Тестовое задание для стажера на позицию «Разработчик Python (проект Machine Learning)»

**Задание:**

Разработать систему машинного обучения, которая по списку статически импортируемых библиотек exe файла предсказывает, является ли этот файл зловредным.

Для выполнения задания предоставляются три выборки: обучающая, валидационная и проверочная. Выборки представлены в виде tsv файлов с тремя колонками – is\_virus – является ли файл зловредным: 1=да, 0=нет; filename – имя файла для ознакомления; libs – через запятую перечисление библиотек, статически импортируемых этим файлом (мы использовали библиотеку LIEF для получения списка).

На обучающей выборке – train.tsv – следует тренировать модель машинного обучения.

На валидационной выборке – val.tsv – требуется подсчитать, насколько хорошо модель справляется с файлами, которые она не видела при тренировке. Характеристики требуется записать в текстовый файл validation.txt со следующим содержанием (изменив значения на ваши):

True positive: 2

False positive: 20

False negative: 18

True negative: 60

Accuracy: 0.6200

Precision: 0.0909

Recall: 0.1000

F1: 0.0952

Проверочная выборка – test.tsv – содержит только колонку libs. Для проверочной выборки требуется создать файл prediction.txt, в котором для каждой строки файла проверочной выборки будет содержаться один символ: либо 1 если модель предсказывает этот файл как зловредный, либо 0 иначе. Первая строка файла, соответствующая заголовку проверочной выборки должна быть “prediction”:

prediction

0

0

1

( … много строк пропущено … )

0

1

0

**Дополнительное задание:**

Для проверочной выборки создайте ещё один файл, explain.txt, где для каждой строки файла проверочной выборки будет содержаться причина (в свободном человеко-читаемом формате) по которой модель посчитала этот файл зловредным. Строки для не зловредных файлов должны быть пустыми.