Demo Day Proposal

**Machine Learning with TensorFlow Training**

**Professional Academy Digital Talent Scholarship 2022**

|  |  |
| --- | --- |
| **Group Number** | **HC\_3** |
| **Dataset** | **Healthcare** |
| **Name - DTS ID** | 1. **Immanuel H G Manurung - 152236035100-1079** 2. **Josua Geovani Pinem - 152236035100-236** 3. **Zarwan Zuhdi - 152236035101-654** 4. **M Alana- 152236035100-700** 5. **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

**Executive 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].

**Project Scope & Deliverables:**

Pengembangan beberapa model dengan pendekatan weak learner apabila digabungkan dapat membentuk model baru yang lebih kuat atau strong learner. Pendekatan ini lebih dikenal dengan ensemble learning. Model-model machine learning dapat dibuat dengan karakteristik berbeda-beda dengan tujuan yang sama. Pengembangan projek dengan pembuatan model independent menjadi mungin untuk dilakukan.

Masing-masing anggota di dalam kelompok memiliki kesebukan dan aktivitas beragam. Kesulitan untuk mencari waktu yang tepat untuk melakukan diskusi dan sharing agak sulit untuk dilakukan. Pendekatan awal yang dilakukan dengan menggunakan ensemble learning akan lebih tepat diterapkan. Adapun ruang lingkup dalam pengembangan project ialah:

* + - 1. Model dikembangkan secara independen.
      2. Tidak ada batasan pada hyperparameter yang digunakan.
      3. Tiap model diharapkan memberikan nilai akurasi terbaik.
      4. Tiap model yang ada kemudian disubmit pada branch github masing-masing anggota kelompok. Update sesuai waktu senggang masing-masing anggota kelompok.
      5. Tiap model independen kemudian digabungkan pada tahap akhir untuk memberikan hasil terbaik.

Adapun tantangan dalam pengerjaan projek ialah:

Pencarian hyperparameter terbaik

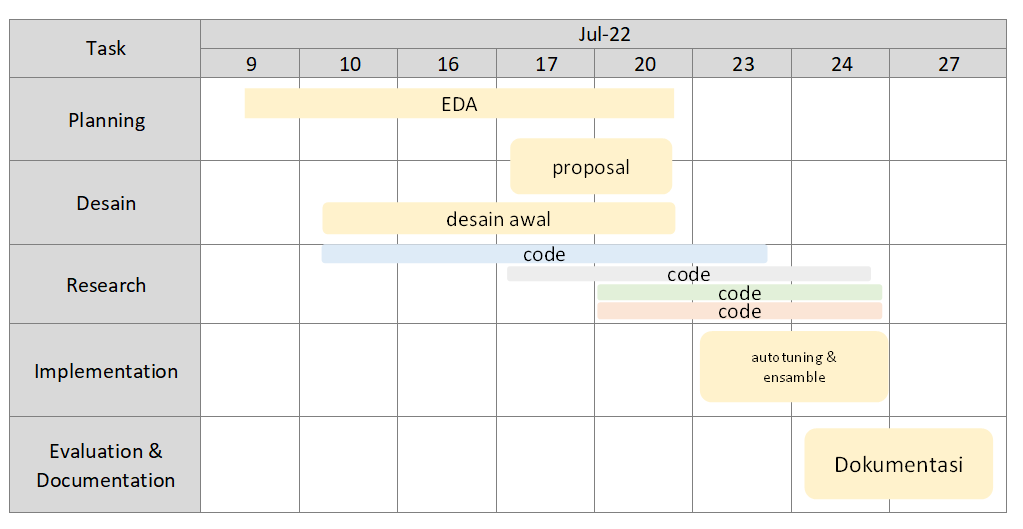
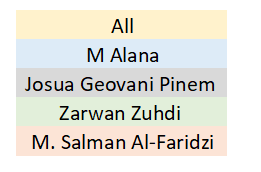
Waktu yang dibutuhkan untuk Exploratory Data Analysis memberikan tantangan tersendiri serta.

Penyesuaian waktu antar anggota kelompok.

Diharapkan dengan pengembangan dengan konsep ensemble machine learning pada project dimana tiap anggota kelompok dapat membuat model terbaik dapat menghadapi permasalahan dan tantangan yang ada. Kemudian untuk pencarian hyper parameter terbaik serta EDA diharapkan dapat diselesaikan sering dengan pengerjaan model.

**Project Schedule:**

Untuk dapat menyelesaikan proyek maka dilakukan definisi pengerjaan proyek dengan time line sebagai berikut:

**Gantt Chart**

**Risk and Issue Management Plan:**

Daftar faktor yang dapat menggagalkan proyek dan rencana bagaimana masalah akan diidentifikasi, ditangani, dan dikendalikan.

**Matrix Risk Management**

|  |  |  |  |
| --- | --- | --- | --- |
| Sangat Mungkin | Bentrok jadwal | Tidak ada progress pengerjaan | Malas |
| Mungkin | Bentrok jadwal | Kekurangan waktu | Bentrok Jadwal |
| Jarang | Network | Kurang referensi dan pengetahuan | Lupa mengumpulkan |
| Kejadian/Dampak | Rendah | Sedang | Tinggi |

**Penjelasan Matrix dan Resolusi**

* **Matrix Value Rendah**

**Network**, permasalahan network kadang terjadi dipengaruhi banyak faktor. Dapat diatasi dengan pengerjaan proyek dengan lebih cepat.

**Bentrok Jadwal,** permasalahan umum yang sering terjadi, solusi tersedia dengan update pengerjaan secara langsung ke github

**Bentrok Jadwal2,** permasalahan umum yang sering terjadi apabila beban individu lebih banyak, solusi tersedia dengan update pengerjaan secara langsung ke github

* **Matrix Value Sedang**

**Kekurangan waktu,** dikarenakan kesibukan kadang tidak dapat menyelesaikan.

**Tidak ada progress pengerjaan,** terhambat pengetahuan dan masalah lainnya, dapat diselesaikan dengan sharing dan diskusi dengan rekan lain apabila ada waktu luang

**Kurang referensi dan pengetahuan** sharing dengan mentor dan sensei

* **Matrix Value Tinggi**

**Malas** bawaan lahir, diperlukan dorongan dari rekan tim

**Bentrok jadwal3,** permasalahan umum yang lebih sering terjadi, solusi tersedia dengan update pengerjaan secara langsung ke github

**Lupa mengumpulkan,** kurang koordinasi sehingga tidak ada yang mengumpulkan. Diperlukan koordinasi antar anggota team lebih sering terjadi.

**Project Reference:**

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

**Referensi:**

[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]