# РЕШЕНИЕ НА ЗАДАЧАТА

## Анализ и предобработка на данни за машинно обучение

| Име | Възраст | Град | Използва ИИ продукти? | Интерес към нови технологии |
|---|---|---|---|---|
| Петър | 34 | София | Да | Висок |
| Иван | 19 | Варна | Не | Среден |
| Мария | 45 | Пловдив | Да | Нисък |

1. Оценка на качеството на данните:
    - Липсващи стойности – NA, празни клетки;
    - Несъответствия.

Николай Вецов: Липсват празни клетки, но информацията в таблицата изглежда недостатъчна. Може да се добави още (напр. фамилия, квартал на съответния град, колко често се използват ИИ продукти, в кои дни се използват и в кои не, към какви технологии конкретно се интересуват изследваните потребители и т.н. Не виждам несъответствия от сорта на правописни грешки или да е изписана информацията по погрешен начин.

2. Почистване:
    - Запълване на липсващи стойности или премахването им (в зависимост от контекста);
    - Уеднаквяване на категориите (пример: всичко „да“ да стане на „Да“ с главна буква);
    - Премахване на ненужни колони (напр. Име, ако не носи смислова стойност)

Не виждам да има липсващи стойности, ненужни колони и разминавания в текстовете между отделните клетки.

3. Трансформация на данните (числови стойности

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Name | Age | Uses AI | Interest in High Technologies | City_Sofia | City_Varna | City_Plovdiv |
| 2 | Peter | 34 | 1 | 2 | TRUE | FALSE | FALSE |
| 3 | Ivan | 19 | 0 | 1 | FALSE | TRUE | FALSE |
| 4 | Maria | 45 | 1 | 0 | FALSE | FALSE | TRUE |
| 5 | | | | | | | |

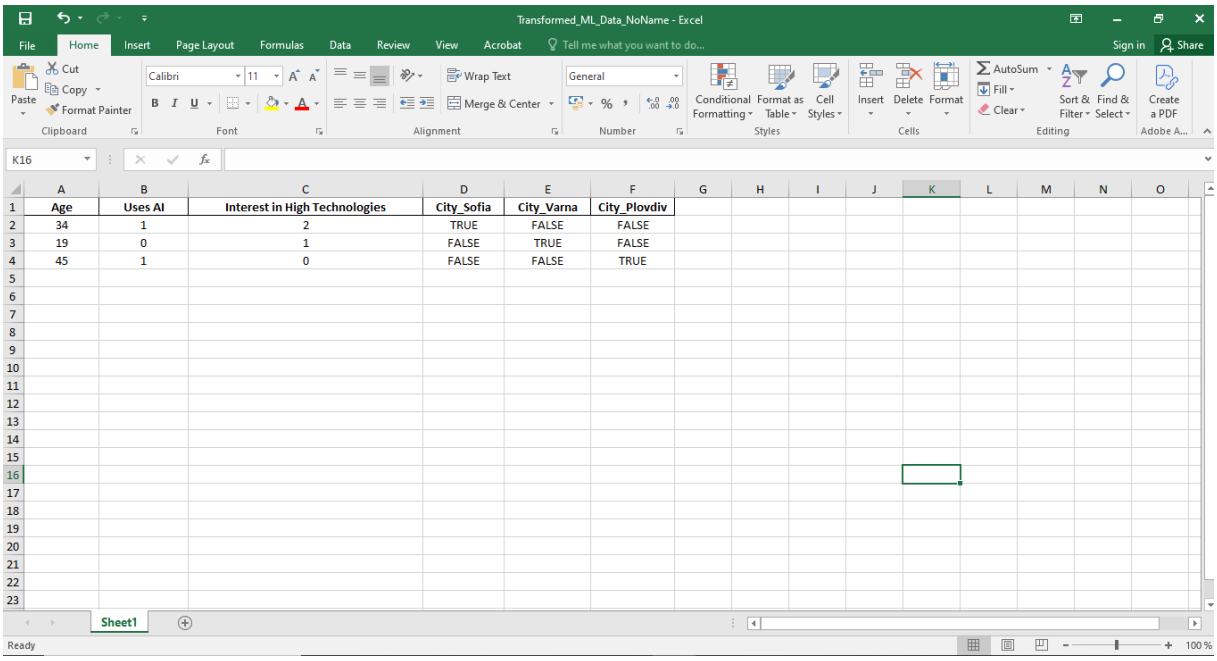
Николай Вецов: Това са трансформираните данни (Transformed_ML_Data.slsx). По този начин, те са подходящи за обучение на модел (машинно обучение). Всичко това е генерирано, чрез употребата на изкуствен интелект, който генерира Python код, имащ за цел да генерира чисто нов Excel файл, съдържащ данните така, че да бъдат подходящи за Машинно обучение.

При колоната – използване на AI, този уникален потребител, който използва (т.е Да) се използва цифрата 1 за „Да“ или цифрата 0 за „Не“ – този потребител, който не използва AI в този пример.

При колоната „Интерес към високи технологии“ – отговорите са три – „Висок“, „Среден“ или „Нисък“ интерес, но в този случай, за да бъде по-математически изградено, се използват отговори като за нисък = 0, за среден = 1 и за висок = 2. Този вариант е подходящ за обучение на модел, чрез алгоритъм. (Машинно обучение).

ВАЖНО: Според това, как реагира даден модел при машинното обучение, е правилно моделите да не бъдат обучавани на имена на потребители, тъй като това е по-лична информация, а не е добре и е много голям риск да се споделят такива данни. Затова, според препоръките, които получавам като Junior AI, е правилно да се скрият имената, както и Name, преди тази информация да бъде поднесена на модела, който ще се употребява за обучение. За тази цел, давам инструкции на изкуствения интелект да ми създаде нов код (Python), който има за цел да промени таблицата вътре в Excel файла (Transformed_ML_Data.xlsx) и да генерира нов

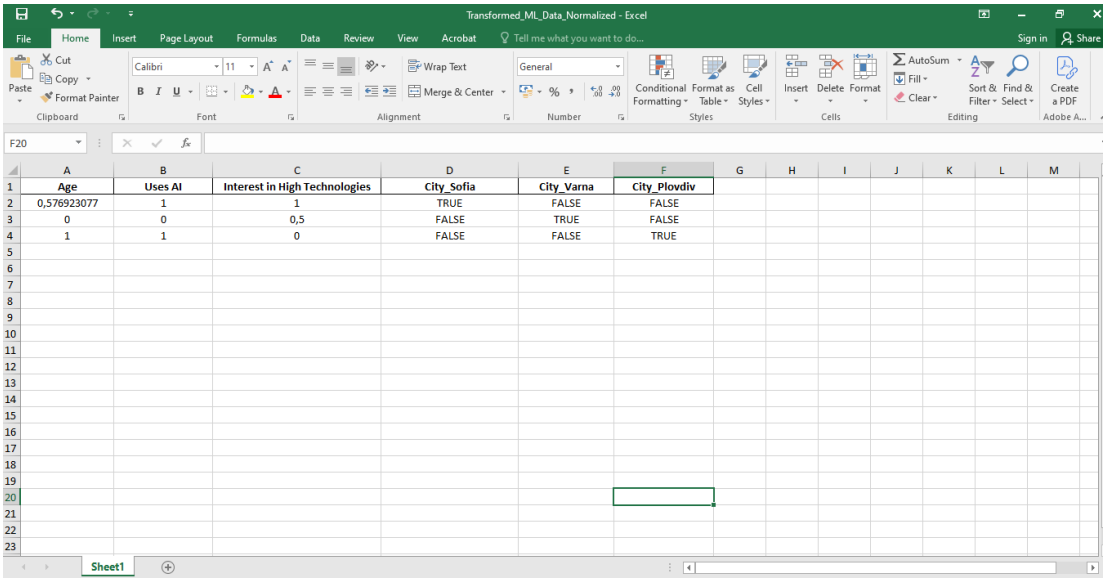
такъв и променен файл (Transformed_ML_Data_NoName.xlsx), който се намира и в директорията на проекта ми. Ето и визуализация на новата таблица в новия Excel файл:



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Age | Uses AI | Interest in High Technologies | City_Sofia | City_Varna | City_Plovdiv |
| 2 | 34 | 1 | 2 | TRUE | FALSE | FALSE |
| 3 | 19 | 0 | 1 | FALSE | TRUE | FALSE |
| 4 | 45 | 1 | 0 | FALSE | FALSE | TRUE |

По този начин, моделът, на база на избрания алгоритъм, може да бъде обучен, без да има рискове за разпространяване на по-лична информация. А иначе TRUE (вярно) и FALSE (невярно) е още един вариант, вместо 0 и 1, което може да бъде по-добър вариант за разбиране на стойности.

4. Нормализация:



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Age | Uses AI | Interest in High Technologies | City_Sofia | City_Varna | City_Plovdiv |
| 2 | 0,576923077 | 1 | 1 | TRUE | FALSE | FALSE |
| 3 | 0 | 0 | 0,5 | FALSE | TRUE | FALSE |
| 4 | 1 | 1 | 0 | FALSE | FALSE | TRUE |

Тази техника служи за улеснение на модела, за да може да разбира информацията, което служи за неговото обучение (Machine Learning). Използва се двоична бройна система (0 и 1). Или при повече варианти например: 0, 0.5 и 1. А това, че се използват и думи като TRUE и FALSE, е още един вариант за разбиране от страна на модела.

5. Експлоративен анализ (анализ на изследването):

| Age | Uses AI | Interest in High Technologies | City_Sofia | City_Varna | City_Plovdiv |
|---|---|---|---|---|---|
| 0,576923077 | 1 | 1 | TRUE | FALSE | FALSE |
| 0 | 0 | 0,5 | FALSE | TRUE | FALSE |
| 1 | 1 | 0 | FALSE | FALSE | TRUE |

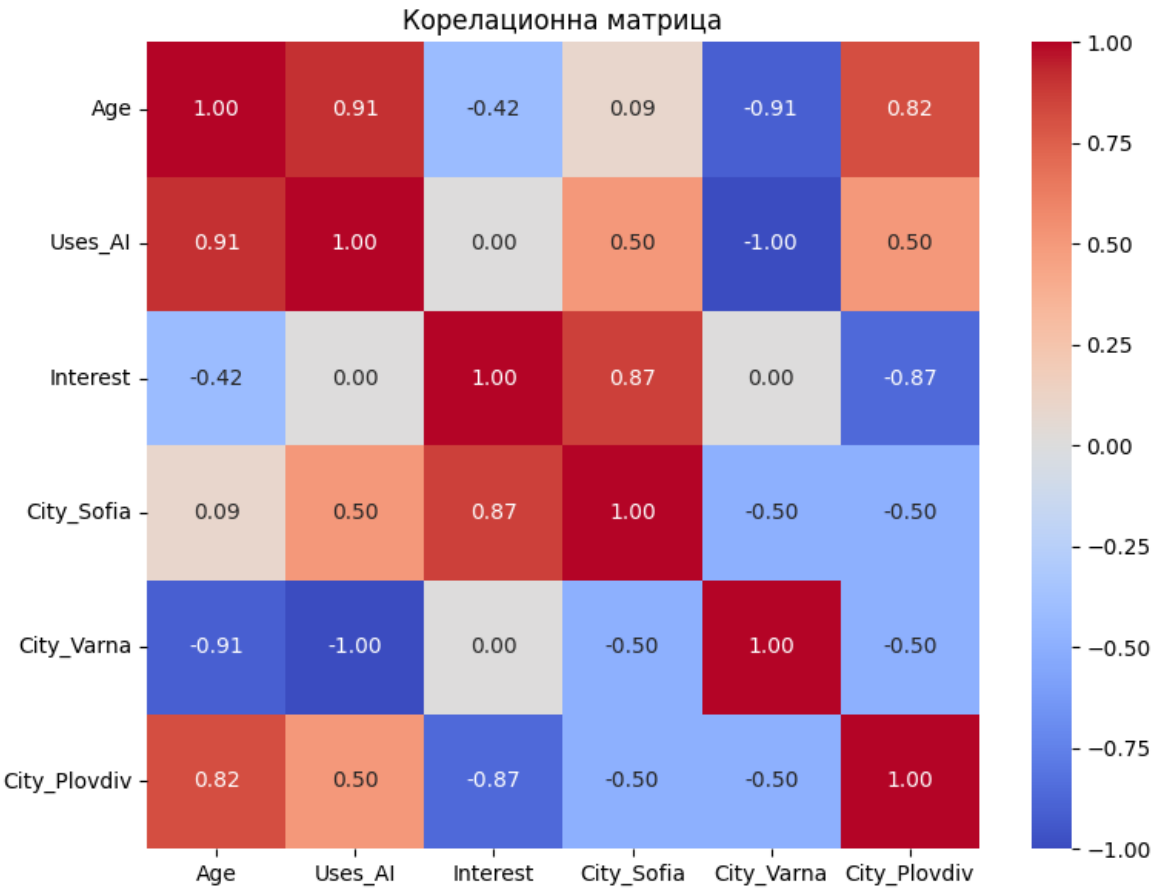
Така, таблицата е попълнена, а стойностите са подходящи за извършване на машинно обучение.

Age (възраст): 34 (0,58), 19 (0) и 45 (1).

***34 + 19 + 45 = 98 / 3 = 32,67 години е средната възраст.***

- Най-възрастният потребител, когото изследваме е на 45 години, най-младият е на 19 години.
- Третият потребител е на 34 години. (по средата).
- Тук, потребителите, които се изследват, са на млада възраст предимно.
- Различни нива на интерес (малък, среден, голям).
- Двама от потребителите използват изкуствения интелект при нужда, докато най-младият – не, въпреки, че има средно голям интерес към технологии (66,6%).
- Най-възрастният използва изкуствен интелект, вероятно е видял за какво става въпрос и е решил, че си струва да го използва, когато прецени, че се налага.
- Всеки един от тях е от голям град (София, Пловдив и Варна).

Корелационна матрица:




| Възраст | Град | Използва ИИ продукти? | Интерес към нови технологии |
|---|---|---|---|
| 34 | София | Да | Висок |
| 19 | Варна | Не | Среден |
| 45 | Пловдив | Да | Нисък |

За съжаление, информацията, с която разполагам, е твърде недостатъчна и един модел ще бъде обучен на много слабо ниво. Но, какво се оказва?

- Интересът към технологии е много нисък за град Пловдив (-0.87) – тук говорим за отрицателна корелация – нисък

резултат, докато например за София е наистина висок, както е описано на матрицата – 0.87 е висока оценка, висока променлива, почти перфектна положителна променлива (почти до 1.00). А вече за Пловдив е по средата – 0, но не се има предвид тук това, че е 0 – има корелация, а просто оценката така е зададена от изкуствения интелект;

- Колкото до употребата на Изкуствен Интелект – най-вероятно хората във Варна наистина не използват ИИ (или поне по-голямата част от тях). Данните са твърде малко, но се предполага, че е така. За момента имаме перфектна отрицатена корелация, докато за София и Пловдив имаме 0.50 – положителна корелация;
- Възрастта – за София имаме корелация от 0.09, която е все пак положителна засега. За Пловдив имаме по-сериозна корелация – 0.82. Тук имаме по-възрастни хора, които изследваме, докато за Варна имаме доста млади хора (19 г.). Корелацията е почти перфектна отрицателна (- 0.91).

6. Подготовка за модела:

   Сега, следва трансформация на данните във вид, подходящ за модела, който ще обучим: Да не забравяме, важното е употребата на двоична бройна система (0 и 1). Това е начинът за успешно обучен модел.

| Възраст | Град | Интерес към технологии | Използва AI |
|---|---|---|---|
| 34 | София | Висок | Да |
| 19 | Варна | Среден | Не |
| 45 | Пловдив | Нисък | Да |

   Данните ще възприемем като X, а последното – дали използват AI – като Y.

Ще създам таблица по следния начин, включващата двоична бройна система, необходима за да разбира моделът:

| Възраст | Интерес | Град София | Град Варна | Град Пловдив | Използва ИИ |
|---|---|---|---|---|---|
| 34 | 2 | 1 | 0 | 0 | 1 |
| 19 | 1 | 0 | 1 | 0 | 0 |
| 45 | 0 | 0 | 0 | 1 | 1 |

English Version

| Age | Interest | City_Sofia | City_Varna | City_Plovdiv | Uses_AI |
|---|---|---|---|---|---|
| 34 | 2 | 1 | 0 | 0 | 1 |
| 19 | 1 | 0 | 1 | 0 | 0 |
| 45 | 0 | 0 | 0 | 1 | 1 |

Така, променяме таблицата така, че да има само цифри от 0 до 1 (и не над 1), за да може да не стане така, че да се объркa моделът или да не направи нещо погрешно:

| Age | Interest | City_Sofia | City_Varna | City_Plovdiv | Uses_AI |
|---|---|---|---|---|---|
| 0.577 | 1 | 1 | 0 | 0 | 1.0 |
| 0.000 | 0.5 | 0 | 1 | 0 | 0.0 |
| 1.000 | 0 | 0 | 0 | 1 | 1.0 |

train_test_split.xlsx – В този файл са разделени данните, за обучение на модела:

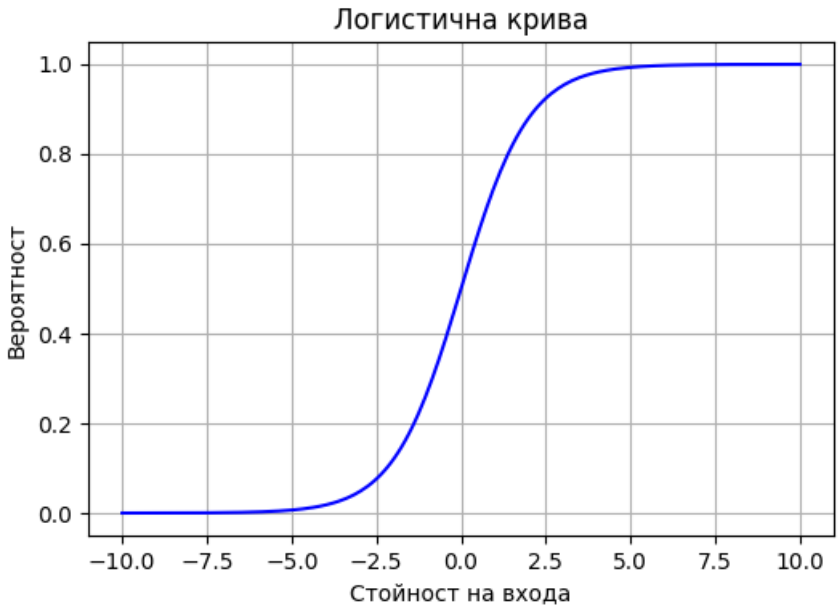
- Използване на **80% от данните за обучение** (`X_train`, `y_train`);
- Оставяне **20% за проверка на точността** (`X_test`, `y_test`);
- Възможност за повтаряемост.

# Резултатът ще изглежда така:

| Набор | Размер (% от данните) | Цел |
|---|---|---|
| `X_train` / `y_train` | ~80% | обучение на модела |
| `X_test` / `y_test` | ~20% | тестване на точността |

Създаваме модел, като използваме алгоритъм – Логистична регресия:

**Резултати от логистична регресия**

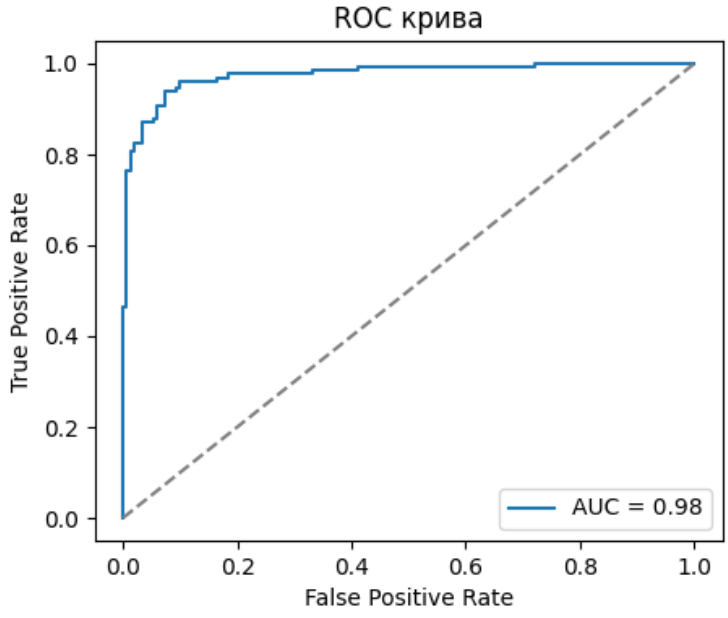



### Оценка на модела

Точност: 0.92, Предсказано: 1

Моделът предсказва с 0.92 от 1.00, във връзка с това дали човек употребява Изкусквен Интелект.

ROC: 0.98 от 1.00 – моделът почти перфектно разграничава потребитеите на ИИ от тези, които не го употребяват.



DATASET: Това е сетът от данни, обработен така, че да е по-подходящ за машинно обучение:

| Age | Interest | Uses_AI | City_Варна | City_Пловдив | City_София |
|---|---|---|---|---|---|
| 0,576923 | 2 | 1 | FALSE | FALSE | TRUE |
| 0 | 1 | 0 | TRUE | FALSE | FALSE |
| 1 | 0 | 1 | FALSE | TRUE | FALSE |

А този е още по-подходящ:

| Age | Interest | Uses_AI | City_Варна | City_Пловдив | City_София |
|---|---|---|---|---|---|
| 0,576923 | 2 | 1 | FALSE | FALSE | TRUE |
| 0 | 1 | 0 | TRUE | FALSE | FALSE |
| 1 | 0 | 1 | FALSE | TRUE | FALSE |

**Ето и още по-уместен и готов за машинно обучение, като тук вече е както трябва (двоична бройна система, подходяща за даден модел, тъй като има само цифри до 1.00 – и никога повече от това:**

| Age | Interest | City_Sofia | City_Varna | City_Plovdiv | Uses_AI |
|---|---|---|---|---|---|
| 0.577 | 1 | 1 | 0 | 0 | 1.0 |
| 0.000 | 0.5 | 0 | 1 | 0 | 0.0 |
| 1.000 | 0 | 0 | 0 | 1 | 1.0 |