi penelitian. Terakhir, visualisasi hasil dapat ditingkatkan untuk memberikan wawasan lebih mendalam dan memperjelas hubungan antar variabel serta kinerja model. Dengan langkah-langkah ini, penelitian di masa depan dapat lebih akurat dan relevan.