LAPORAN TUGAS BESAR MACHINE LEARNING

# **CLASSIFICATION**



Disusun oleh :

Adabi Raihan Muhammad 1301180379

M Wanda Wibisono   1301184339

Telkom University

Bandung

2021

1. Formulasi Masalah

Tahap kedua tugas besar machine learning ini, kami melakukan classification dari dataset salju yang sudah disiapkan oleh dosen agar dapat memprediksi hari ini dan besok akan turun salju atau tidak.

Dengan model classification ini diharapkan kedepannya dapat memprediksi tidak hanya hari ini atau besok aja, tetapi bisa memprediksi hingga satu bulan kedepan bahkan hingga satu tahun kedepan, agar masyarakat dapat mempersiapkan diri dalam menghadapi cuaca bersalju dari jauh hari.

1. Eksplorasi dan Persiapan Data

* Data Eksploration
* Jumlah baris dalam dataset salju berjumlah 127277, dengan 24 kolom.
* Dari 24 kolom(variabel), terbagi menjadi 16 variabel bertipe float, 1 variabel bertipe integer, dan bertipe object berjumlah 7.
* Float

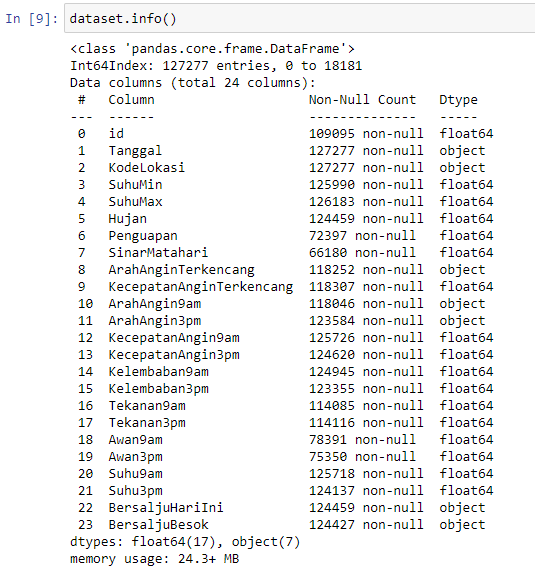
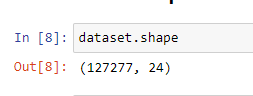
SuhuMin, SuhuMax, Hujan, Penguapan, SinarMatahari, Kecepatan angin Terkencang, Kecepatan Angin 9am, Kecepatan Angin 3pm, Kelembaban 9am, kelembaban 3pm, Tekanan 9am, Tekanan 3pm, Awan 9am, Awan 3am, Suhu 9am, Suhu 3am

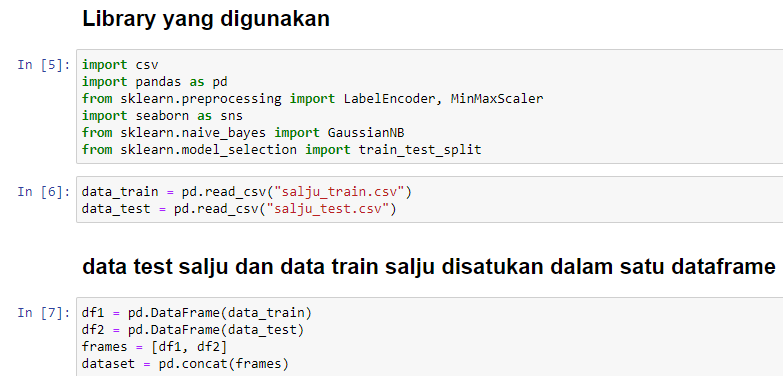
* Integer

Id

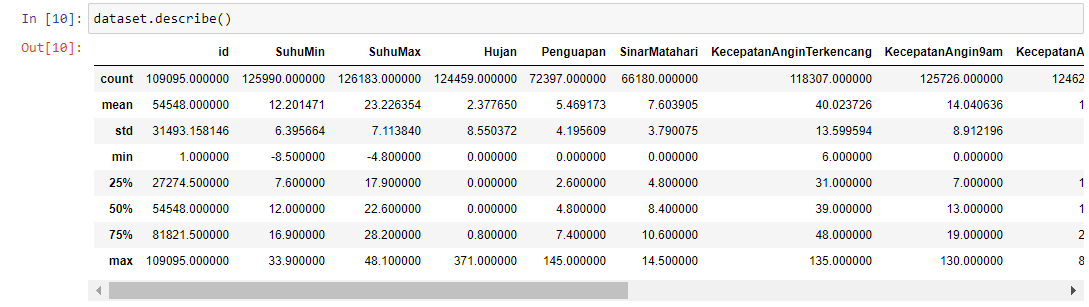
* Object

Tanggal, KodeLokasi, Arah Angin Terkencang, Arah Angin 9am, Arah Angin 3pm, Bersalju Hari ini, Bersalju besok.

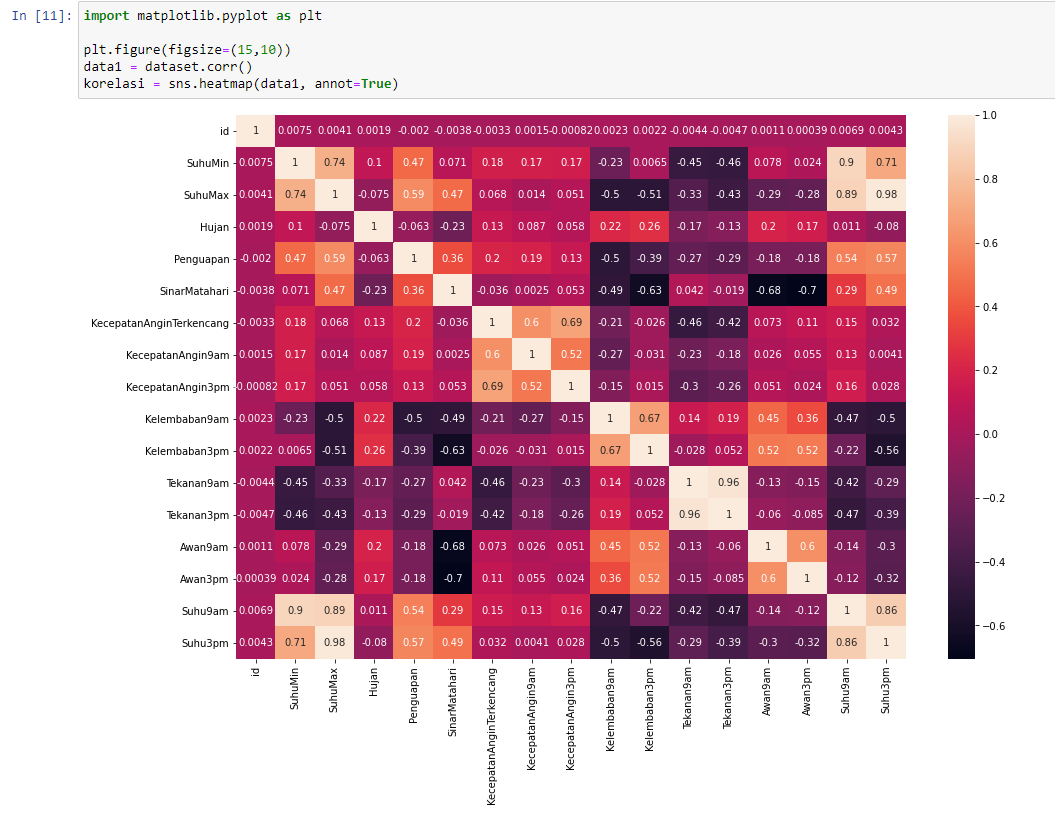




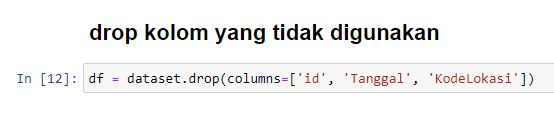
* Simple Statistics (mean, standard deviasi, min, max, count)



* Korelasi keseluruhan data

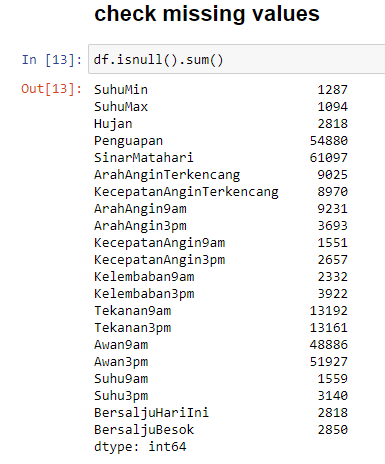


Dari grafik heatmap diatas, dapat disimpulkan bahwa kolom id, tanggal, dan kode lokasi memiliki nilai korelasi yang sangat kecil dengan kolom lainnya, maka dari itu kami melakukan drop kolom terhadap kolom id, tanggal, dan kode lokasi.

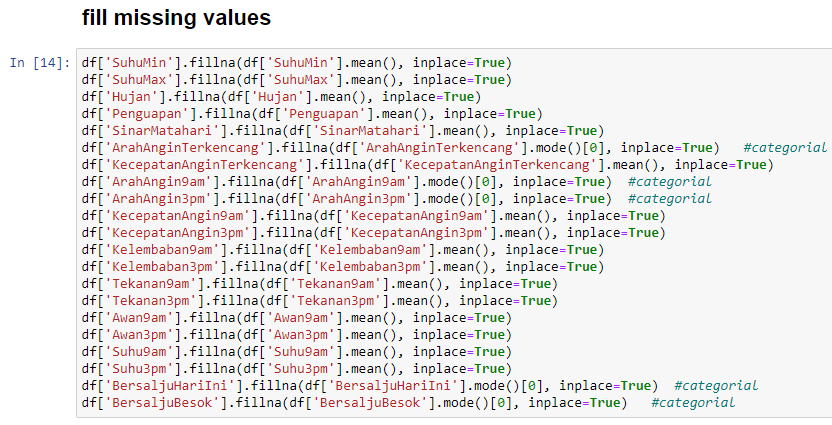


* Data cleansing

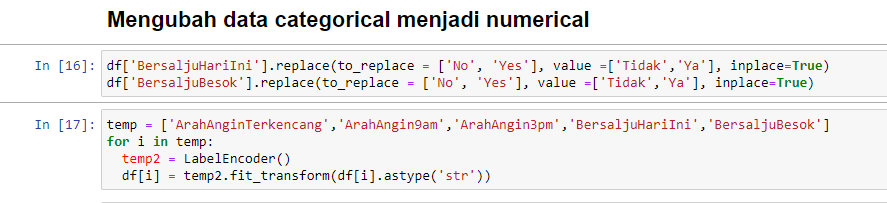
Pada tahap data cleansing, pertama saya akan melakukan check missing values



Untuk menangani missing values, kami mengisi missing values dengan dua acara, pertama untuk kolom yang isinya numerical akan diisi dengan nilai mean dari kolom tersebut, sedangkan untuk kolom categorial akan diisi dengan nilai modus dari kolom tersebut.

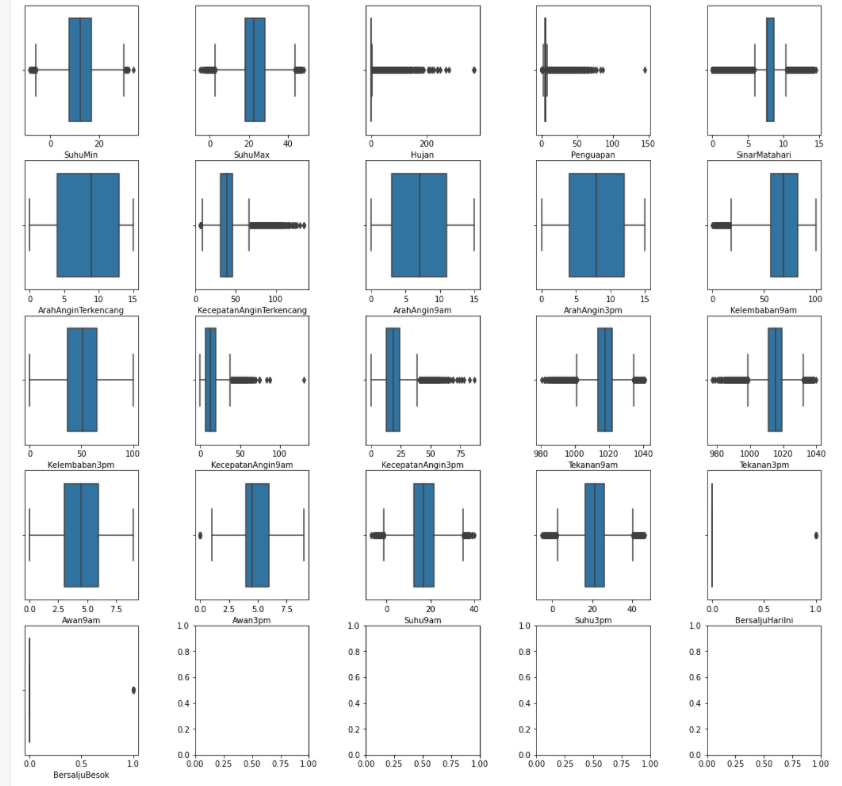


* Selanjutnya dilakukan pengubahan data categorical menjadi numerical



* Setelah itu kami melakukan check data outliers menggunakan boxplot

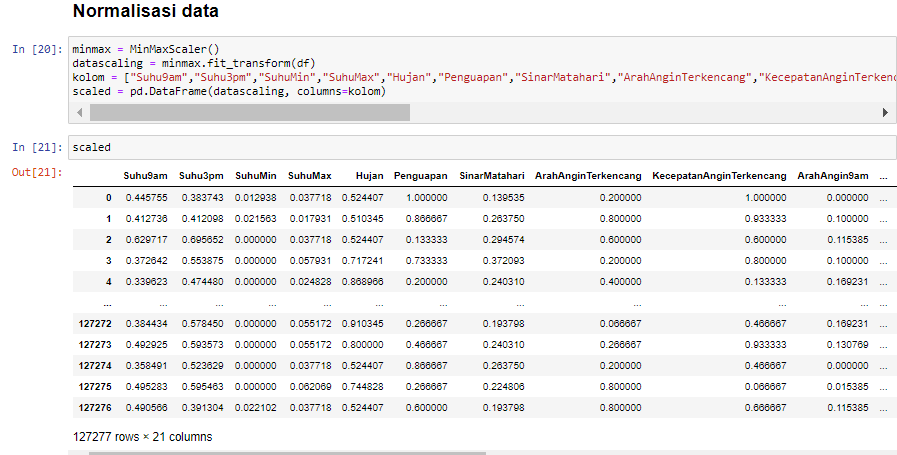




Hasil dari grafik menunjukan di dalam dataset terdapat data outliers, , kami memutuskan untuk tidak akan menghapus data outliers, karena menurut kami data tersebut masih masuk akal, contohnya ada data suhu max yang melebihi data interquartile range, karna mungkin saja suhu itu tinggi karena matahari sedang berada di titik terdekat dengan bumi.

* Feature Engineering
* Scaling Method

Scaling method adalah metode untuk menormalisasikan data agar semua data berada dalam skala yang sama. Pada kasus ini kami menggunakan Normalisasi MinMax menjadikan rentang nilai dari 0 hingga 1.

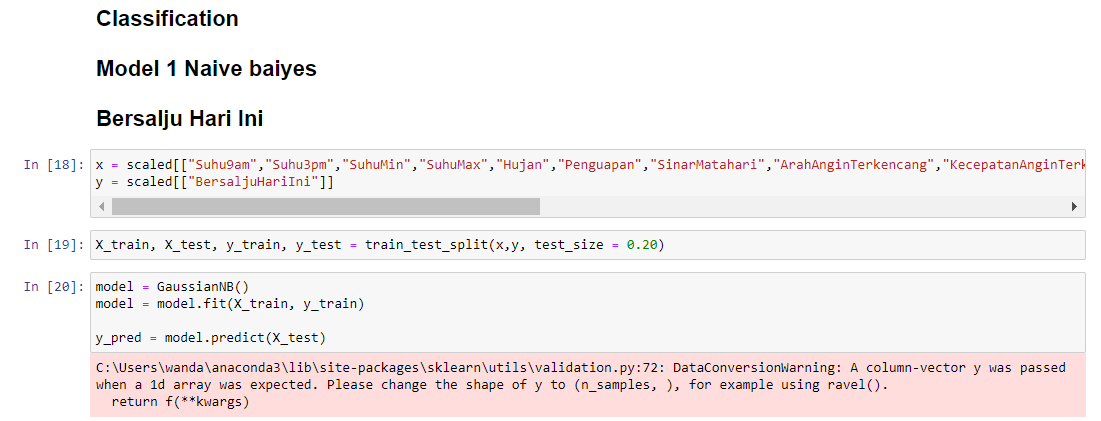


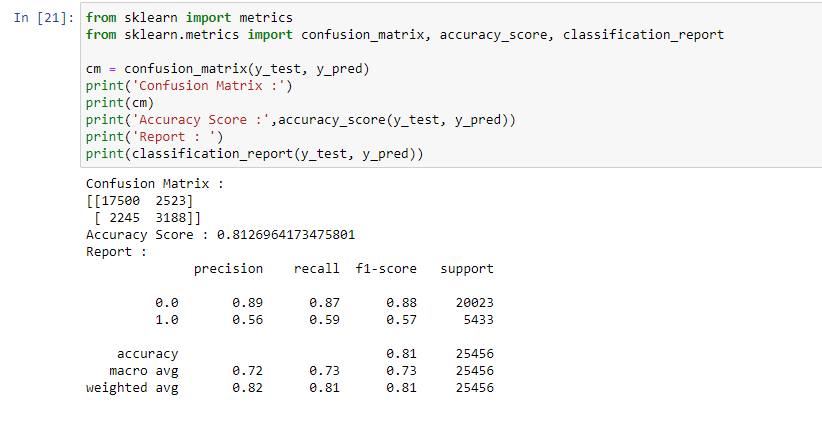
* Splitting data

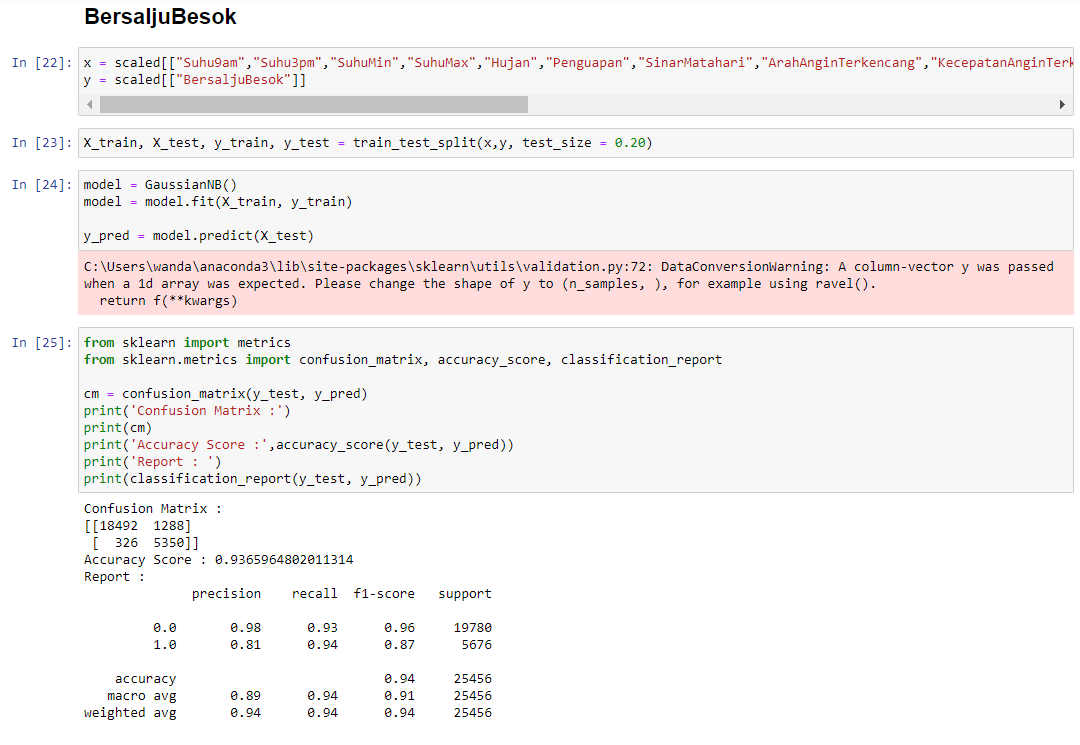
Untuk splitting data kami menggunakan perbandingan rasio 80:20, karena pada dasarnya tidak ada perbandingan rasio yang paling bagus, tergantung dari banyaknya jumlah data yang kita miliki, karena pada dataset salju ini tergolong masih tidak terlalu banyak datanya, maka menurut kami perbandingan rasio 80:20 sudah memberikan akurasi yang baik.

1. Pemodelan Classification

Untuk tugas classification ini kami diperbolehkan untuk menggunakan library dari python, untuk model pertama kami menggunakan algoritma naive baiyes dan untuk model kedua buat eksperimen kami menggunakan algoritma ID3.



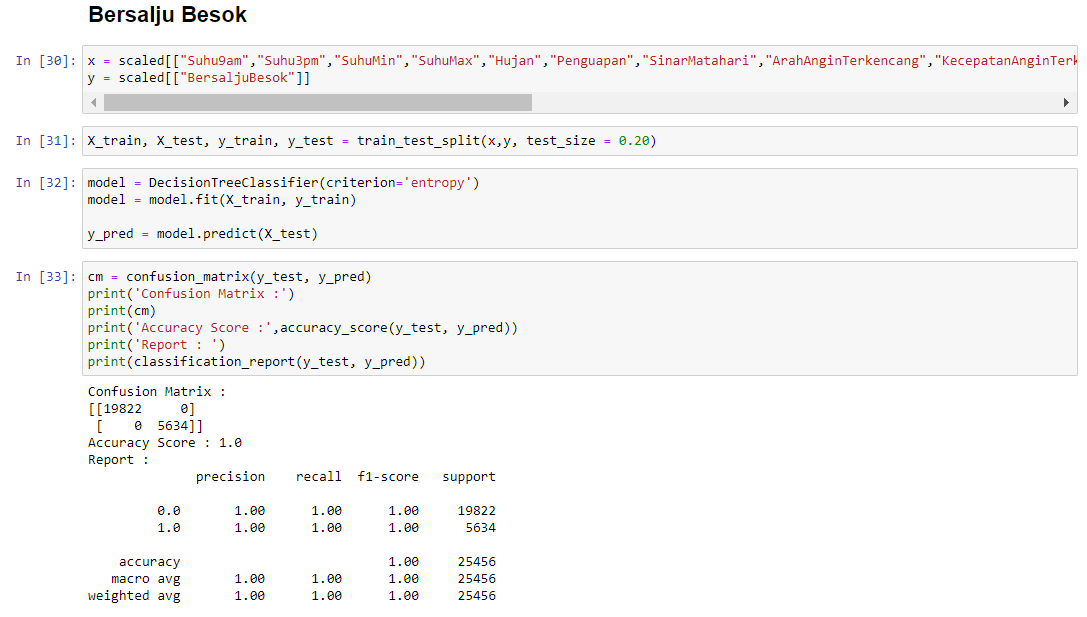




1. Eksperimen

Untuk eksperimen kami menggunakan dataset yang sama dengan yang dilakukan terhadap algoritma naive baiyes, tetapi menggunakan algoritma ID3.





1. Evaluasi

Setelah tahapan pemodelan, selanjutnya masuk tahapan terakhir yaitu evaluasi, diperuntukan melihat hasil akurasi dari pemodelan yang telah dilakukan. Evaluasi yang dilakukan menggunakan confusion matrix, di dalamnya menggunakan akurasi precision, recall, dan f1-score. Precision untuk memprediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Recall untuk memprediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Dan f1-score untuk membandingkan rata-rata precision dan recall yang dibobotkan.

1. Kesimpulan

Dari hasil akurasi yang telah dilakukan, dapat disimpulkan atribut BersaljuHariIni dengan algoritma naive baiyes lebih tinggi tingkat akurasinya dibandingkan dengan algoritma ID3, sedangkan untuk atribut BersaljuBesok untuk algoritma naive baiyes lebih rendah tingkat akurasinya dibandingkan dengan algoritma ID3. maka pemilihan algoritma classification yang digunakan dapat mempengaruhi hasil tingkat akurasi.

Referensi:

<https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8>

<https://www.researchgate.net/post/can_someone_recommend_what_is_the_best_percent_of_divided_the_training_data_and_testing_data_in_neural_network_7525_or_8020_or_9010>