Demo Day Final Report

**Machine Learning with TensorFlow Training**

**Professional Academy Digital Talent Scholarship 2022**

|  |  |
| --- | --- |
| **Group Number** | **HC\_3** |
| **Dataset** | **Healthcare** |
| **Name - DTS ID** | 1. **Josua Geovani Pinem - 152236035100-236** 2. **M Alana- 152236035100-700** 3. **M. Salman Al-Faridzi-152236035101-336** |

**Selected Theme**:

Ensemble Machine Learning Approach: Case Study Healthcare

**Title of the Project**:

Boosting Machine Learning Prediction by Averaging Ensemble of Learning Model

**Background Summary:**

Kondisi dan posisi data pada saat ini dapat memberikan keuntungan hingga nilai tambah pada semua aspek kehidupan. Hal ini dapat disimak secara langsung dari berbagai aspek terlebih pada bidang kesehatan. Pemanfaatan semua data yang pada bidang kesehatan diharapkan dapat meningkatkan kualitas kesehatan di Indonesia. Berdasarkan data Kementrian Kesehatan jumlah tenaga medis di Indonesia mencapai 124.449 orang pada 2020[1]. Namun terjadi kesenjangan terhadap distribusi tenaga medis pada provinsi tertentu. Hal ini dapat menyebabkan keterlambatan dalam pengecekan kondisi pasien, terlebih apabila kondisi pasien dalam kondisi yang membahayakan. Kondisi ketika tempat berobat penuh juga menjadi hal yang perlu dipertimbangkan.

Permasalahan ini membutuhkan pengambilan keputusan yang cepat, dalam hal ini tenaga medis tentunya memerlukan pengalaman untuk dapat mengambil keputusan. Rekam jejak dan pengalaman ini kemudian menjadi data yang kemudian menjadi informasi yang dapat digunakan untuk membuat sistem yang dapat mengambil keputusan. Sistem yang dapat mengambil keputusan dapat direalisasikan dengan menggunakan kecerdasan buatan. Kecerdasan buatan dalam kasus ini dapat menggunakan machine learning hingga deep learning apabila data tersedia dalam jumlah yang sangat besar.

Untuk dapat menyelesaikan permasalahan ini maka perlu dikembangkan model pembelajaran mesin yang dapat bekerja dengan sangat baik hingga kesalahan prediksi terhadap penyakit maupun kondisi pasien dapat diminimalisir. Hal ini tentunya menjadi tujuan utama ketika ingin meningkatkan kualitas kesehatan.

Berdasarkan dataset yang dimiliki serta tujuan dan arah kedepannya agar memberikan hasil yang maksimal maka dilakukan pendefinisian terhadap masalah yang ada.

* + - 1. Apa pendekatan yang dilakukan berdasarkan dataset yang dimiliki? Apakah cukup dengan pendekatan Machine Learning atau hingga Deep Learning?
      2. Bagaimana melakukan Exploratory Data Analysis pada data Healthcare yang terdiri dari beberapa kelas yang akan digunakan sebagai subjek penelitian?
      3. Bagaimana membuat model yang dapat menyesuaikan permasalahan yang dihasilkan dari Exploratory Data Analysis yang kemudian memberikan hasil yang efektif dan efisien?

Dari sini kemudian diajukan sebuah pengembangan model pembelajaran yang belajar berdasarkan model-model lain. Model-model ini kemudian digabungkan yang kemudian berjalan bersama-sama dan memberikan hasil terbaik berdasarkan voting dari model lain. Adapun model yang dapat digunakan beragam dimulai dengan teknik tree-based machine learning, hingga deep learning dengan Multi-Layer Perceptron hingga CNN [2][3][4][5].

**Please specifically mention what you’ve done:**

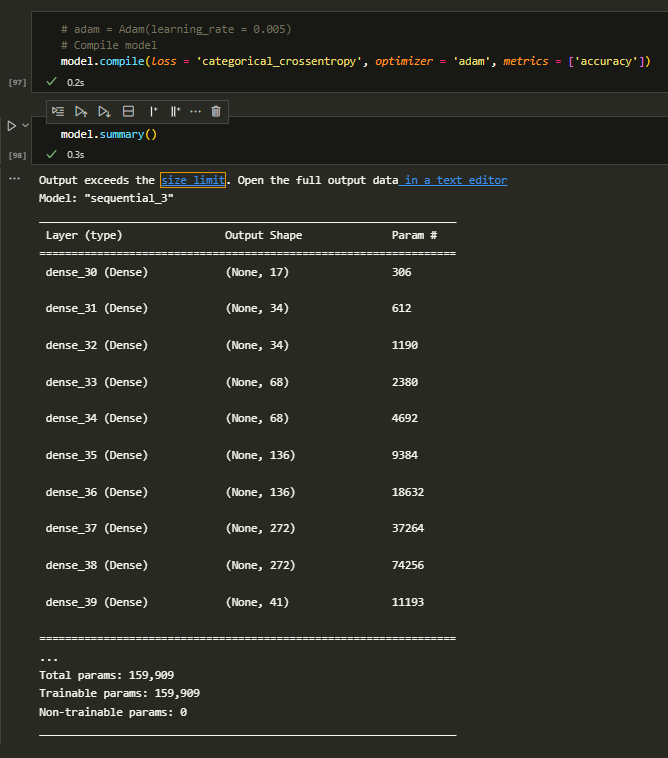
1. **Josua Geovani Pinem - 152236035100-236 :** Membuat proposal, mencari referensi jurnal, dan men-tuning hyperparameter dari model yang sudah dibuat. Hasil tuning kemudian digunakan untuk 3 model baru dan digabung dengan ensemble.
2. **M Alana - 152236035100-700 :** Melakukan pencarian referensi jurnal, melakukan pencarian referensi Code di github dan Kaggle, melakukan preprocessing data, dan membuat Model machine learning utama menggunakan Tensorflow dengan Fully-Connected Layer yang dikomparasikan dengan model yang menggunakan GradientBoostingClassifier (GBC).
3. **M. Salman Al-Faridzi - 152236035101-336 :** Mencari referensi jurnal dan code, Melakukan pembersihan data, membuat model dengan algoritma Support Vector Machine (SVM) yang kemudian dikomparasikan dengan model XGBoost

**Screenshot Output Model :**

****

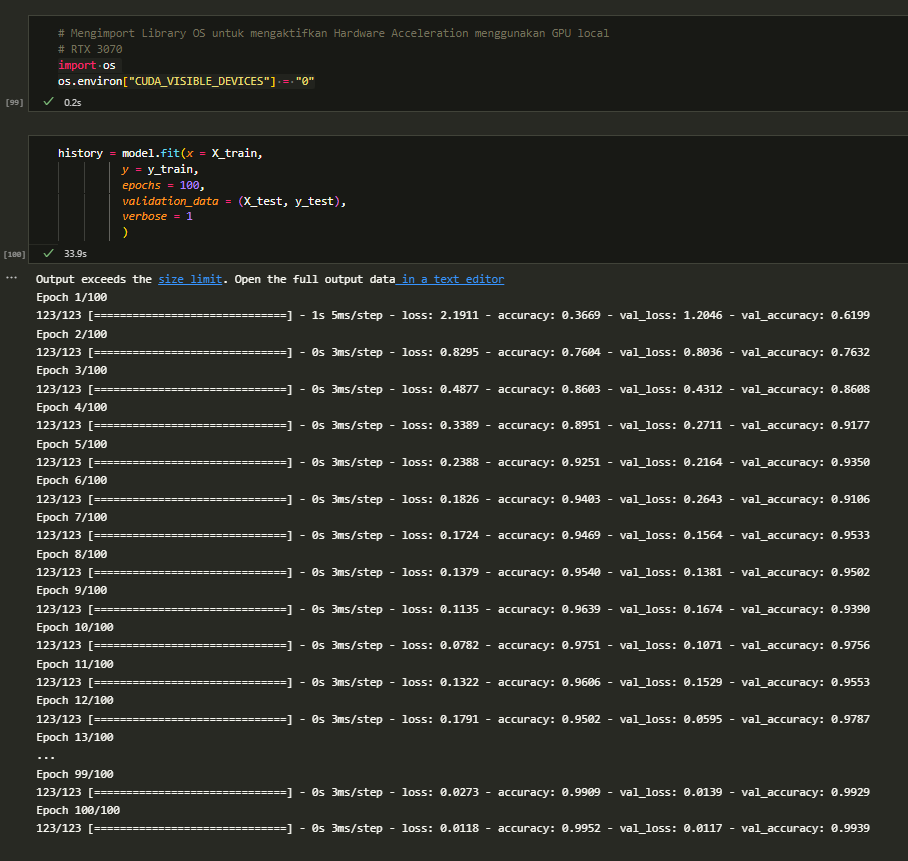
Gambar 1. Arsitektur Model

Struktur model yang di gunakan adalah 1 *input* *layer* dengan input dimensi sebesar 17 dan aktivasi ‘relu’, kemudian 8 *fully-connected layers* dengan aktivasi ‘relu’, dan satu *output layer* dengan aktivasi ‘sigmoid’.

****

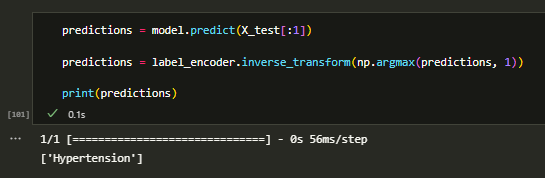
Gambar 2. Arsitektur Model 2

Parameter compiling model menggunakan *loss function* “categorical\_crossentropy”, *optimizer* “adam”, dan *metrics* “accuracy”.

****

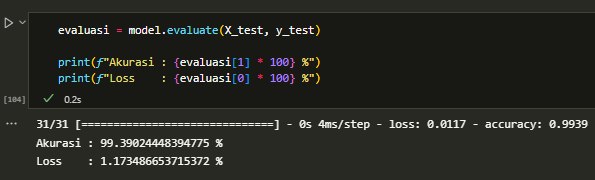
Gambar 3. Model Fit

Proses training dilakukan dengan menggunakan GPU lokal dengan *epoch* sebesar 100 dan validation data menggunakan X\_test dan y\_test.

****

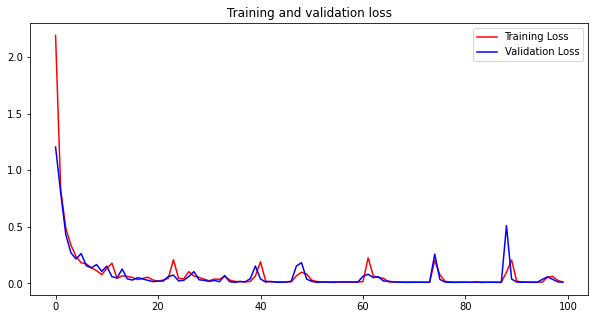
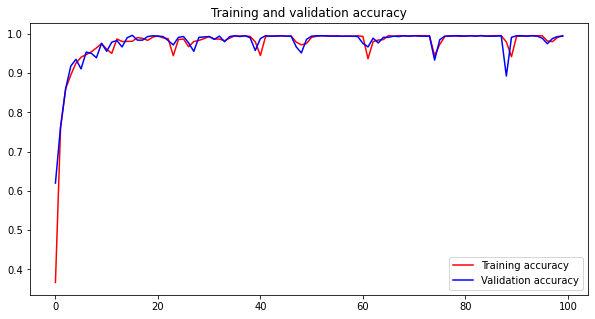
Gambar 4. Uji coba prediksi

Kemudian melakukan proses testing model menggunakan data X\_test.

****

Gambar 5. Evaluasi ke semua data

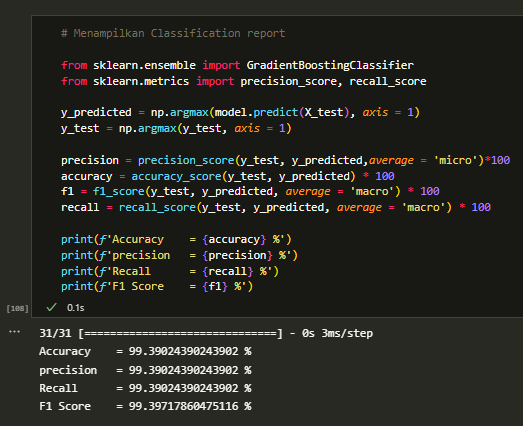
Mengevaluasi model menggunakan data X\_test dan y\_test, setelah dilakukan evaluasi, hasil akhirnya menunjukkan akurasi sebesar 99,39% dan loss sebesar 1,17%.

****

Grafik 1. (a) Perbandingan akurasi pada tahapan training dan validasi

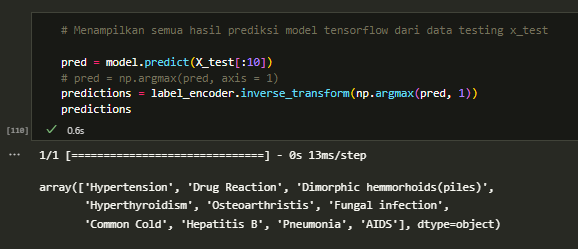
(b) Perbandingan akurasi pada tahapan training dan validasi

Hasil dari plotting grafik Train Accuracy, Validation Accuracy, Train Loss, dan Validation Loss ditampilkan pada Grafik 1(a) dan 1(b).

****

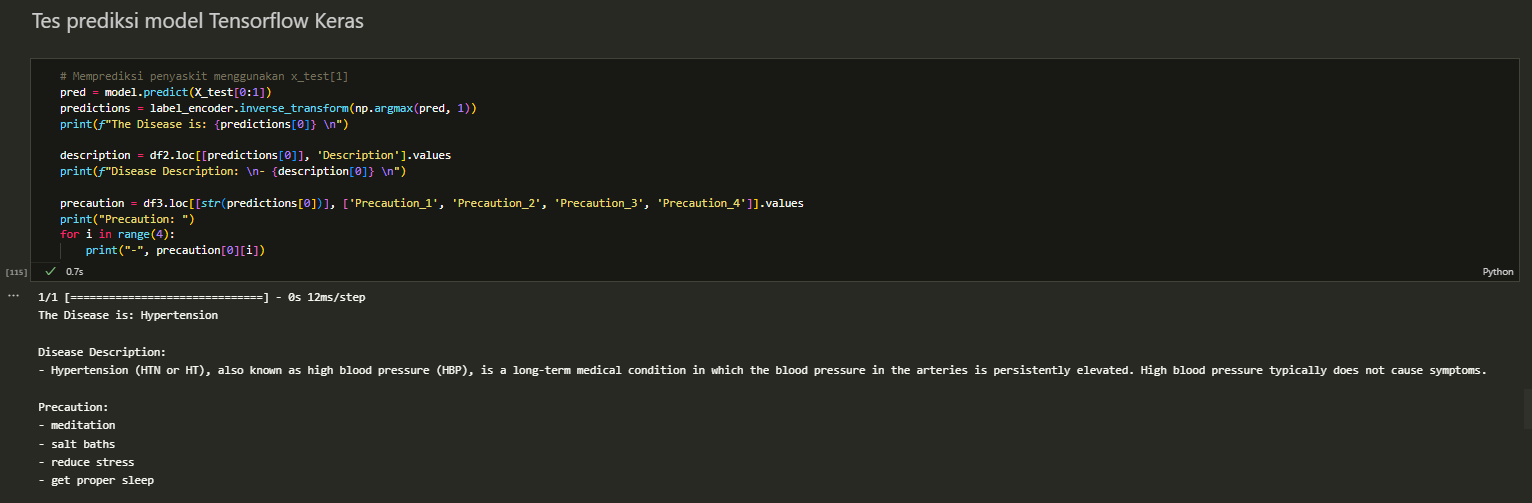
Gambar 6. Hasil Evaluasi

Menampilkan *metrics Classification Report* untuk menampilkan hasil *Accuracy* sebesar 99.39%, *Precission* sebesar 99.39%, *Recall* sebesar 99.39%, dan *F1 Score* sebesar 99.39%.

****

Gambar 7. Uji coba prediksi

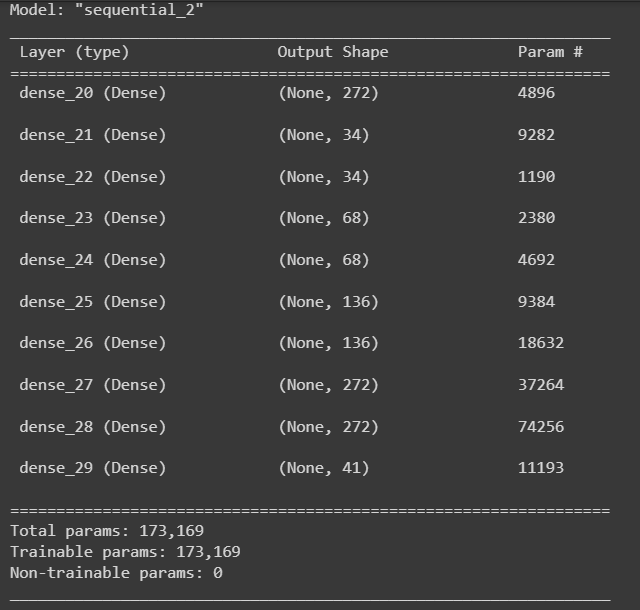
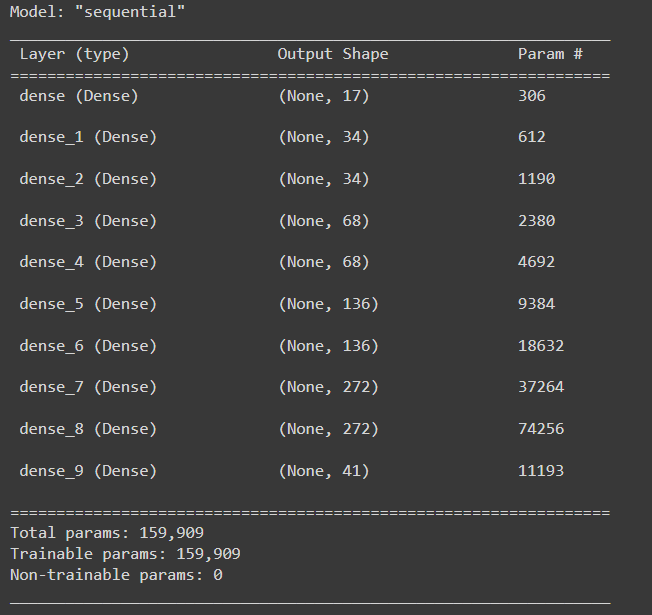
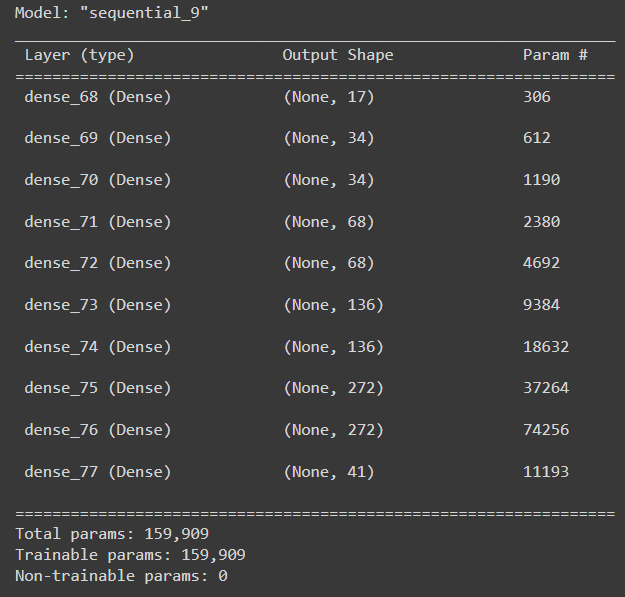
Menampilkan hasil prediksi menggunakan 10 data pertama dari X\_test.

****

Gambar 8. Tes Prediksi

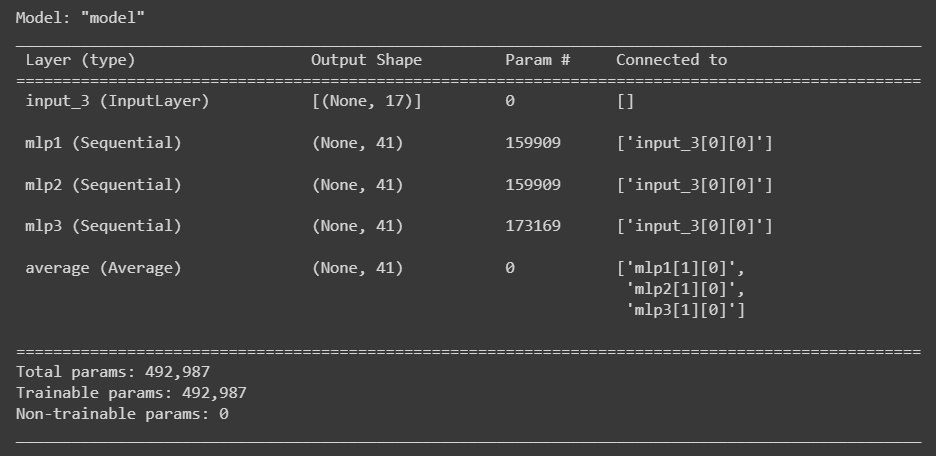
Menampilkan hasil final dari prediksi model yang juga menggunakan dataset symptom\_Description.csv dan symptom\_precaution.csv untuk menampilkan deskripsi penyakit yang diprediksi dan menampilkan pencegahan-pencegahan yang diperlukan.

**Ensemble Approach**

Kemudian berdasarkan model ini dibuat model lain yang diharapkan dapat menutupi kekurangan yang dimiliki model awal. Tujuannya ialah agar data dengan nilai false positive dan true negative dapat diminimalisir. Adapun desain untuk yang digunakan ialah sebagai berikut:

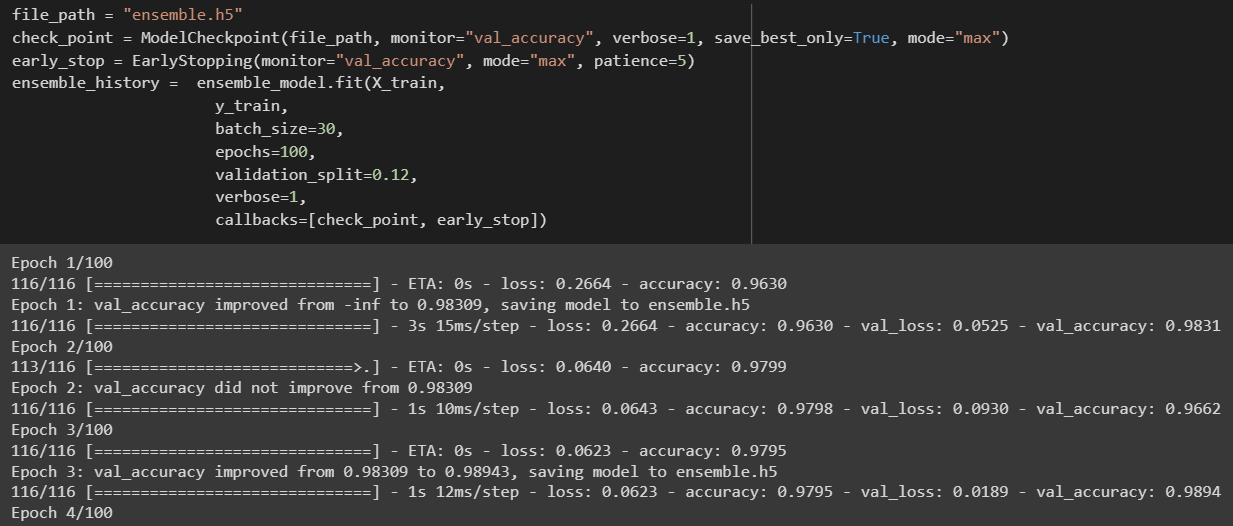
Gambar 9. (a)MLP (b)MLP2 (c)MLP3

Ketiga model kemudian digabung dengan average layer menjadi:



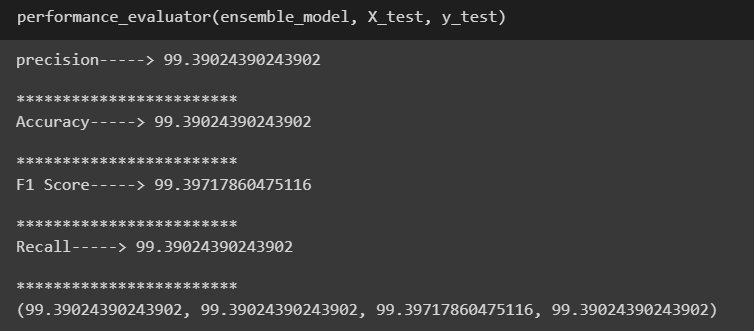
Gambar 10. Ensemble

Dan diperoleh keluaran sebagai berikut:



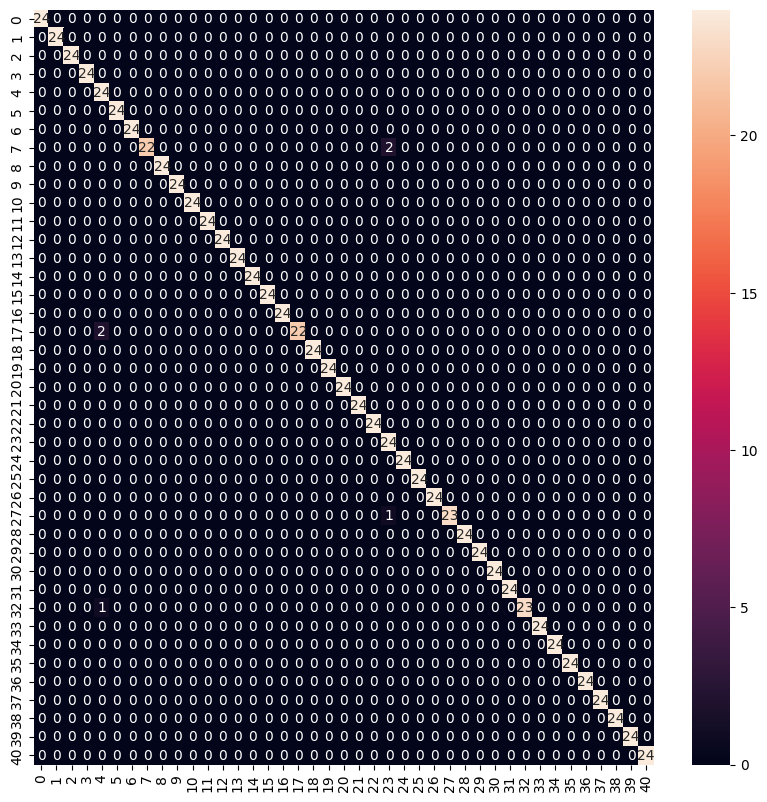
Gambar 11. Ensemble Fit

Proses berjalan menggunakan callback dengan patience 5, apabila tidak terjadi perubahan dari nilai max proses iterasi berhenti. Dapat diperhatikan epoch satu langsung memperoleh nilai akurasi yang tinggi 96%. Adapun hasil akhir dari proses ensemble pada data test ialah sebagai berikut.



Gambar 12. Evaluasi ensemble

Dengan confusion matrix sebagai berikut:



Gambar 13. Confusion Matrix Ensemble Learning

**Github Repo Link:**

[Disease Symptom Prediction](https://github.com/ARRARIAKU2/Disease-Symptom-Prediction)

**10-Min Video Presentation Link:**

**Quotes: *Jagalah kesehatanmu, jangan sampai kebahagiaan dirimu direnggut oleh penyakit tetanggamu!!***

**Project Reference:**

**Dataset**: [Healthcare](https://drive.google.com/drive/folders/1dhF8lX6yY4TpOQC6Z_3mt36Zvb3DE_7q)

[1] Kemenkes RI. 2021. Profil Kesehatan Indonesia Tahun 2020. Kementrian Kesehatan Republik Indonesia. Diakses online pada 19 Juli 2022. [Available]

<https://pusdatin.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Profil-Kesehatan-Indonesia-Tahun-2020.pdf>

[2] Dietterich, Thomas G. 2002. Ensemble Learning. Oregon State Universty. Diakses online pada 20 Juli 2022. [Available] <https://courses.cs.washington.edu/courses/cse446/12wi/tgd-ensembles.pdf>

[3] Yongquan Yang, Haijun Lv, Ning Chen. 2022. A Survey on Ensemble Learning under the era of Deep Learning. ACM. Diakses Online pada 20 Juli 2022 [Available] <https://arxiv.org/ftp/arxiv/papers/2101/2101.08387.pdf>

[4] Fionn Murtagh. 1991. Multilayer Perceptrons for Classification and Regression. Neurocomputing, Volume 2, Issues 5-6 July 1991 Pages 183-197.

[5] Peter Ghavami. 2019. Big Data Analytics Methods Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing. [e-book]

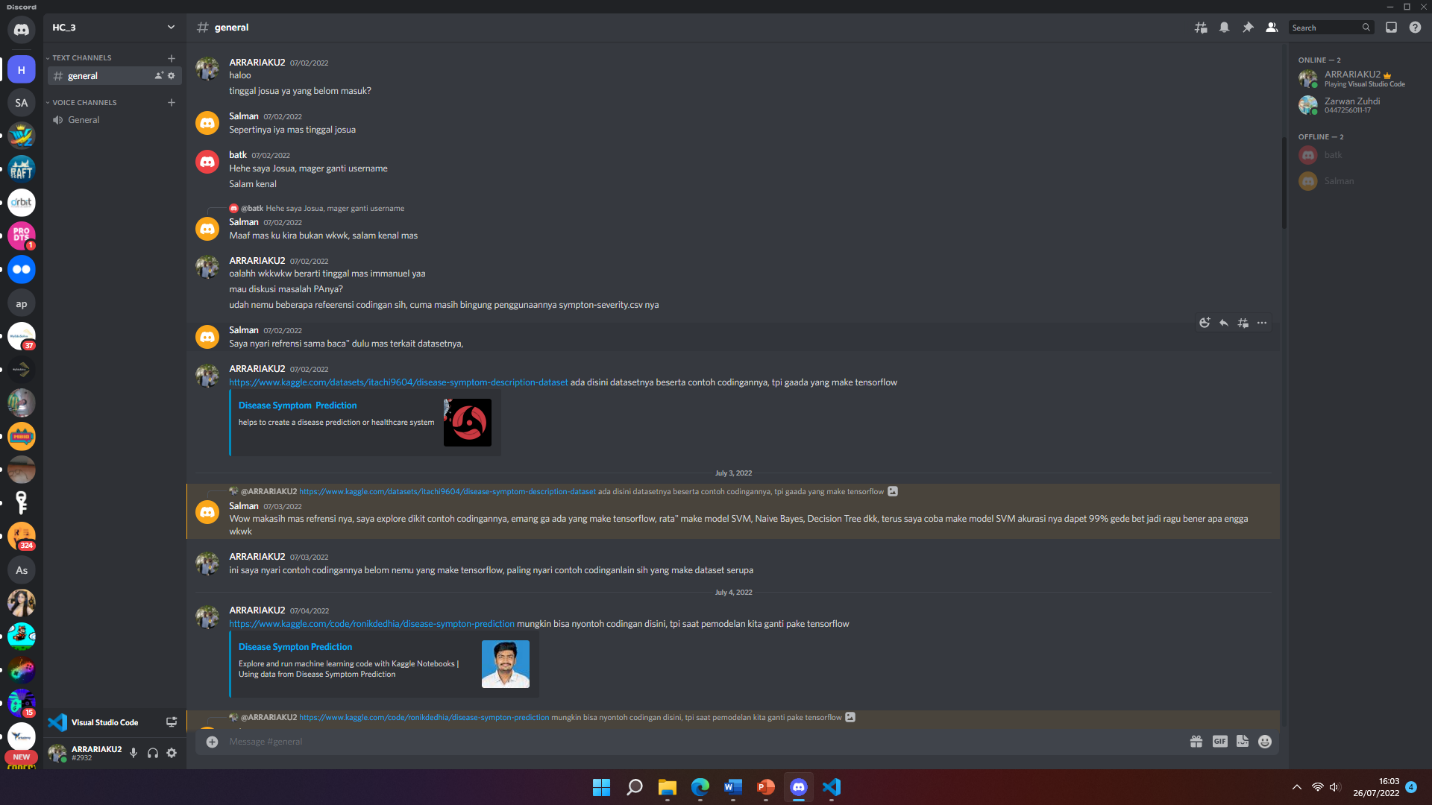
[6] Whidiastuti, Efrika. 2021. “Implementasi Data Mining Untuk Memprediksi Penyakit Hipertensi Dalam Kehamilan Menggunakan Algoritma C4.5”. Diakses online pada 18 Juli 2022. [Available]

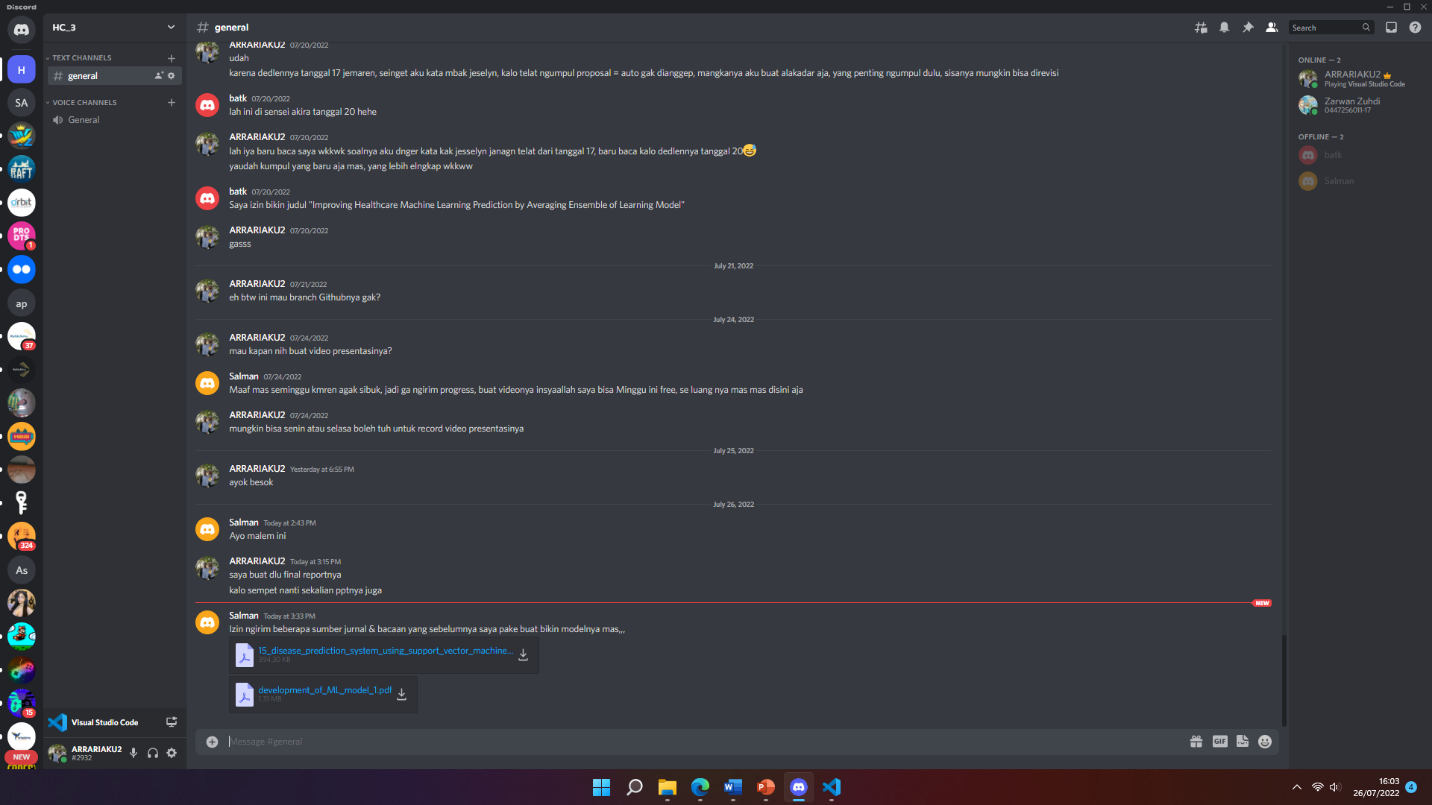
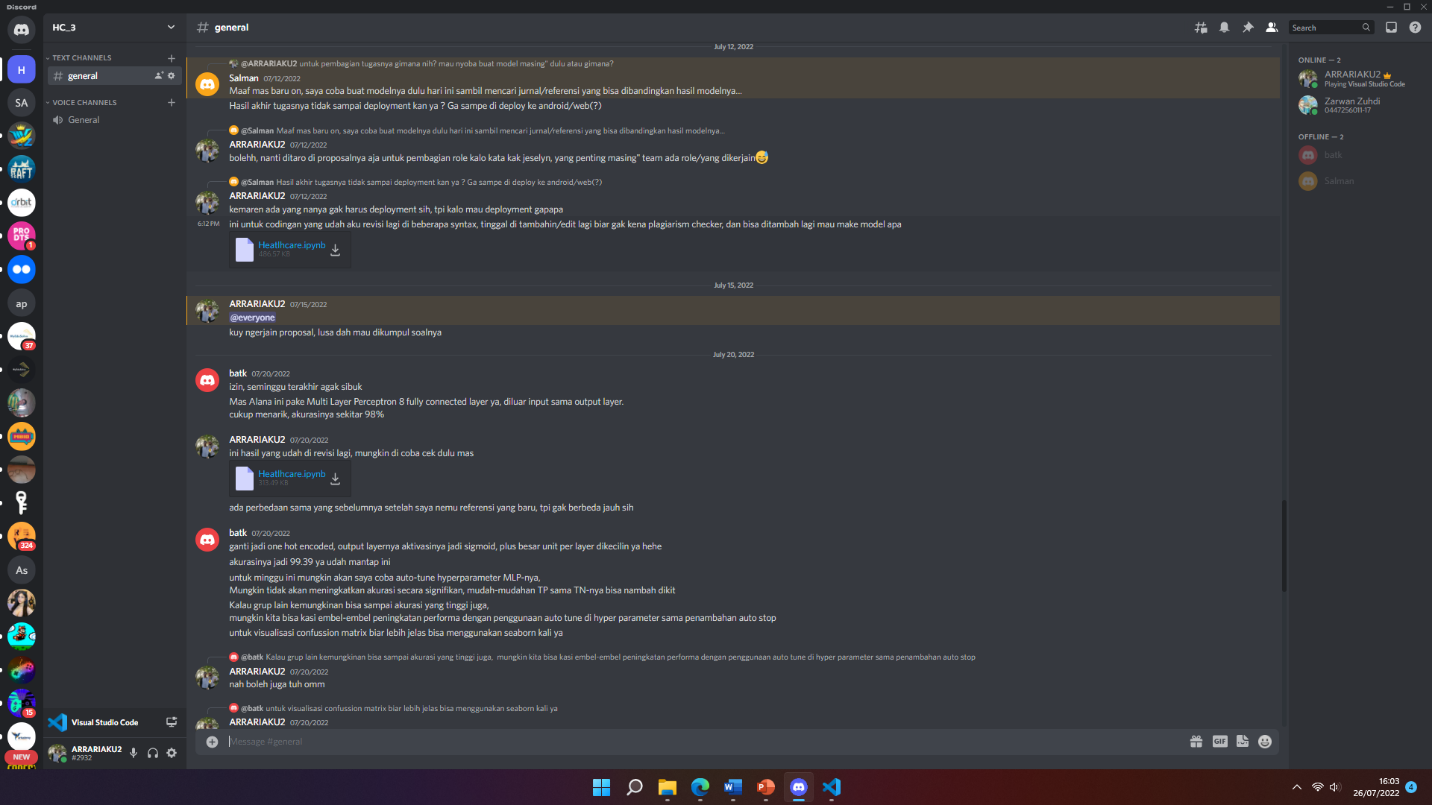
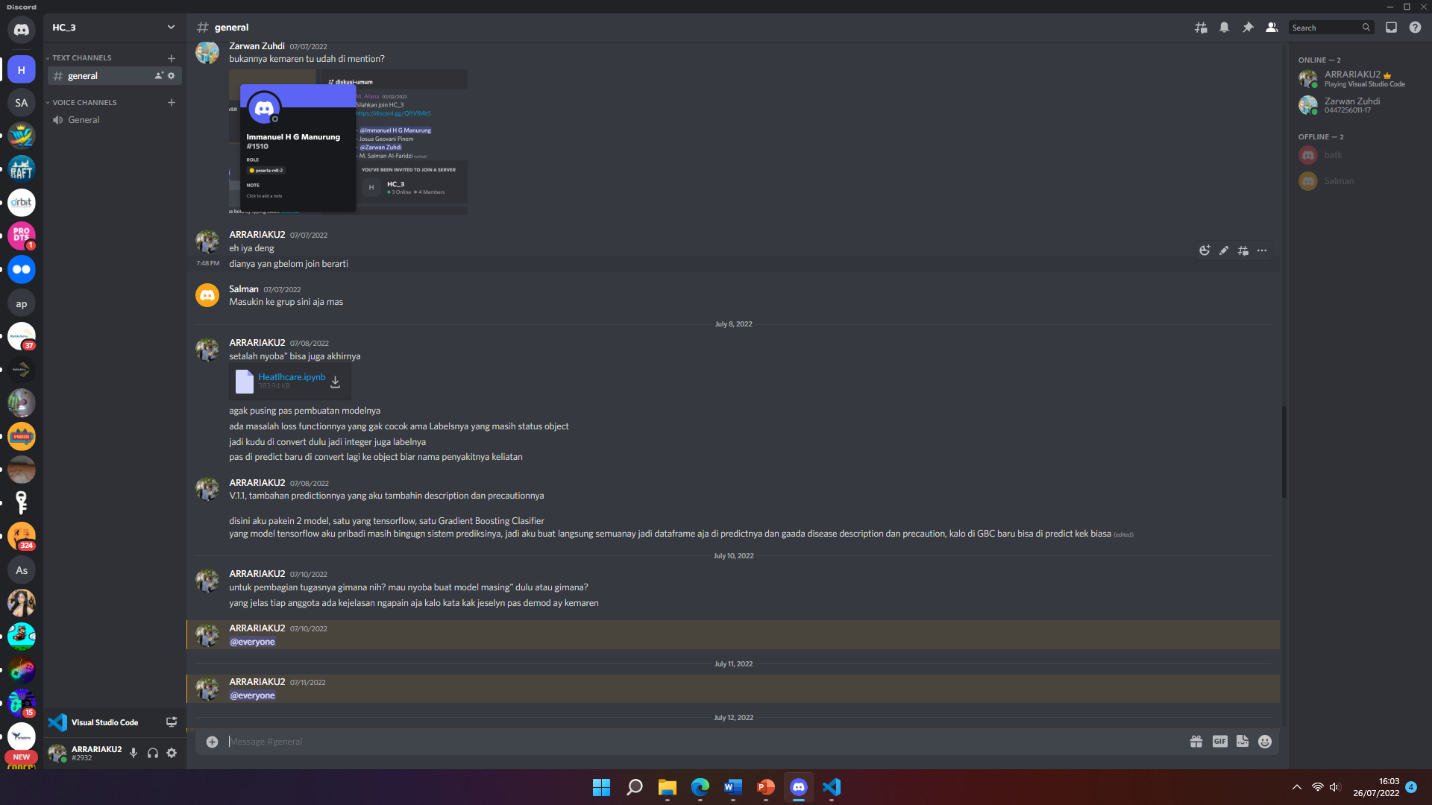
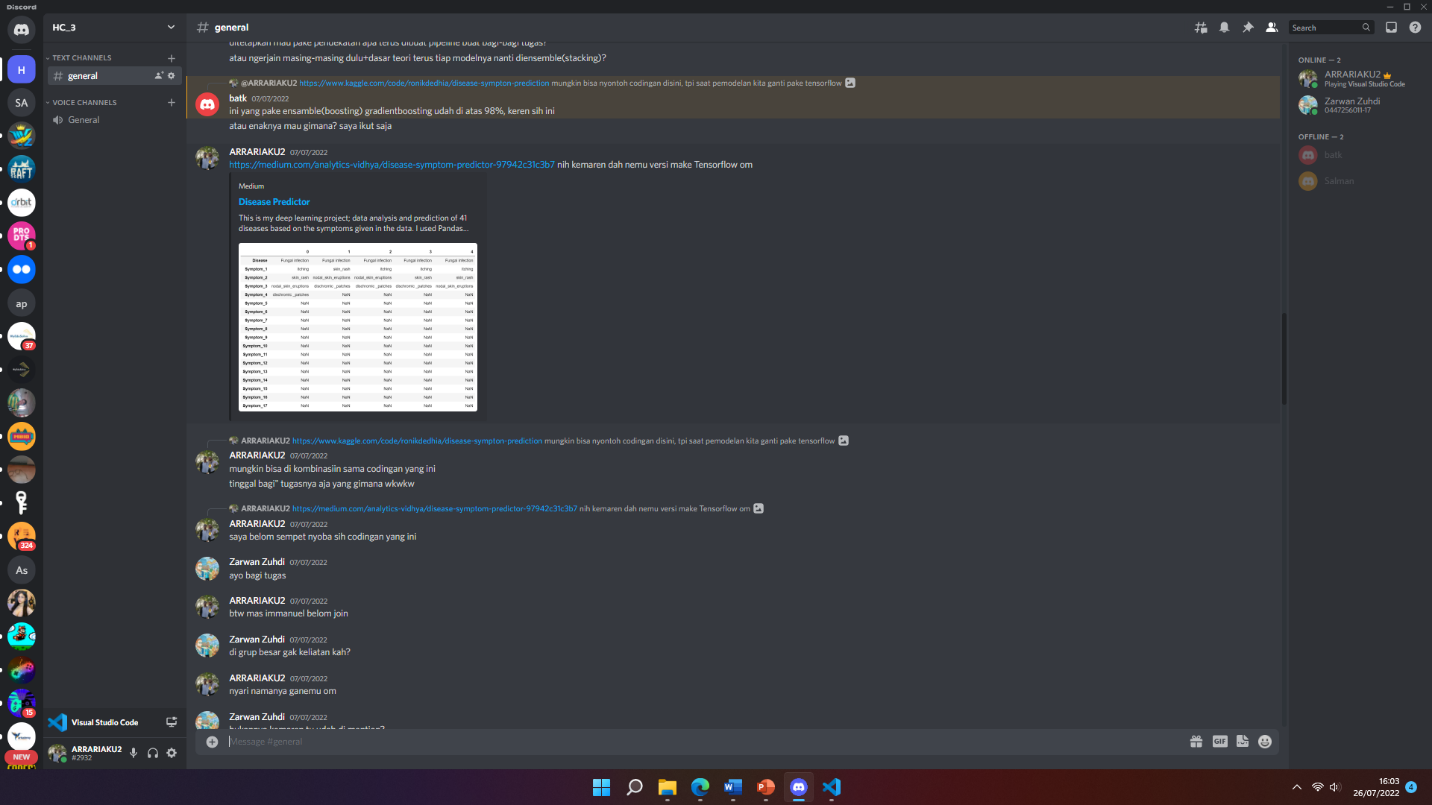
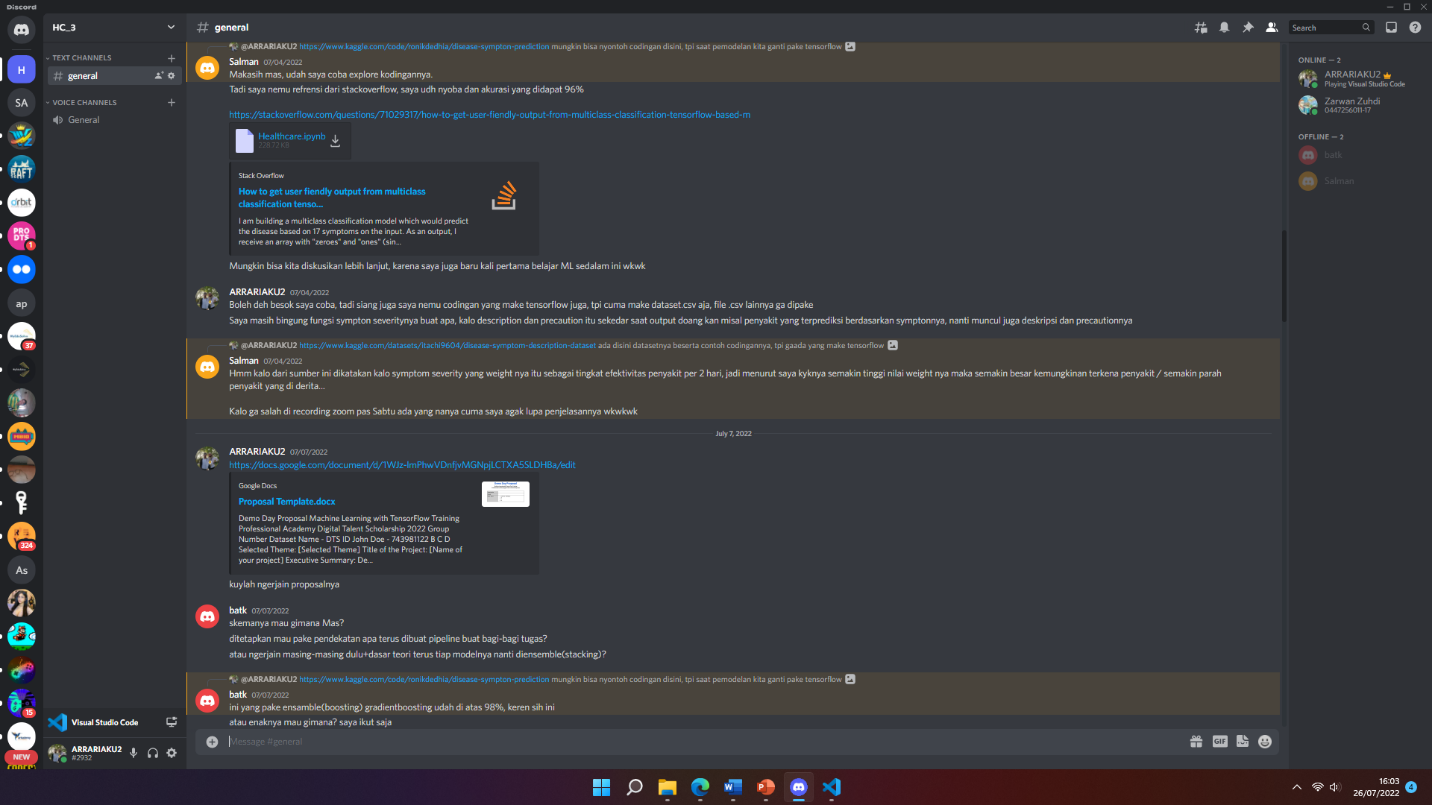
[7] Noor Affan Anshori, Moch. 2018. “Prediksi Gagal Ginjal Kronik Menggunakan Improved C4.5”. Universitas Islam Negeri Sunan Ampel Surabaya. Diakses online pada 18 Juli 2022. [Available]

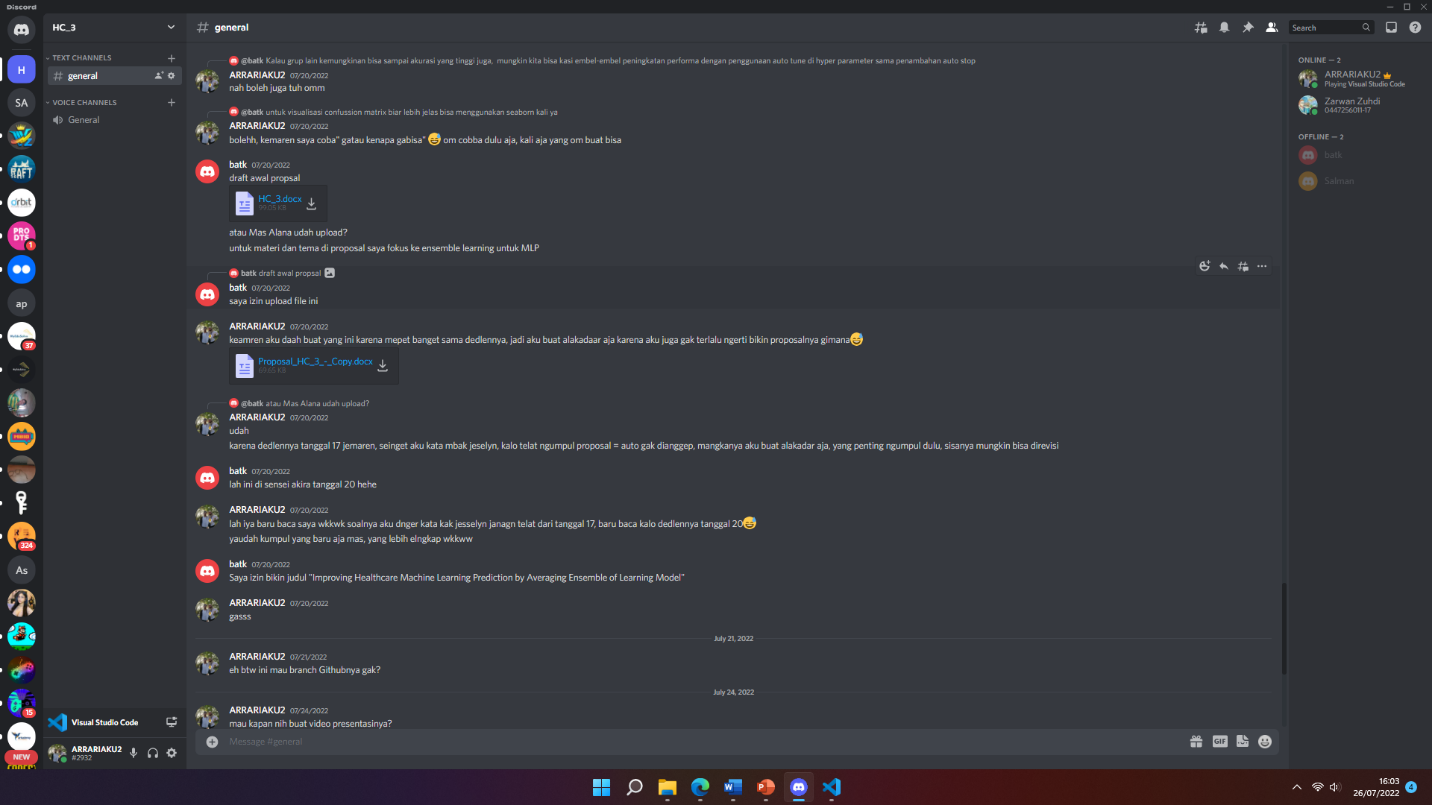
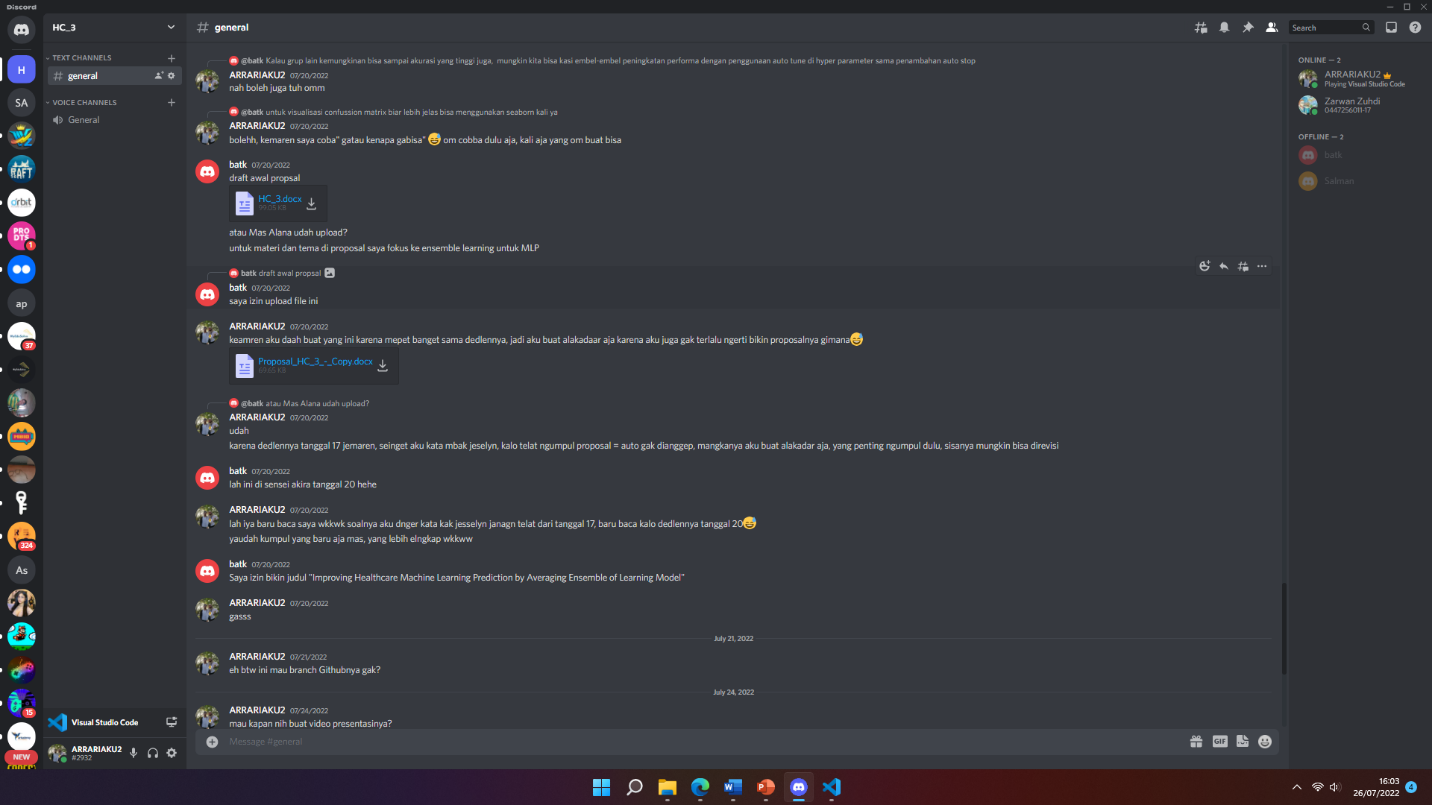
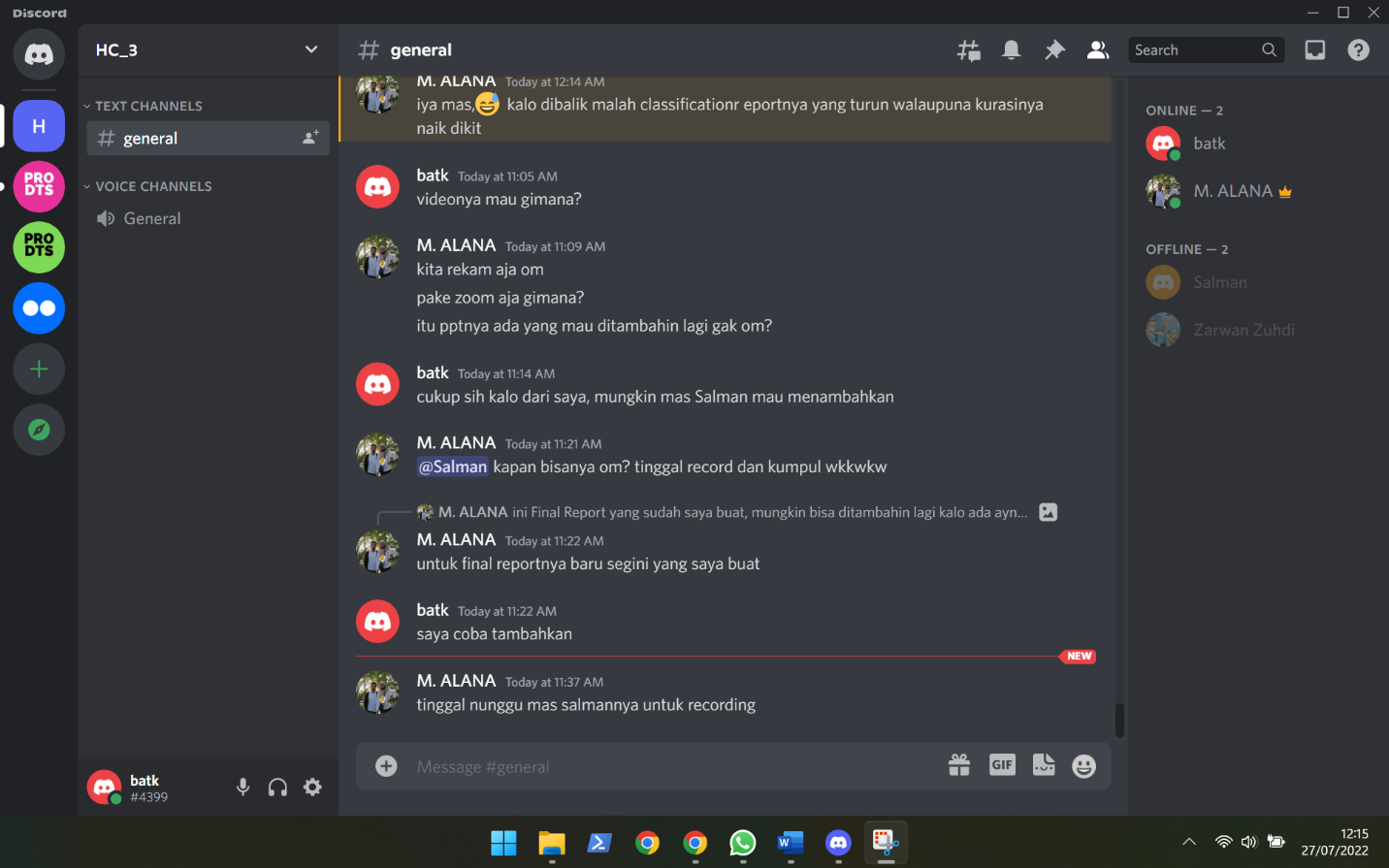
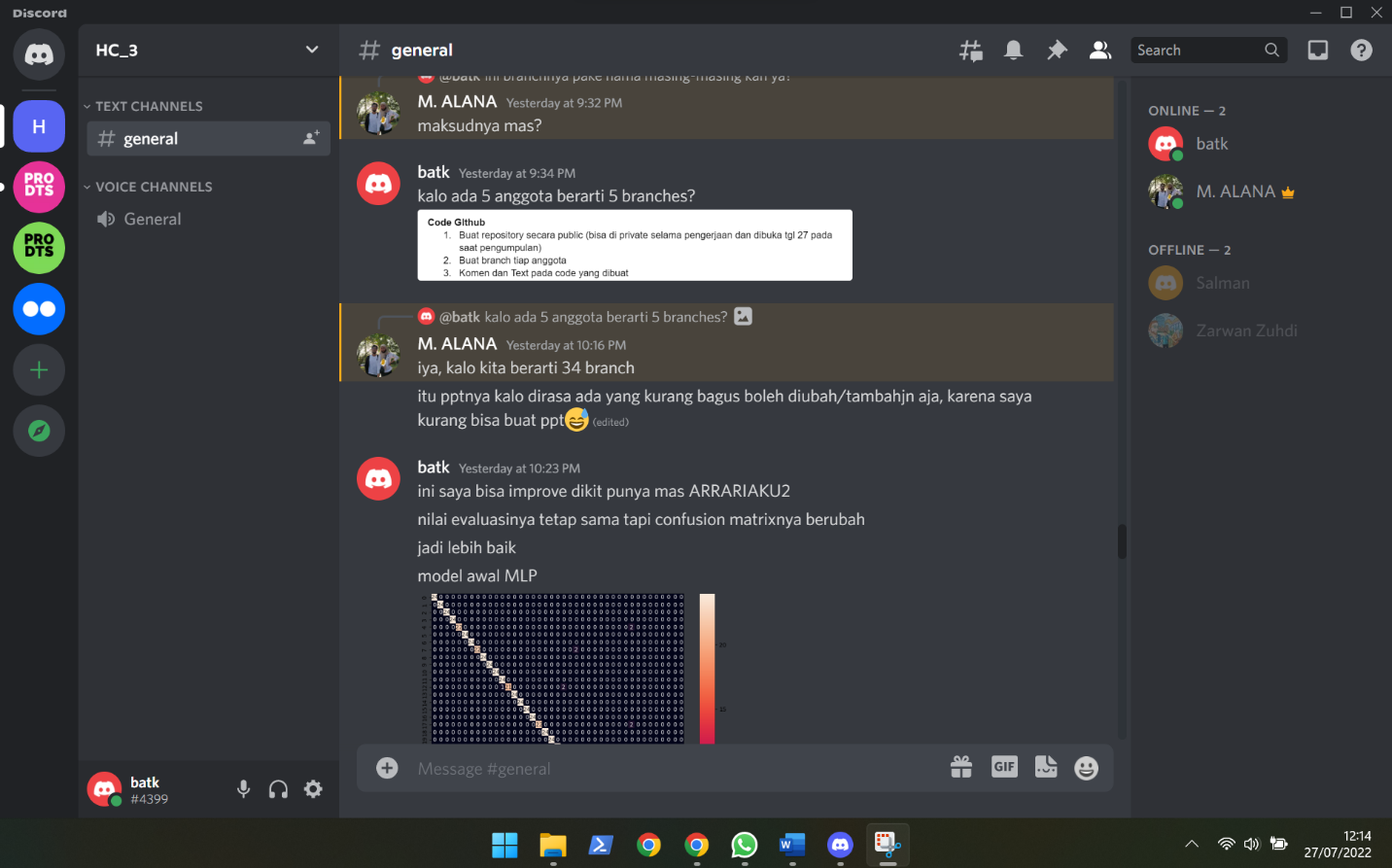
[8] Dong Jin Park, Min Woo Park, Homin Lee, Young‑Jin Kim, Yeongsic Kim & Young Hoon Park. 2021. Development of machine learning model for diagnostic disease prediction based on laboratory tests. University of Korea. Diakses Online pada 26 Juli 2022 [Available]

[9] Md. Ehtisham Farooqui, Dr. Jameel Ahmad. 2020. Disease Prediction System using Support Vector Machine and Multilinear Regression. International Journal of Innovative Research in Computer Science & Technology (IJIRCST). Diakses Online pada 26 Juli 2022.

**Documentation:**

****

**

******