# **Техническое задание к проекту М1**

## **1. Общая информация и цель проекта**

В этом проекте предполагается работа с датасетом из открытого источника. Вам потребуется разработать модель машинного обучения, а также подготовить библиотеку и интерфейс к ней для предсказания на тестовой выборке. В начале проекта вам будет доступен обучающий датасет и тестовый датасет, не содержащий правильных ответов. Метрика на тестовом датасете будет измерена в ходе проверки работы. Также вам будет доступен скрипт, с помощью которого будет измеряться финальная метрика вашей модели.

## **2. Задачи проекта**

**Предполагается, что в ходе работы над проектом будут решены следующие задачи:**

* Исследование датасета (предполагается поиск дополнительной информации для лучшего понимания природы данных)
* Предобработка данных. Будьте внимательны: в данных могут быть бесполезные признаки, сильно скорелированные признаки а также косвенные утечки целевого признака. Некоторые колонки потребуется преобразовать к нужному типу данных.
* Обучение модели
* Подготовка предсказания на тестовой выборке. Обратите внимание. Файл предсказаний должен быть в csv формате и содержать 2 столбца: “id” и “prediction”.
* Подготовка скриптов и библиотеки для обработки данных и предсказания на тестовой выборке
* Написание инструмента для тестирования
* Оформление документации

## **3. Описание исходных данных**

**Данные пациентов для предсказания риска сердечных приступов**

* **id - id**
* Антропометрические параметры (вес, возраст, рост)
* Привычки (курение, качество сна и т.д)
* Давление
* Наличие хронических заболеваний
* Биохимия крови
* **Таргет - высокий или низкий риск поражения сердца**

[**Train**](https://drive.google.com/file/d/1QYkOGKlHtugjJcccN7UhwPAvfsfyck7C/view?usp=sharing)

[**Test**](https://drive.google.com/file/d/1qJduj_-_nWCNhfTdICReepAbUs8OHC_Y/view?usp=sharing)

[**Скрипт, которым будет тестироваться качество вашей модели**](https://drive.google.com/file/d/12sMCamQieVwGXa-VIKjbjXdJqeUYSd-D/view?usp=share_link)

## **4. Используемая метрика оценки качества и сравнения моделей**

В ходе анализа датасетов вам предстоит самостоятельно определить наилучшую метрику оценки качества модели и обосновать ваш выбор.

## **5. Стек**

* Python ([ООП](https://www.dmitrymakarov.ru/python/oop/))
* Pandas
* numpy
* Scikitlearn
* CatBoost
* FastAPI
* HTML

## **6. Таймлайн проекта**

Проект длится 2 недели и включает в себя:

* Установочную встречу
* Семинар по FastAPI + QA-сессия
* Финальная встреча

## **7. Результат**

В результате у вас должно получиться приложение на FastAPI (или аналогичном фреймворке) которое принимает на вход путь к csv файлу тестовой выборки, выполняет предсказание (например, путем POST запроса к сервису) и возвращает ответ в формате JSON. Допускается для тестирования написать скрипт, который посылает запрос к запущенному ранее приложению, либо реализовать веб-интерфейс.

Обратите внимание, что оцениваться будет помимо качества модели структурированность, [понятность](https://peps.python.org/pep-0257/#what-is-a-docstring) и [чистота кода,](https://peps.python.org/pep-0008/) архитектура приложения. Обязательным является использование ООП, то есть составные части приложения должны быть реализованы в виде классов. Проект должен быть оформлен в виде [репозитория](https://docs.github.com/en/repositories/creating-and-managing-repositories/best-practices-for-repositories) на github (gitlab), должна быть написана документация к сервису с описанием для пользователя и описанием классов и методов в приложении, как если бы далее ваш проект вы передавали пользователю. В jupiter ноутбуках могут быть только исследования и эксперименты, если демонстрацию результатов работы приложения вы также выносите в ноутбук - то кода приложения в ноутбуке быть не должно: он должен импортироваться в качестве внешних модулей.

**Формат сдачи работы**

В репозитории должны быть:

* Jupiter Notebook с исследованием, обучением моделей и выводами
* Код вашего приложения
* Предсказание на тестовой выборке в требуемом формате
* Инструкция по запуску вашего приложения ИЛИ Jupiter Notebook с демонстрацией работы приложения ИЛИ выступление на финальной встрече где вы продемонстрируете работу приложения вживую

Критерии оценки

* Исследование и результаты обучения моделей
* Метрика на тестовой выборке (измеряется при проверке)
* Чистота кода и структурированность кода
* Оформление репозитория
* Воспроизводимость результатов