TASK1:

Github

TASK2:

Background:
Mainly I consider any medical personnel, as potential users of these model and also, depending on the outcome of my work, possibly it could also help non-medic related people detect the chance of them potentially having a glaucoma. But still, the most precise results will be available only with results of medical analysis, tools for which are usually not available for the majority of people.

Business goals:
Mainly it could help to achieve the amount of cases, when glaucoma is detected before later stages of this disease, thus potentially helping to escape any irreparable eye damage and obstructed vision, which happens at later stages of glaucoma.

Business success criteria:
Main goal is to gain a model, which with all the available dataset features can predict all cross validation(k=5) cases glaucoma diagnosis and stage with recall above or equal 0.95, while keeping accuracy and precision above 0.8.

Inventory of resources:
1000 entries available on kaggle with parameters:
Patient ID, Age, Gender, Visual Acuity Measurements, Intraocular Pressure (IOP), Cup-to-Disc Ratio (CDR), Family History, Medical History, Medication UsageVisual Field Test Results, Optical Coherence Tomography (OCT) Results, Pachymetry, Cataract Status, Angle Closure Status, Visual Symptoms, Diagnosis, Glaucoma Type

Requirements, assumptions, and constraints:
It is required to provide the model which will be able to achieve results, listed in business success criteria before the 11th of december 2023.

Risks and contingencies:
Inability of the only “team” member to complete the task due to any unpredictable problem, rendering the person incapable of achieving the required results.

Terminology:
Glaucoma: a group of eye diseases that can cause vision loss and blindness by damaging a nerve in the back of an eye called the optic nerve.

model: a program that can find patterns or make decisions from a previously unseen dataset.

Costs and benefits:

No benefits
Costs: ~50 euro, for living in campus, including electricity;

Data-mining goals:

Final results should include at least one model, which is able to give the results, as they are stated in the "Business success criteria" paragraph. Also, if possible make this model able to work with incomplete data, or different model(s) to work with incomplete data, this depends on the importance of different features, listed in the initial dataset, and whether it is possible to exclude some of them, while still keeping acceptable quality of the model's output.

Data-mining success criteria:

If the main model is able to give the results required in "Business success criteria". Criterias to discern the success of any additional model will depend on its existence, data provided, and out of accuracy, precision and recall, mainly recall will be a decisive factor while appraising these models.

TASK3:

As of now the only source of data I have is kaggle dataset, and for now I do not plan on seeking any additional data for this project, though if during data processing and evaluation I will find any data resource, which I deem relevant and helpful for my task, I may use it during training, but model testing will always be done with initial dataset, taken from kaggle.
Now, the second part of this task I will describe all the data available from the initial dataset, while trying to evaluate its importance for the task, and preliminary deciding which parts should be kept, and which are not really important for achieving the required result, which later should also show whether it is possible to create additional possibilities for prediction with incomplete data, preferably with missing information, for which specific tools unavailable for majority of people required.

1. **Patient ID:** Outright unnecessary piece of data, which should be excluded from the training dataset.
2. **Age:** glaucoma usually becomes noticeable at later age, so people at their later years definitely have a higher chance of glaucoma chance, but it does not help the idea of predicting it, and will rather obstruct it, so probably should be deleted.
3. **Gender:** From some research I got that there are more cases of women having glaucoma worldwide, so that feature will definitely be used for training.
4. **Visual Acuity Measurements:** For now I am not sure whether it will be important, so it will be kept.
5. **Intraocular Pressure (IOP):** Glaucoma is closely related with pressure inside the eye, so this feature has a high chance of being relevant.
6. **Cup-to-Disc Ratio (CDR):** This parameter is widely used to check if there is a glaucoma, while by itself it cannot point whether person has glaucoma or not, as it rather requires continuous checks and if this ratio is increasing or one eye has it

noticeably bigger than the other one, both of these variants usually rise suspicion for the reason being glaucoma.

7. **Family history:** As was pointed out before women seem to have a larger chance of having glaucoma, and usually it is related to the chromosomes, which are hereditary, it is possible to assume that the chance will be bigger if someone from patient's relatively late ancestors had glaucoma.
8. **Medical History:** For now I am not sure, but maybe only leaving the parameter of any medical history related to the eyes, instead of this initial one will be productive. But maybe there are some related statuses.
9. **Medication Usage:** This highly depends on the previous one, as if there are any medical states sequential for initial stages of glaucoma, this parameter will be useful, otherwise possibly should be changed to whether the patient takes any eye-related medications.
10. **Visual Field Test Results:** As glaucoma's most noticeable consequence is reduction of visual field this parameter will be helpful when deciding on the later stages of glaucoma.
11. **Optical Coherence Tomography results:** This is a test which shows state of the back-part of the eye, and as main reason of glaucoma is high blood pressure in this part, at least in glaucoma's later stages it could be easily seen that this part is clotted with blood and mucus, so while it will definitely help with later stages of glaucoma, it is possible that before it becomes noticeable there are some other signs.
12. **Pachymetry:** It is a quick and simple test to discern thickness of the part of an eye, called cornea, and people with thin corneas have higher risk of having a glaucoma, so this will definitely be an important feature, as it allows to also predict glaucoma in its earliest stages, effectively helping obstruct its progression.
13. **Cataract Status:** For now it is believed that cataract and glaucoma have no direct relation, but both usually become noticeable at a later age, so probably just like age it will only obstruct training, so should also be deleted unless I will find some other info.
14. **Angle Closure Status:** This parameter may seem similar to the vision field, but this one indicates whether eye field deteriorated compared to the previous state, so it may point to the later stages of glaucoma.
15. **Visual Symptoms:** This parameter represents different easily noticeable conditions, and some of them will definitely be helpful with predicting glaucoma

TASK4:

1. Dataset preprocessing. ~5h
2. Feature selection. ~5h
3. Model building. ~10h
4. Evaluation. ~5h
5. Documentation. ~5h

All will be done by Andrei Kuzmin, using tools such as: laptop, kaggle notebook, or jupiter notebook. Also, I maybe should make the additional idea clear, as I am not sure if it is possible or not it is an additional task, which I did not put as the main one, but overall I want to create a model which will be able to show a chance of having glaucoma without any additional tests, only using person's own knowledge and feelings, but whether it is possible

depends highly on both data provided in dataset, and also possibility and viability of it, as it may be not possible to do at all with just this data, or prediction will not be high enough to be relevant.