##### Лабораторна робота № 1

**Байєсівський класифікатор, машина опорних векторів(SVM) та градієнтний бустінг**

**Варіант 3**

**Тема:** дослідження різних методів для задач класифікації.

**Мета:** провести класифікацію за допомогою баєсівського классифікатора, SVM, градієнтного бустінга та провести аналіз якості їх роботи у відповідності до даних варіанта.

**Виконали:**

**студенти групи КА-03мп**

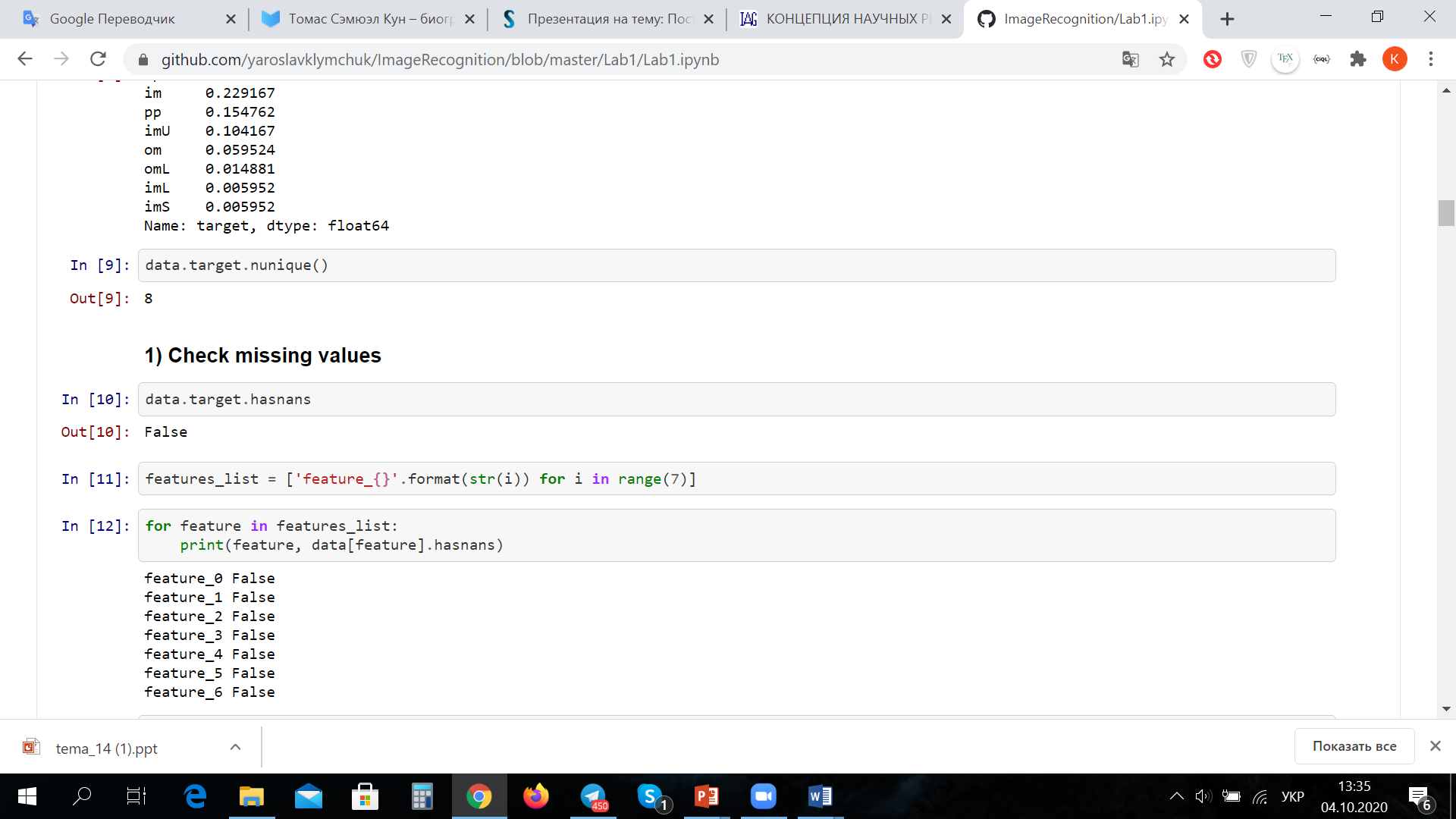
**Климчук Я.І.**

**Кочмар К.С.**

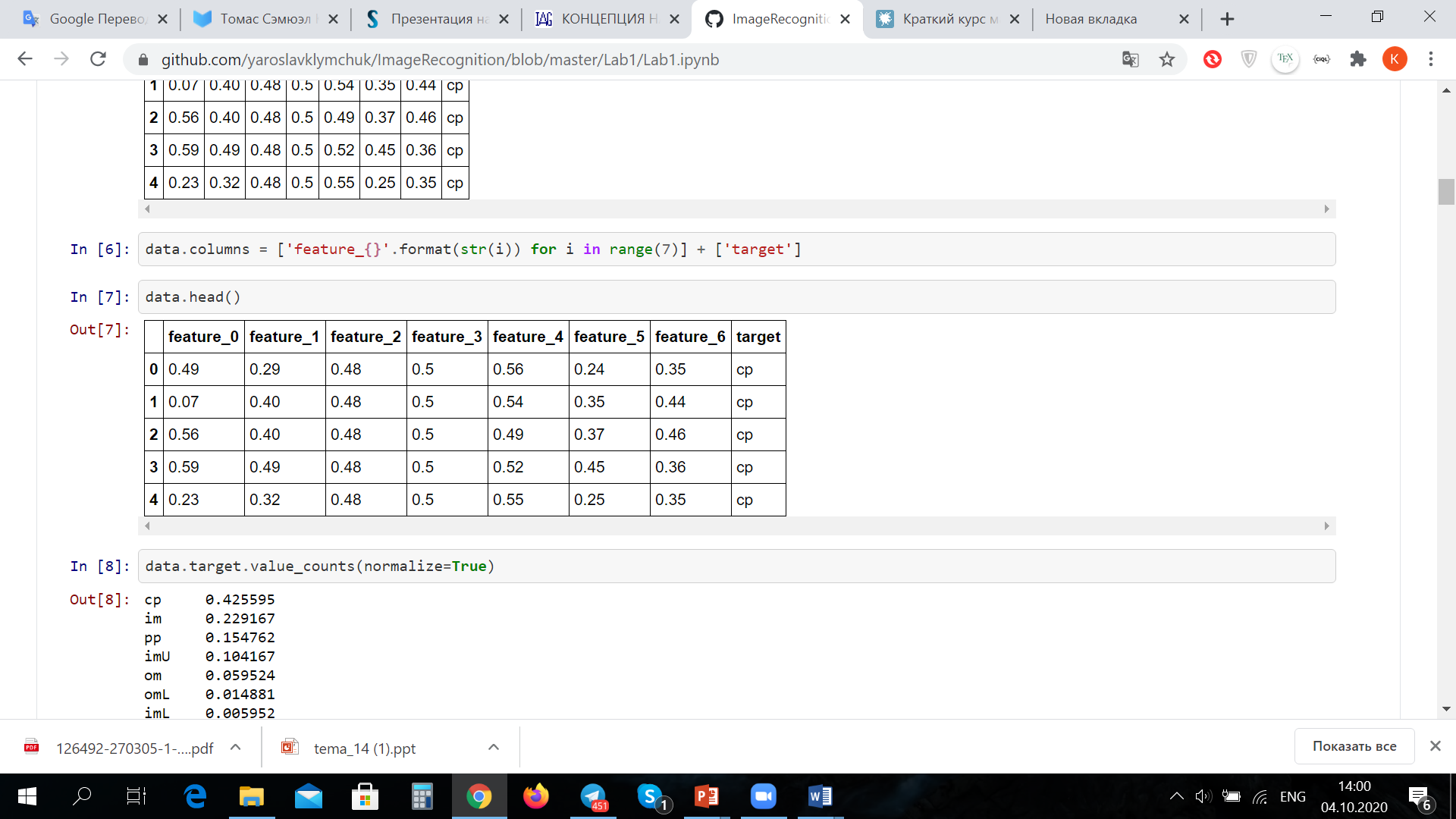
**Жуковський А.В.**

**Хід роботи**

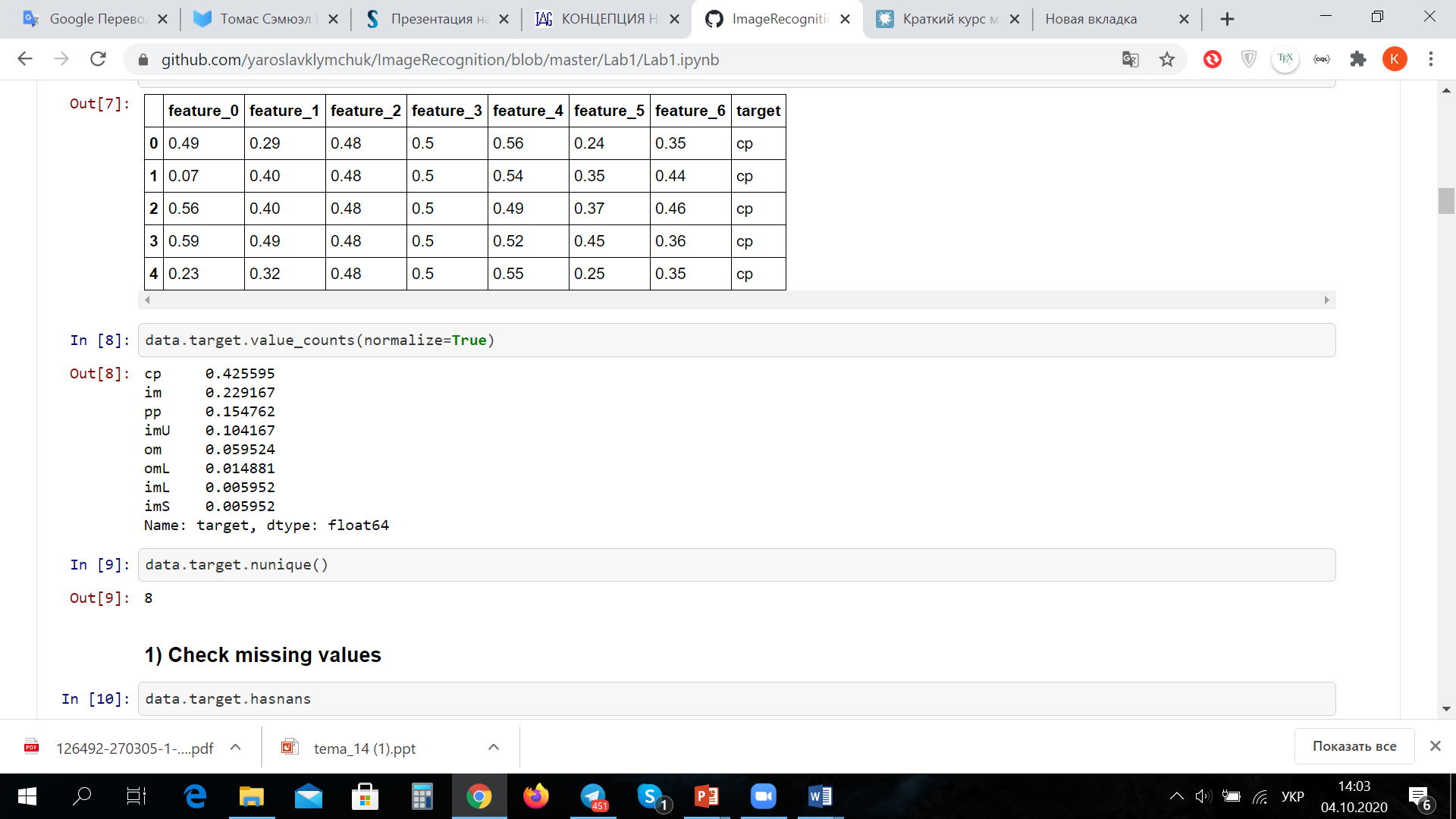
1. Результати препроцесингу:
2. Перевірка на пропущені дані виявила, що вибірка не містить пропусків



1. Вибірка містить 336 екземплярів, які характеризуються 7 ознаками (feature\_0, feature\_1, …, feature\_6), і 1 вихідну змінну (target). Для подальшої зручності стовпці датасету було названо у відповідності до їх змісту:

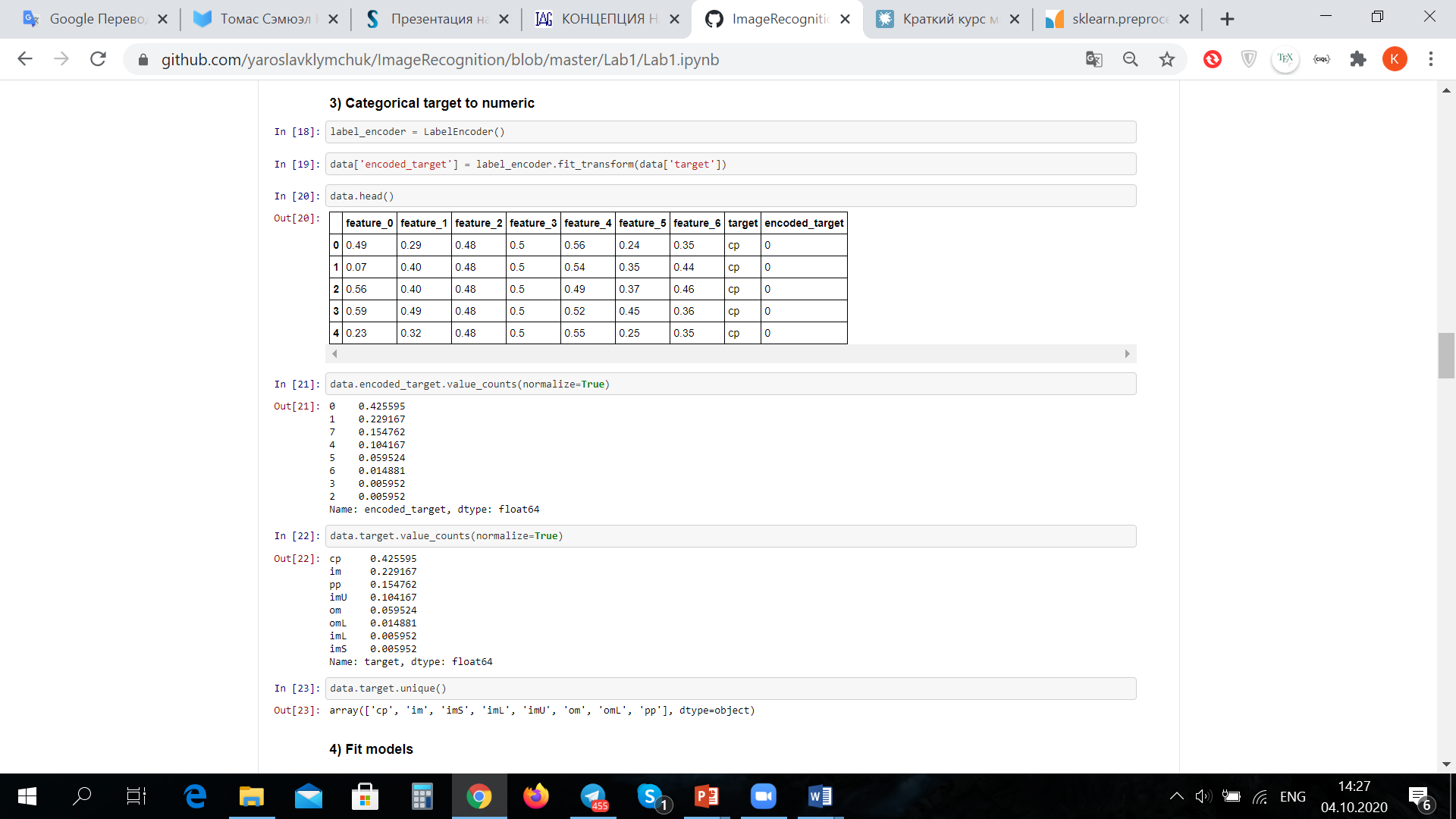


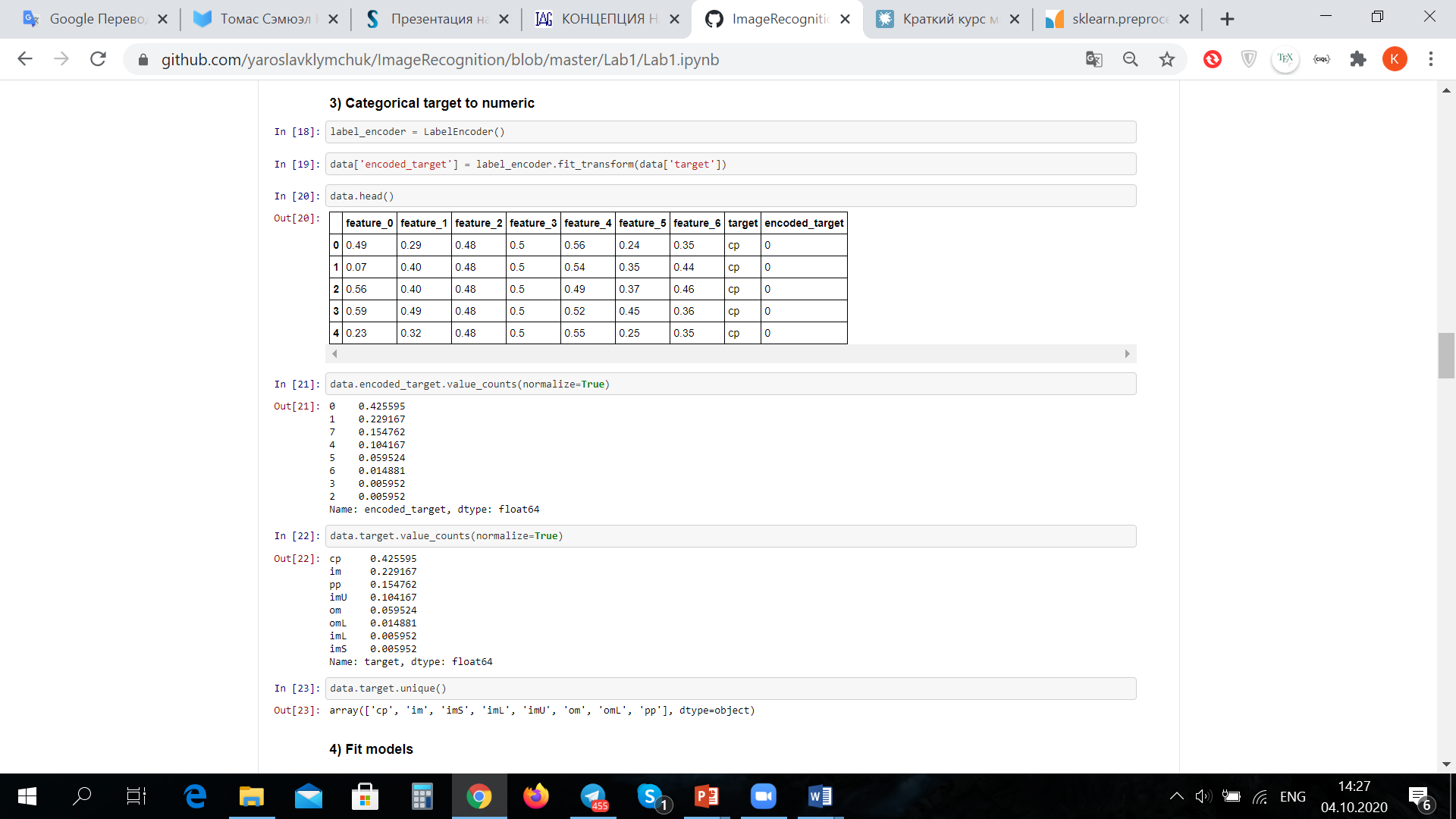
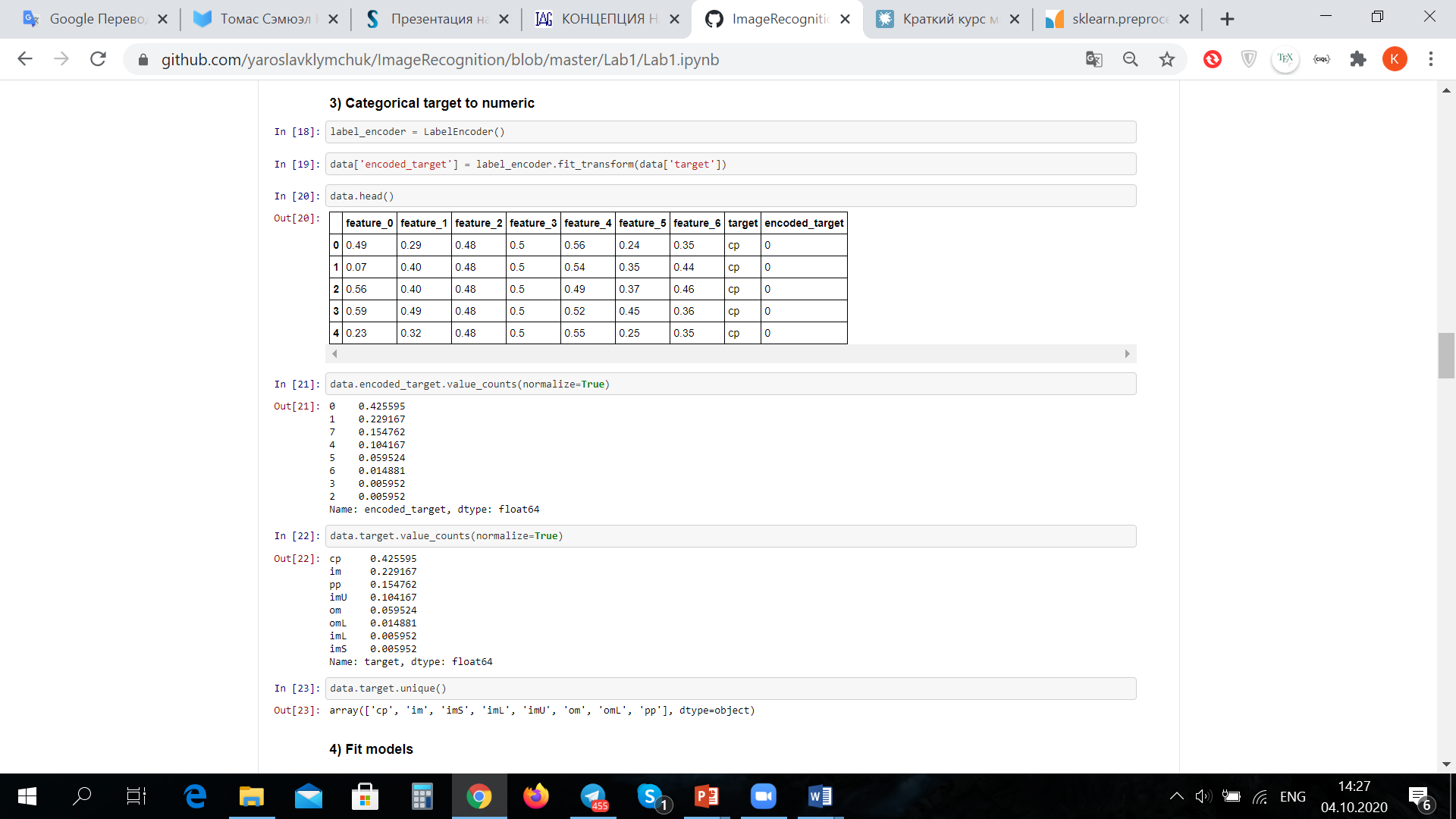
1. Вихідна змінна поділяється на 8 класів, частка яких у загальному наборі даних розподілилась наступним чином:



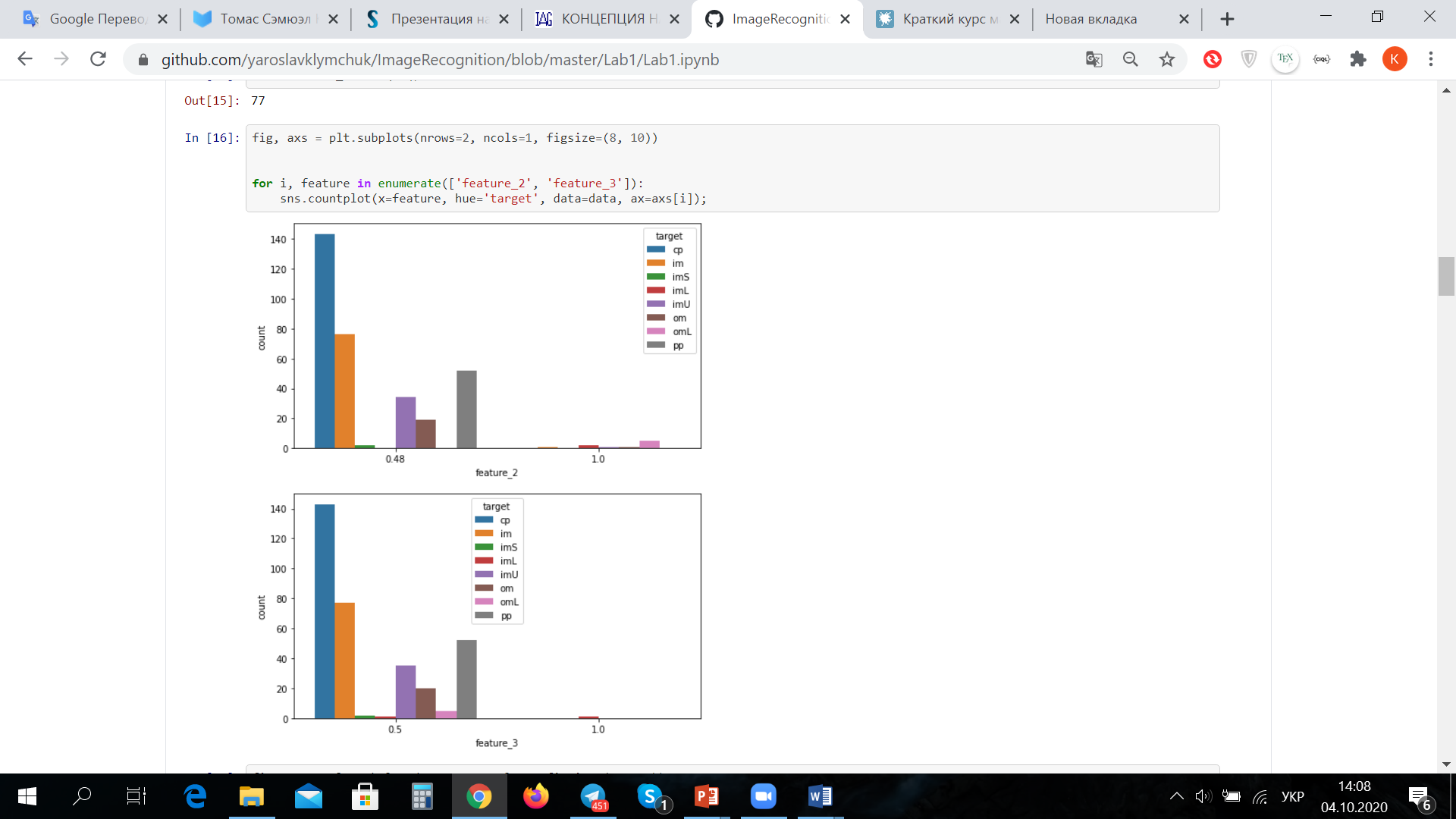
Як бачимо, початкова вибірка у переважній більшості складається з екземплярів класу cp, а представники класів imL та imS становлять меншість.

1. Оскільки вихідна змінна подана у якісному записі, то переводимо її у кількісний вид, застосовуючи функцію fit\_transform класу LabelEncoder із бібліотеки sklearn. Тобто відбувається кодування цільової змінної значенням від 0 до n\_classes-1:



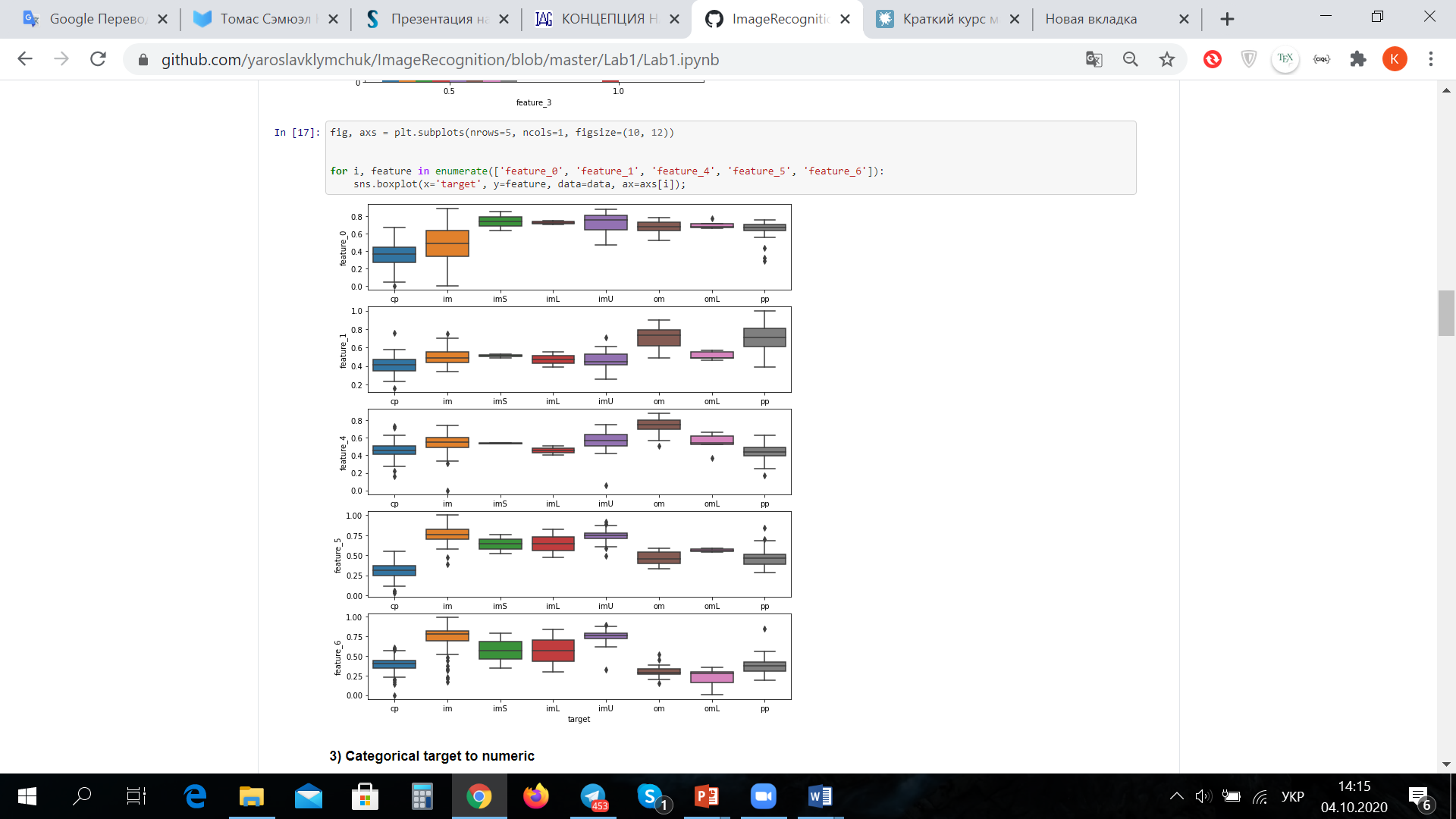
 

1. Результати аналізу даних
2. З’ясували, що ознаки feature\_2 та feature\_3 для всіх екземплярів всіх класів можуть приймати по два значення відповідно.

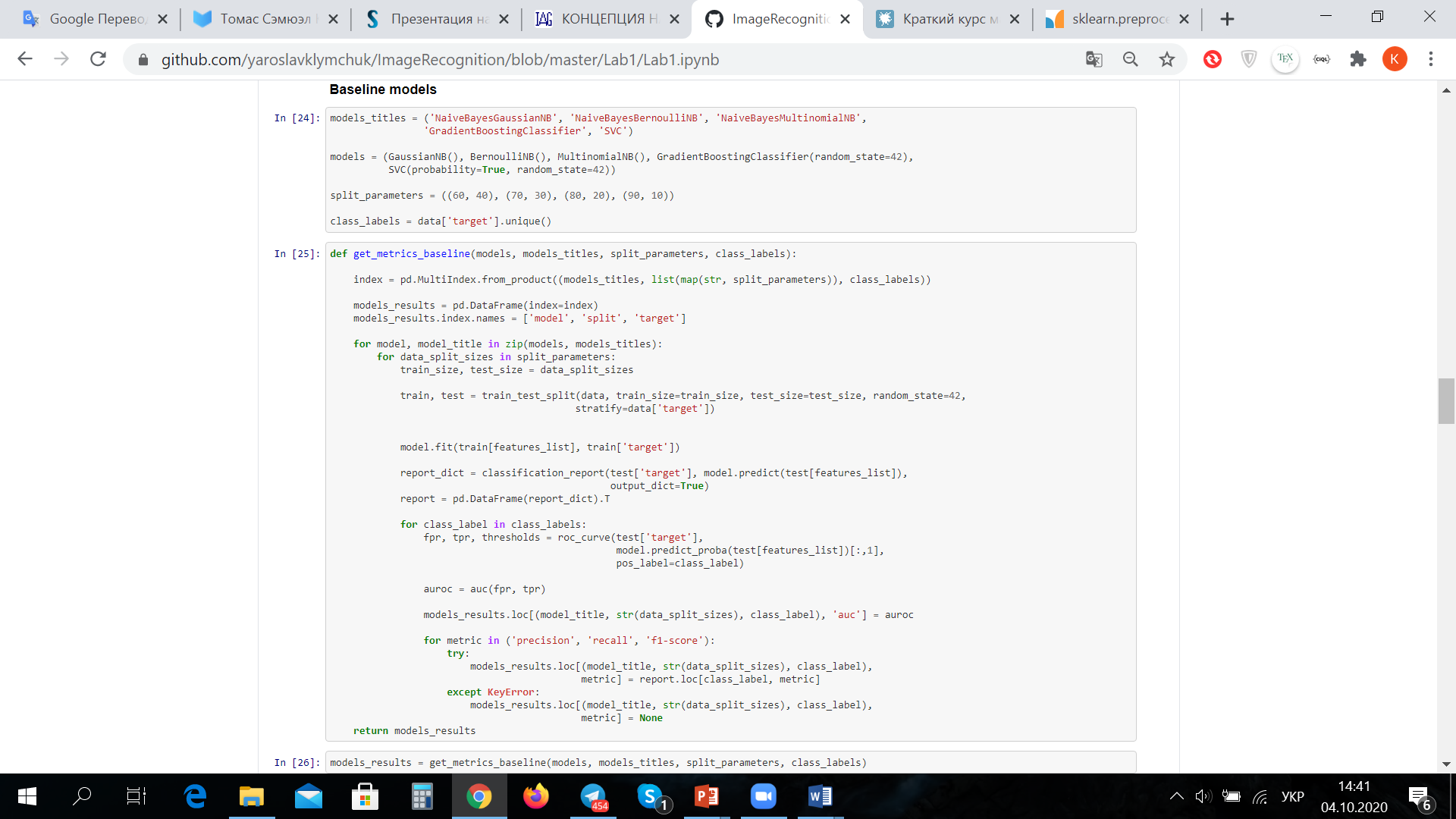


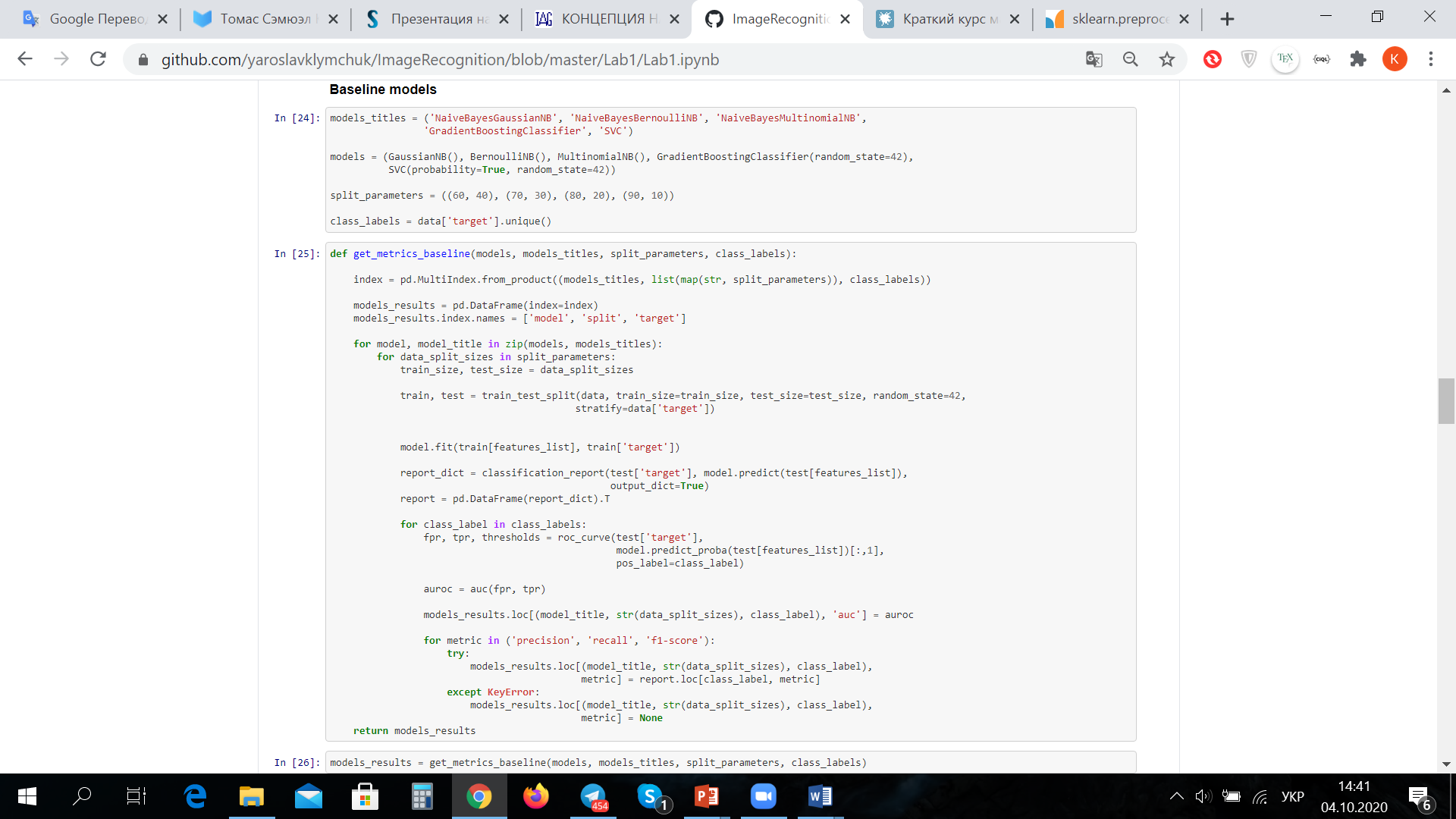
Із результатів виведеної візуалізації можемо стверджувати, що більшість класів за ознакою feature\_2 набувають значення 0,48, а для ознаки feature\_3 переважає значення 0,5.

1. Що ж до інших ознак, то вони є більш варіативними: кожному класу притаманний власний набір значень за відповідною ознакою:



1. Результати побудови baseline моделей
2. Проводимо класифікацію за 5 моделями – GaussianNB, [BernoulliNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#_blank), [MultinomialNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#_blank), GradientBoostingClassifier та SVC. Для кожної із зазначених моделей проводимо розбиття вибірки на навчальну і тестову за декількома варіантами: 60/40, 70/30, 80/20 та 90/10. Задля оптимізації процесу створюємо функцію get\_metrics\_baseline, яка для кожного класифікатора та варіанту розбиття прорахує якість одержаної класифікації:





1. Результати якості класифікації оцінюємо за критеріями: precision, recall, f1-score та auc.
2. Одержали такі результати по моделям:

*Байєсівський класифікатор GaussianNB*



*Байєсівський класифікатор* [*BernoulliNB*](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#_blank)



*Байєсівський класифікатор MultinomialNB*



*Класифікатор GradientBoostingClassifier*



*Класифікатор SVC*

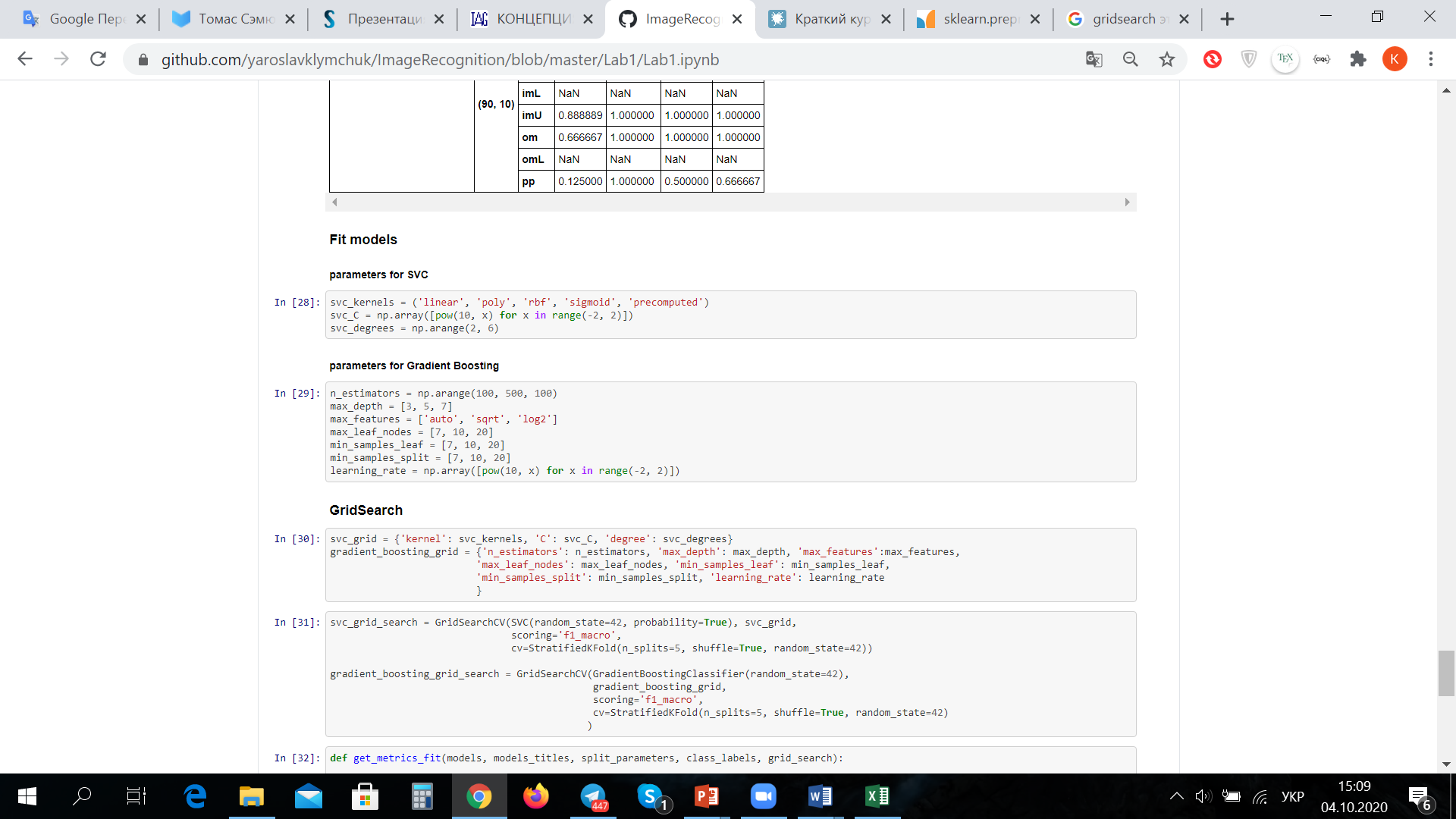


1. Висновки до baseline моделей:

Найкращі результати показали, незалежно від розбиття на тренувальну та тестову вибірки, такі моделі: SVC, GradientBoostingClassifier, NaiveBayess(GaussianNB), найгірші — NaiveBayess(MultinomianlNB) та NaiveBayess(BernoulliNB).

Характер розбиття на тренувальну та тестову вибірки також суттєво впливає на метрики на тестовій вибірці: чим більше даних використовуємо для навчання, тим більш сильно відбувається процес перенавчання моделі.

1. Результати для моделей з підбором кращих параметрів (GridSearch):
2. Щоб покращити результати класифікації застосували метод знаходження кращих гіперпараметрів моделі – GridSearch. Оскільки для байєсівського класифікатор вибір параметрів досить обмежений, то підбір параметрів будемо здійснювати лише для GradientBoostingClassifier та SVC:



1. Розраховуємо оцінки якості одержаних класифікаторів:

*Класифікатор GradientBoostingClassifier*



*Класифікатор SVC*



1. Висновки до моделей з підбором параметрів:

Як бачимо, в цілому підбір гіперпараметрів в моделях не має великого впливу на метрики класифікації. Це зумовлене маленькою кількістю об’єктів в вибірці, що використовується для навчання моделей.