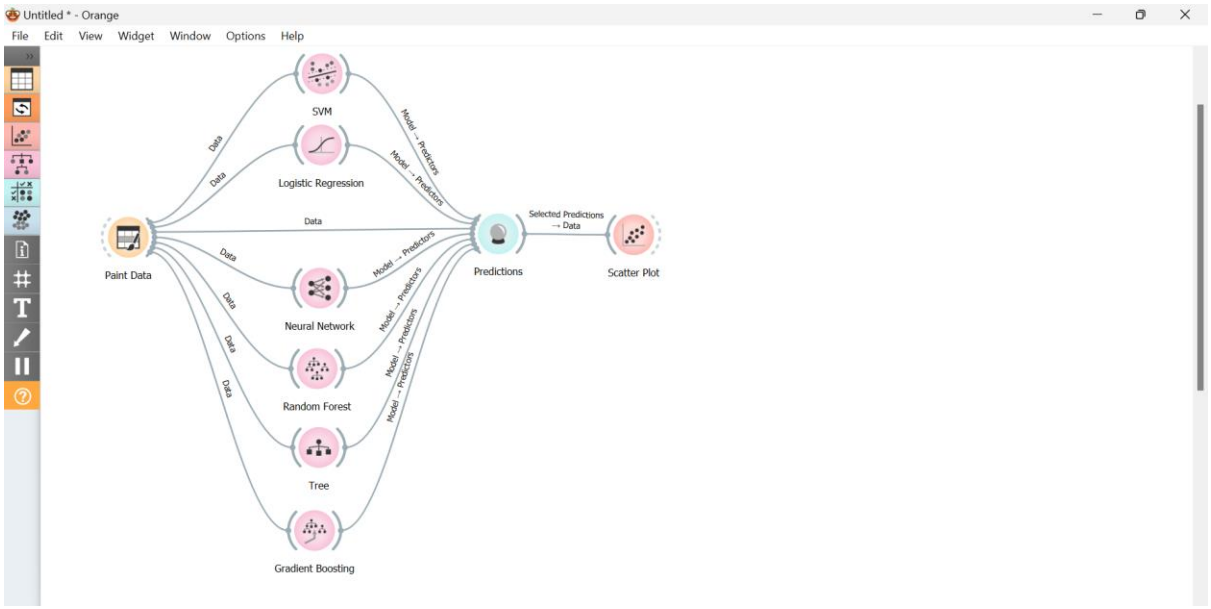
M.Asjaun

1103210181

Task 10



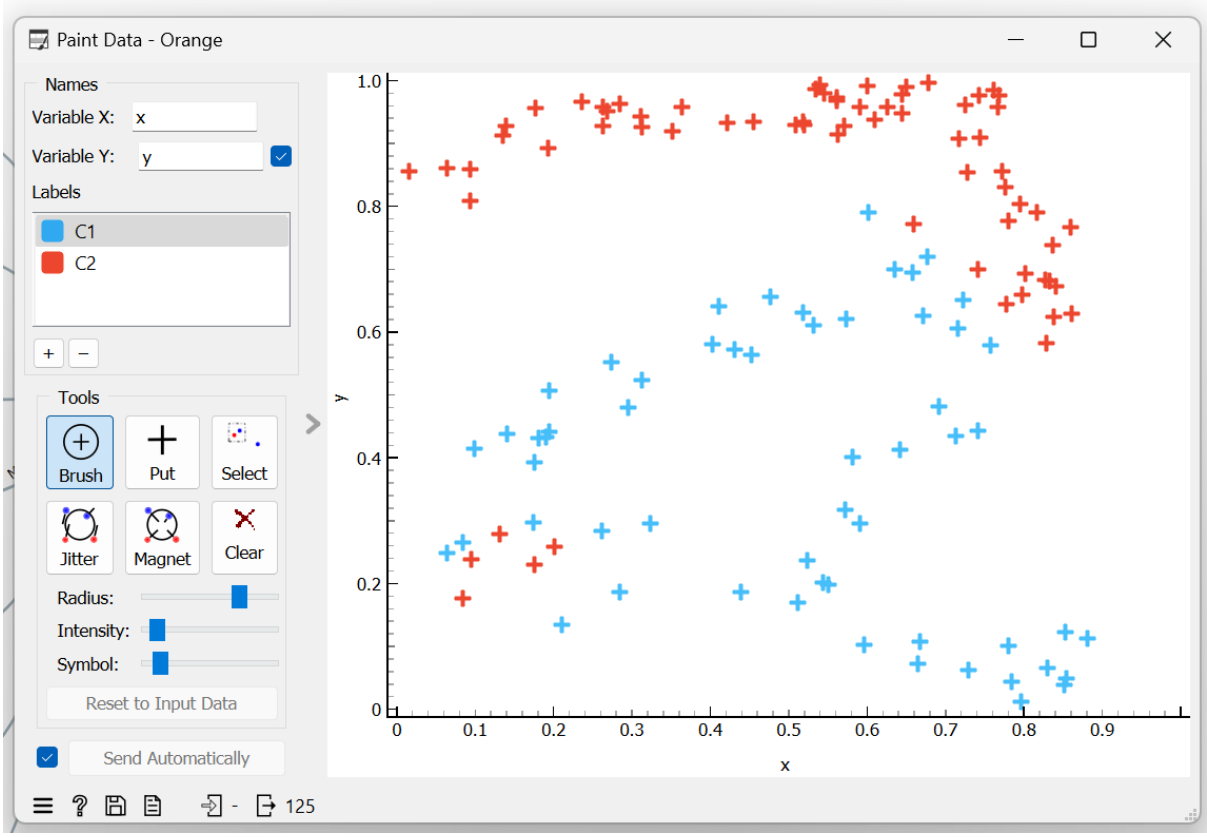
Workflow ini bertujuan untuk membandingkan performa beberapa model machine learning terhadap dataset tertentu yang dibuat melalui Paint Data.

Penjelasan Komponen:

1. Paint Data:
    - Merupakan modul awal yang digunakan untuk membuat atau memasukkan dataset secara manual.
    - Di sini, Anda bisa menggambar dataset dengan label tertentu untuk tujuan eksplorasi atau pengujian model.
2. Model yang Digunakan:
    - SVM (Support Vector Machine): Digunakan untuk klasifikasi berdasarkan margin maksimum.
    - Logistic Regression: Model regresi untuk klasifikasi biner atau multi-kelas.
    - Neural Network: Model berbasis jaringan saraf tiruan untuk menangani masalah kompleks.
    - Random Forest: Algoritme ensemble yang menggabungkan banyak decision tree untuk meningkatkan akurasi.
    - Tree: Algoritme decision tree sederhana.
    - Gradient Boosting: Model ensemble yang membangun pohon keputusan secara bertahap untuk mengurangi error.

Setiap model menerima data dari Paint Data dan mengolahnya untuk menghasilkan prediksi.

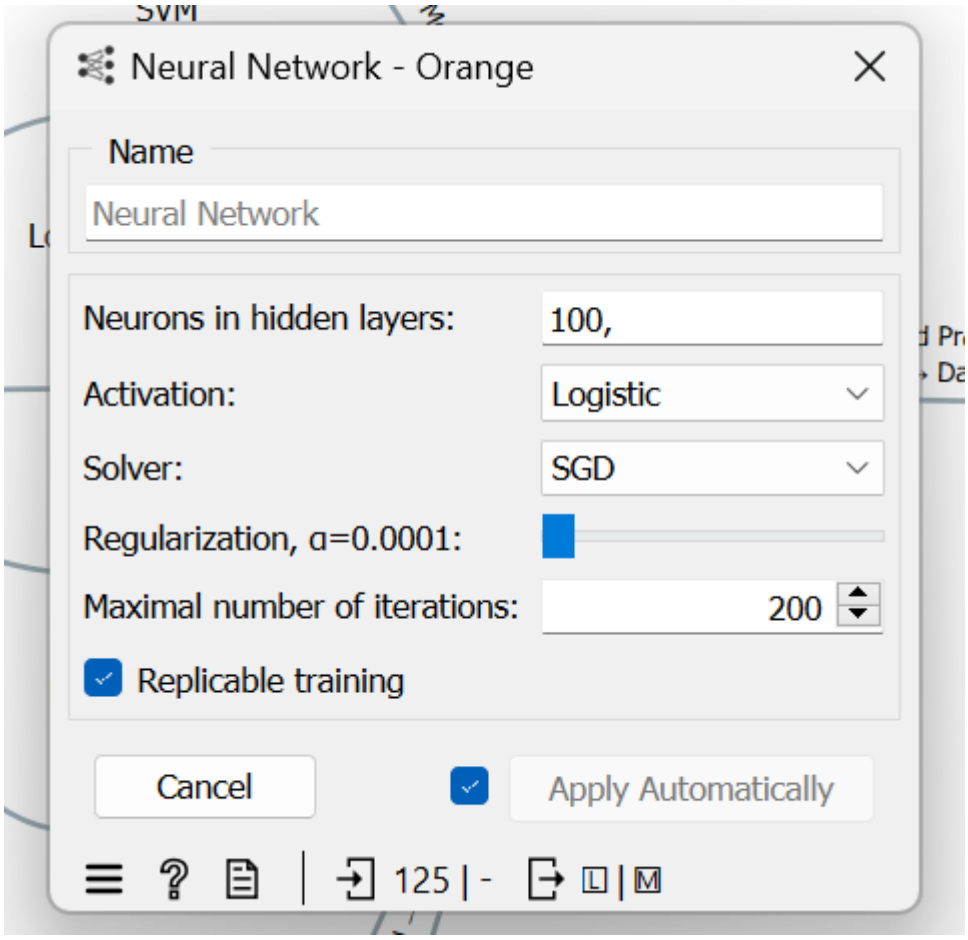
3. Predictions:
    - Modul ini mengumpulkan hasil prediksi dari semua model yang dihubungkan.
    - Prediksi dari berbagai model dapat dibandingkan untuk analisis lebih lanjut.
4. Scatter Plot:
    - Modul ini digunakan untuk memvisualisasikan prediksi yang dihasilkan oleh model dalam bentuk scatter plot.
    - Scatter plot membantu dalam memahami distribusi prediksi dan membandingkan hasil dari berbagai model.



Gambar tersebut adalah tampilan modul Paint Data di Orange, yang digunakan untuk membuat dataset sintetis secara manual. Pada sumbu horizontal (X) dan vertikal (Y), Anda dapat menggambar titik-titik data dengan mengelompokkan mereka ke dalam kelas yang berbeda. Pada gambar ini:

1. Titik-titik data diwakili oleh simbol silang (+) dengan warna yang berbeda untuk masing-masing kelas:
    - Warna biru (C1) menunjukkan satu kelas.
    - Warna merah (C2) menunjukkan kelas lainnya.
2. Pengelompokan data:
    - Titik-titik merah sebagian besar berada di area atas grafik, menunjukkan distribusi data untuk kelas C2.

- Titik-titik biru tersebar di sebagian besar area, dengan beberapa data mendekati kelompok merah.

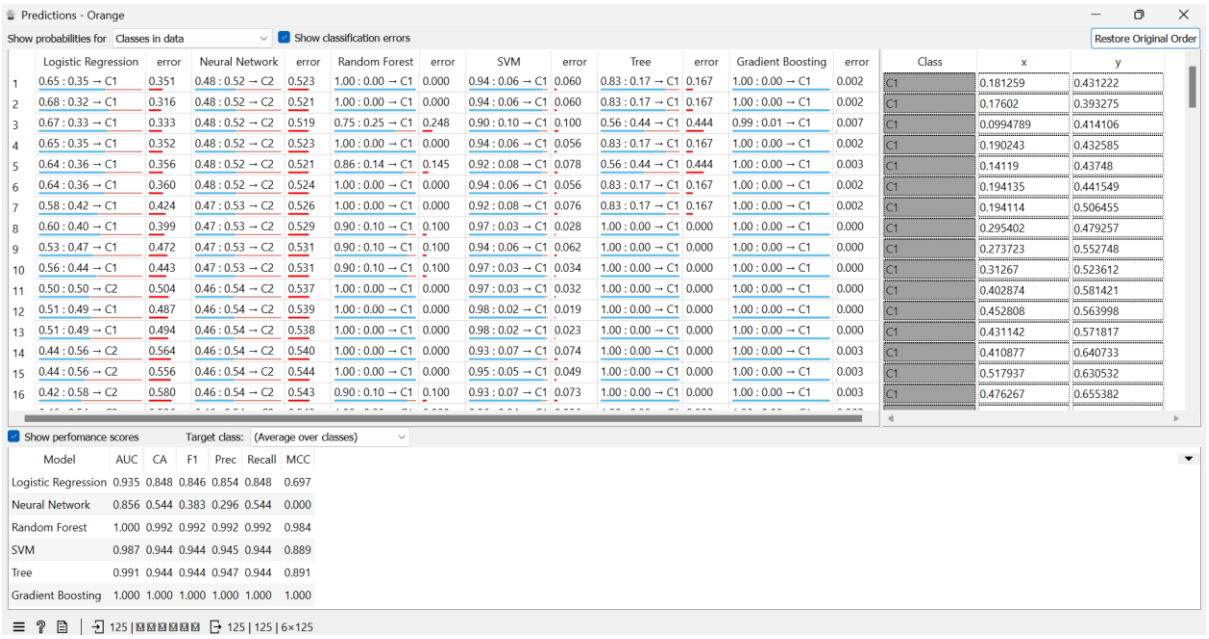


Gambar di atas menunjukkan konfigurasi Neural Network yang digunakan. Berikut adalah penjelasan dari setiap parameter:

1. Neurons in hidden layers:
    - Menggunakan 100 neuron dalam lapisan tersembunyi. Ini berarti jaringan neural memiliki satu lapisan tersembunyi dengan 100 neuron, yang dapat menangkap pola kompleks dari data.
2. Activation:
    - Fungsi aktivasi yang digunakan adalah Logistic (Sigmoid). Fungsi ini sering digunakan untuk tugas klasifikasi biner karena menghasilkan output antara 0 dan 1, merepresentasikan probabilitas.
3. Solver:
    - Algoritma optimisasi yang digunakan adalah SGD (Stochastic Gradient Descent). Ini adalah algoritma optimisasi yang sederhana namun efektif untuk pembaruan parameter, terutama untuk dataset besar.
4. Regularization (α=0.0001):

- Regularisasi digunakan untuk mencegah overfitting dengan menambahkan penalti terhadap bobot besar dalam jaringan. Parameter regularisasi α=0.0001 menunjukkan penalti yang sangat kecil, memberikan kontrol ringan terhadap overfitting.

5. Maximal number of iterations:
    - Jaringan neural akan dilatih hingga maksimum 200 iterasi atau sampai konvergensi tercapai lebih awal. Ini membatasi jumlah waktu yang dihabiskan untuk pelatihan.
6. Replicable training:
    - Opsi ini diaktifkan, yang memastikan bahwa hasil pelatihan dapat direplikasi. Ini dilakukan dengan mengatur seed random number generator, sehingga pelatihan menjadi deterministik.



Gambar di atas menunjukkan hasil prediksi dan evaluasi model menggunakan Orange, yang mencakup berbagai algoritma machine learning seperti Logistic Regression, Neural Network, Random Forest, SVM, Decision Tree, dan Gradient Boosting. Berikut penjelasannya:

1. Hasil Prediksi:
    - Pada bagian atas, tabel menampilkan prediksi masing-masing model untuk setiap instance data. Kolom pertama menunjukkan probabilitas kelas yang diprediksi oleh model, sedangkan kolom kedua menunjukkan error (jika prediksi salah).
    - Prediksi yang benar ditandai dengan nilai error 0.000, sementara error diwakili oleh nilai probabilitas yang salah (highlight merah).
2. Kinerja Model:

- Pada bagian bawah, terdapat tabel evaluasi performa model dengan metrik seperti AUC, Accuracy, F1 Score, Precision, Recall, dan Matthews Correlation Coefficient (MCC).
- Model Gradient Boosting memiliki performa terbaik dengan nilai AUC, Accuracy, dan metrik lainnya sebesar 1.000, menunjukkan bahwa model ini memberikan prediksi sempurna untuk dataset ini.
- Model Random Forest dan SVM juga menunjukkan performa yang sangat tinggi dengan AUC sekitar 0.987 - 0.992.
- Neural Network memiliki performa terendah dibanding model lainnya, dengan AUC sebesar 0.856 dan Accuracy lebih rendah.

3. Analisis Kesalahan:
   - Dengan melihat nilai error di setiap baris data, kita bisa menganalisis area di mana model membuat kesalahan, seperti prediksi Neural Network yang cenderung lebih banyak mengalami error dibanding model lainnya.
   - Model Gradient Boosting dan Random Forest menunjukkan error yang sangat kecil, hampir mendekati nol, membuatnya sangat andal untuk dataset ini.
4. Kesimpulan:
   - Gradient Boosting adalah model terbaik untuk dataset ini, dengan performa sempurna di semua metrik evaluasi.
   - Model sederhana seperti Logistic Regression memiliki performa yang cukup baik, tetapi tidak seunggul Gradient Boosting atau Random Forest.
   - Model dengan akurasi dan AUC yang lebih rendah, seperti Neural Network, dapat ditingkatkan dengan tuning lebih lanjut, seperti menyesuaikan hyperparameter atau struktur jaringan.