Laporan Eksperimental : Pengaruh Dimensionality Reduction dan Regularization pada Model Supervised

Kelompok :

1. Rayhan Firdaus Ardian (23/519095/PA/22279)
2. Amalia Muti'ah Khairunnisa
3. Kartika Adhi N.W
4. Atika Dwi Aryanti
5. Iffa Hesti

Ringkasan

Laporan ini menyajikan hasil eksperimen menggunakan empat algoritma machine learning (Logistic Regression, Naive Bayes, SVM, dan KNN) dengan tiga variasi pendekatan: baseline, dimensionality reduction dengan PCA, dan kombinasi PCA dengan regularisasi/optimasi. Dataset yang digunakan berupa adalah Breast Cancer Wisconsin yang didapat dari kaggle.

Algoritma yang Dievaluasi

1. Logistic Regression (LR) : Model linear untuk *biner classification.*
2. Naive Bayes (NB) : Classifier Probabilistik berdasarkan Bayes Theorem.
3. Support Vector Machine (SVM) : Mencari hyperplane optimal untuk pemisah kelas.
4. K-Nearest Neighbor : Klasifikasi berdasarkan kemiripan dengan tetangga terdekat.

Variasi Eksperimental

1. Baseline : Model dasar
2. PCA : Principal Component Analysis (Dimensional Reduction) dengan 95% variance
3. PCA + Regularization/Optimization :
    - LR : Regularisasi L2 dengan GridSearch untuk Parameter C yang optimal
    - NB : Optimasi var_smoothing
    - SVM : Optimasi parameter C dan gamma
    - KNN : Optimasi n_neighbors, weights dan parameter jarak (p)

Metriks Evaluasi

1. Accuracy
2. ROC AUC → Area Under Receiver Operating Characteristic curve, mengukur kemampuan diskriminatif model

Hasil dan Analisis

Perbandingan Performa Algorithm

Hasil eksperimen menunjukkan bahwa Logistic Regression menggunakan PCA + Regularization memberikan performa terbaik, dengan Accuracy 0.982 dan ROC AUC 0.998.

	Model	Algorithm	Variant	Accuracy	ROC AUC
1	PCA + Reg LR	Logistic Regression	PCA + Reg	0.982456	0.998016
2	PCA LR	Logistic Regression	PCA	0.973684	0.997024
3	Baseline LR	Logistic Regression	Baseline	0.964912	0.996032
4	PCA + Opt SVM	SVM	PCA	0.964912	0.995370
5	PCA SVM	SVM	PCA + Opt	0.964912	0.995370
6	Baseline SVM	SVM	Baseline	0.973684	0.994709
7	Baseline NB	Naive Bayes	Baseline	0.921053	0.989087
8	PCA KNN	KNN	PCA	0.956140	0.983796
9	PCA + Opt KNN	KNN	PCA + Opt	0.964912	0.982474
10	Baseline KNN	KNN	Baseline	0.956140	0.982308
11	PCA NB	Naive Bayes	PCA	0.894737	0.961310
12	PCA + Opt NB	Naive Bayes	PCA + Opt	0.894737	0.961310

Analisis

1. Logistic Regression
    - PCA mengurangi dimensi tanpa kehilangan performa signifikan
    - Accuracy dan ROC AUC tertinggi dicapai oleh PCA + Reg Logistic Regression
    - Accuracy dan ROC AUC seiring meningkat dengan menambahkan variasi.
2. Naive Bayes
    - Menambah variasi PCA dan PCA + Opt justru menurunkan accuracy dan ROC AUC
    - PCA dan PCA + Opt memiliki accuracy dan ROC AUC yang sama
3. SVM
    - Tidak ada perbedaan antara PCA dan PCA + Opt baik di Accuracy ataupun ROC AUC
    - Baseline SVM memiliki accuracy paling tinggi namun ROC AUC paling rendah dibanding PCA dan PCA + Opt
4. KNN
    - PCA + Opt KNN memberikan accuracy tertinggi (0.956) dibanding PCA KNN dan baseline KNN, namun ROC AUC milik PCA + Opt KNN relatif lebih rendah (0.956) dibanding PCA KNN
    - Accuracy baseline KNN dan PCA KNN sama (0.956140)
    - PCA KNN memiliki ROC AUC Paling Tinggi

Mengapa Respons Tiap Model berbeda?

Karakteristik Model & Asumsi

1. Logistic Regression dan SVM adalah model linear (atau kernel-based) yang profit dari reduksi dimensi:
    a. PCA menghilangkan fitur yang berisik/kolinear tanpa mengorbankan informasi penting,
    b. L2-regularization (LR) atau tuning C/γ (SVM) mengendalikan kompleksitas sehingga generalisasi membaik.
2. Naive Bayes mengasumsikan fitur saling independen. Komponen PCA justru merupakan kombinasi linear yang saling berkorelasi, sehingga transformasi ini merusak asumsi dasar NB dan malah menurunkan performa meski parameter var_smoothing di-optimasi.

Trade-off Akurasi vs ROC AUC

1. Accuracy cuma mengukur “benar/salah” prediksi, sedangkan ROC AUC menilai kualitas ranking probabilitas.
2. KNN yang sudah menghasilkan label dengan benar (accuracy tinggi) sering kali kurang terkalibrasi pada skala probabilitas, sehingga AUC-nya lebih rendah meski akurasinya bagus.

Sensitivitas terhadap Dimensi

1. Sebagian besar model (LR, SVM) robust terhadap jumlah fitur yang besar; PCA memotong dimensi tanpa banyak kehilangan kapabilitas pemisahan.
2. KNN sangat sensitif pada metrik jarak:
    a. PCA mengubah struktur ruang jarak, sehingga tetangga terdekat bisa berubah; optimasi hyper-parameter (n_neighbors, weights, p) membantu memulihkan akurasi, namun estimasi confidence masih berbeda.

Overfitting vs Underfitting

1. Pada baseline tanpa regularisasi, model cenderung fit terlalu ketat pada fitur asli (terutama pada SVM/RF), yang bisa menurunkan generalisasi tetapi menerjemahkannya ke probabilitas yang “tegas”, meningkatkan AUC tapi tidak selalu akurasi.
2. Penambahan regularisasi menghaluskan decision boundary, menyeimbangkan bias–variance trade-off, sehingga LR dan SVM membaik pada kedua metrik