TalTech Virumaa kolledž

ОБРАБОТКА ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ НА БАЗЕ PYTHON

Лектор: Ольга Дунаева olga.dunajeva@taltech.ee

24.05.2022

Содержание

- Модель машинного обучения.
- Среда Google Colab. Загрузка и обработка данных.
- Создание модели машинного обучения.
- Валидирование модели.
- Underfitting. Overfitting.
- Алгоритм Random Forest. Представление результатов.
- Kaggle: соревнования по машинному обучению

Использованные материалы:
https://www.kaggle.com/

Расписание и материалы

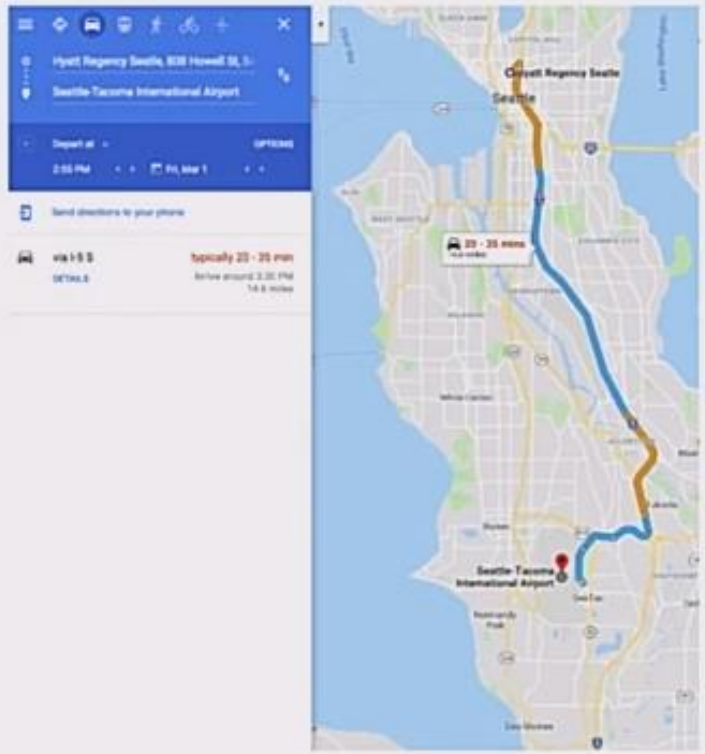
- 08:35 – 11:35 Andmetöötlus ja masinõpe Pythoni baasil, 4 акад. часа
- 12:00 -15:00 Robotite off-line programmeerimine, 4 акад. часа

Материалы

- https://data.vk.edu.ee/Andmetootlus/

ML IS EVERYWHERE

- Google predicts commute times.
- Apple predicts facial matches.
- Dozens of other ML-powered models in your phone today.

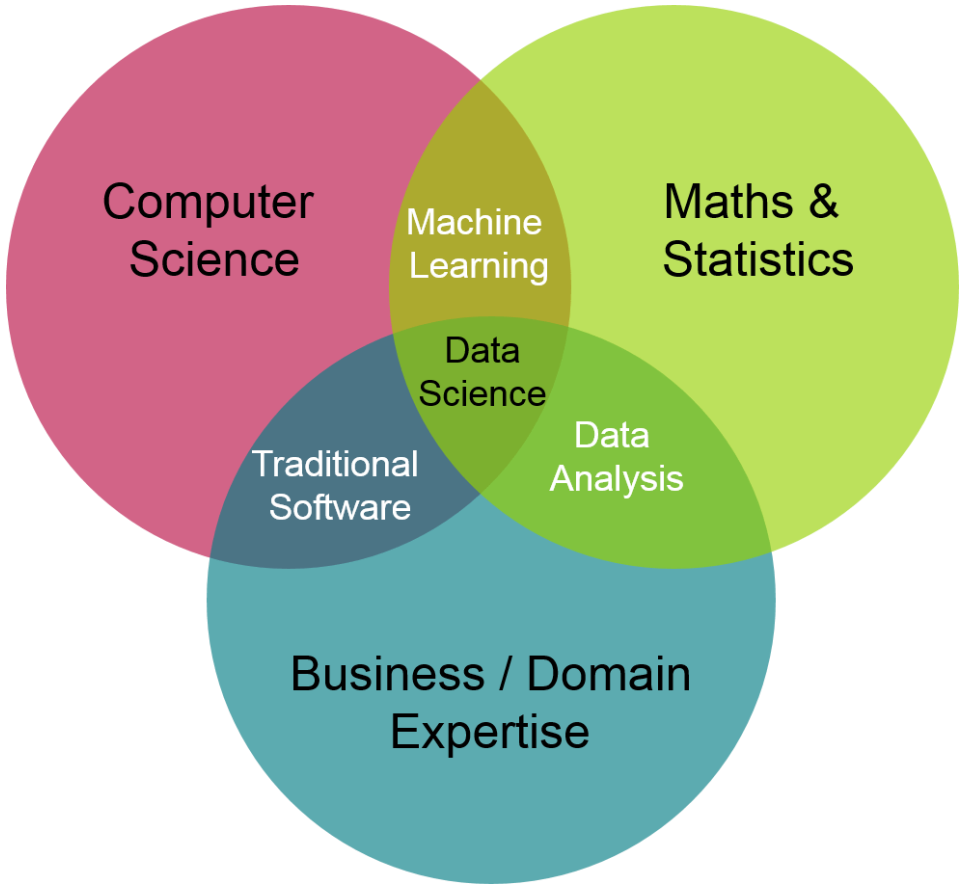


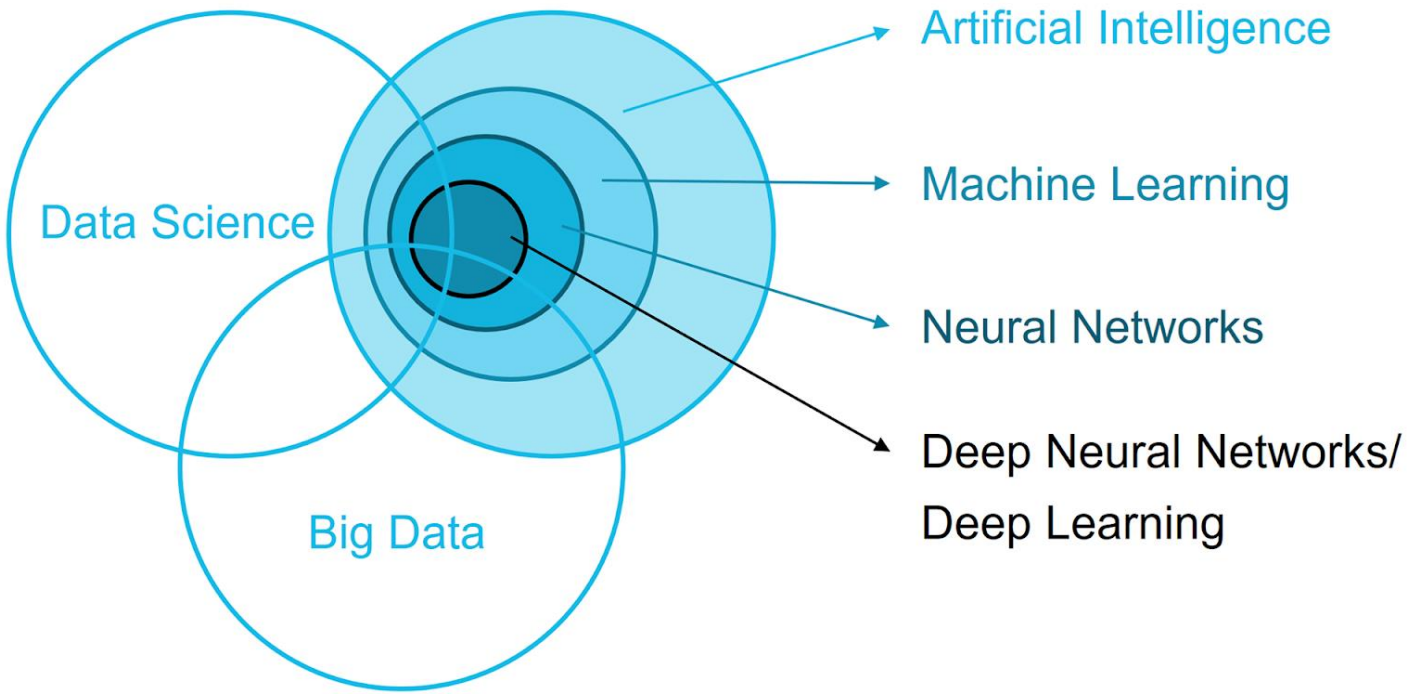
Google didn't set out to make a traffic tool.

Apple isn't in the facial recognition business.

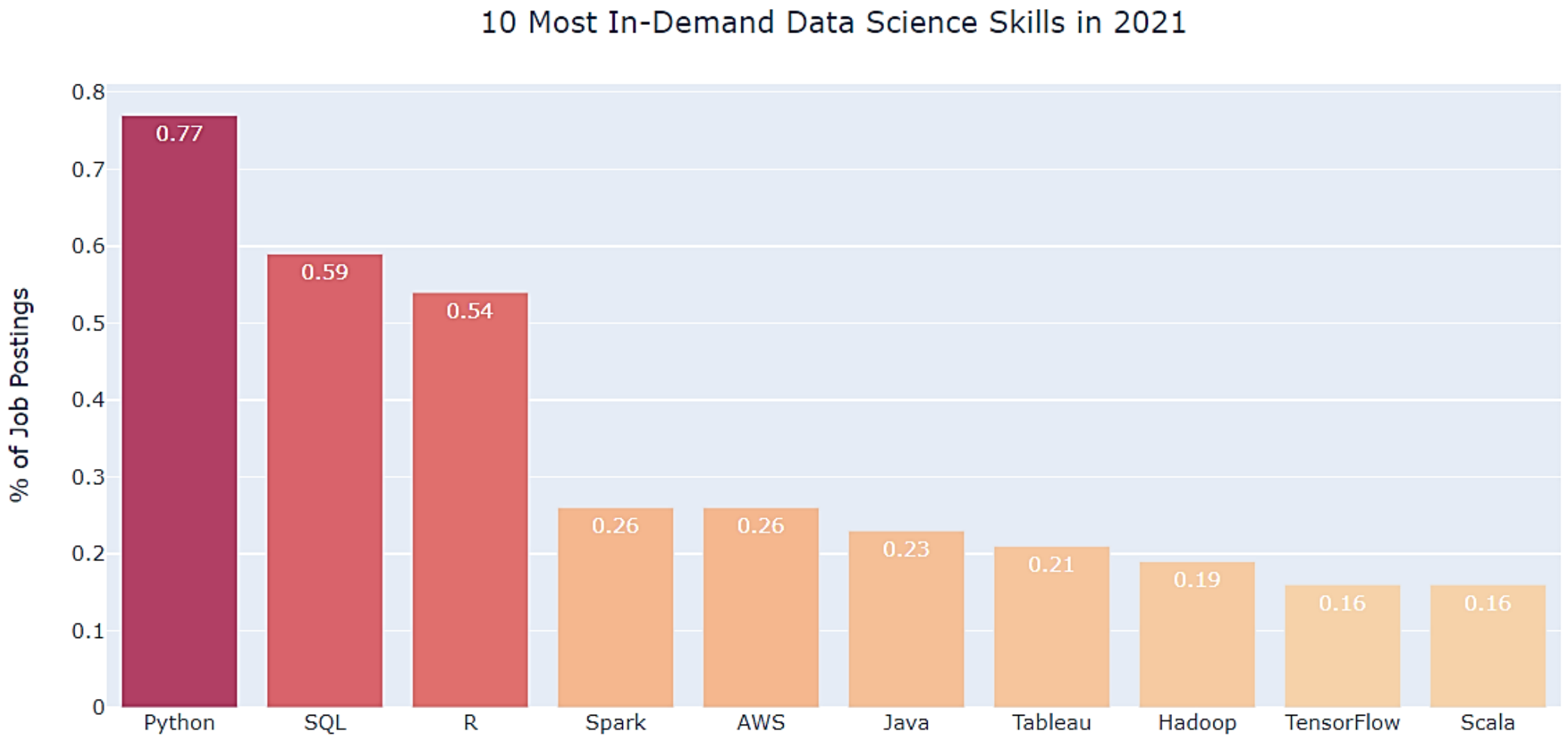
Face ID

Машинное обучение: навыки и разделы





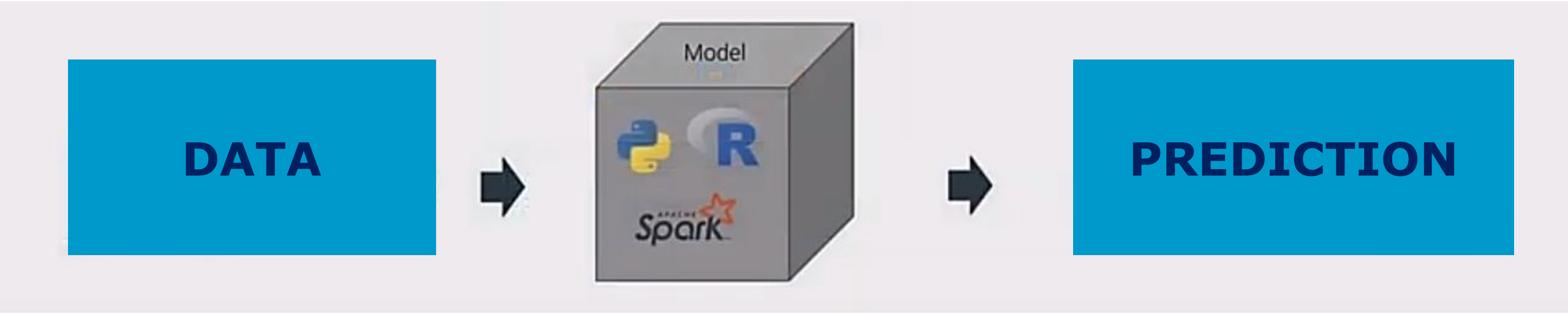
Инструменты работы с данными



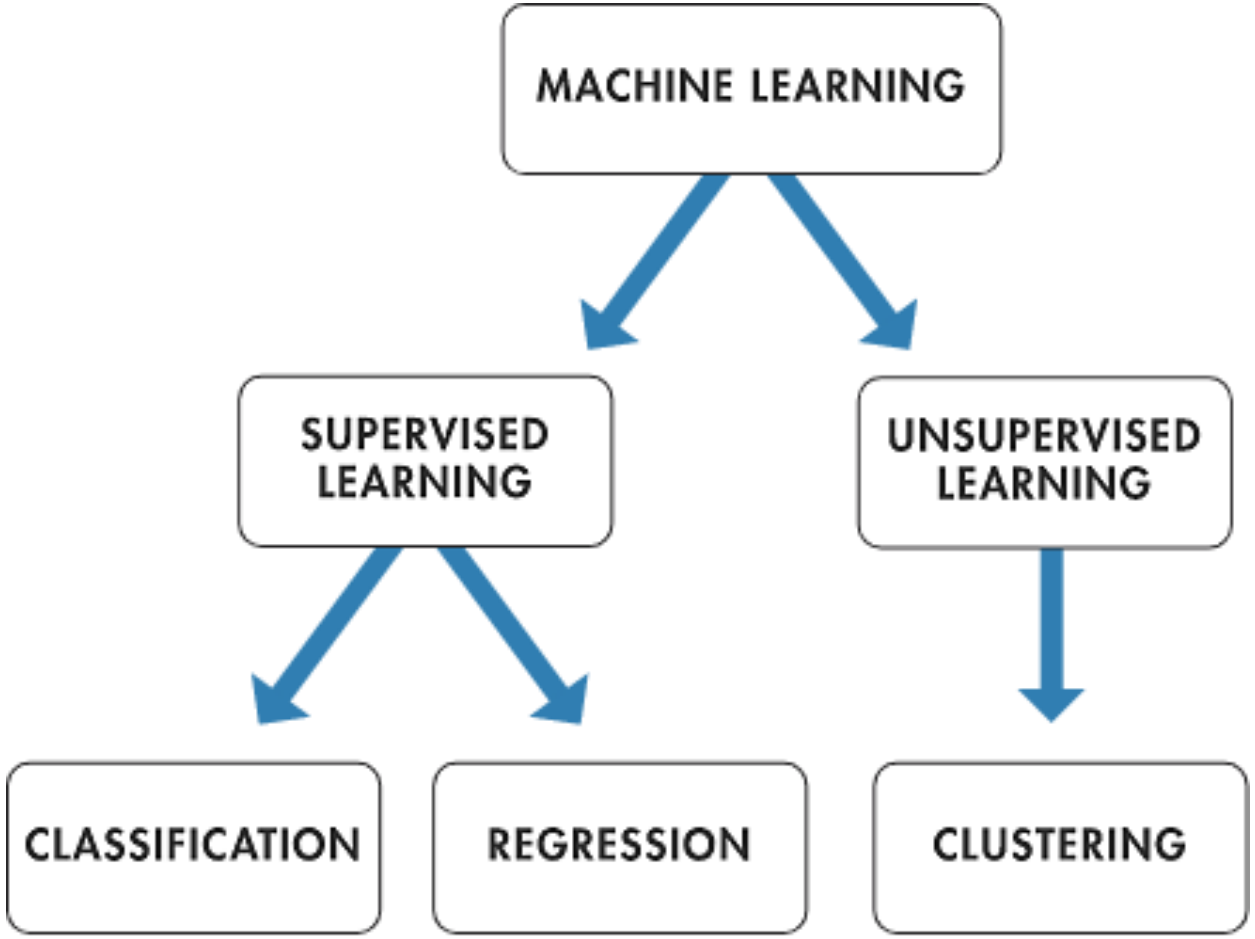
Results from webscraping over 15,000 Data Scientist job postings

https://towardsdatascience.com/

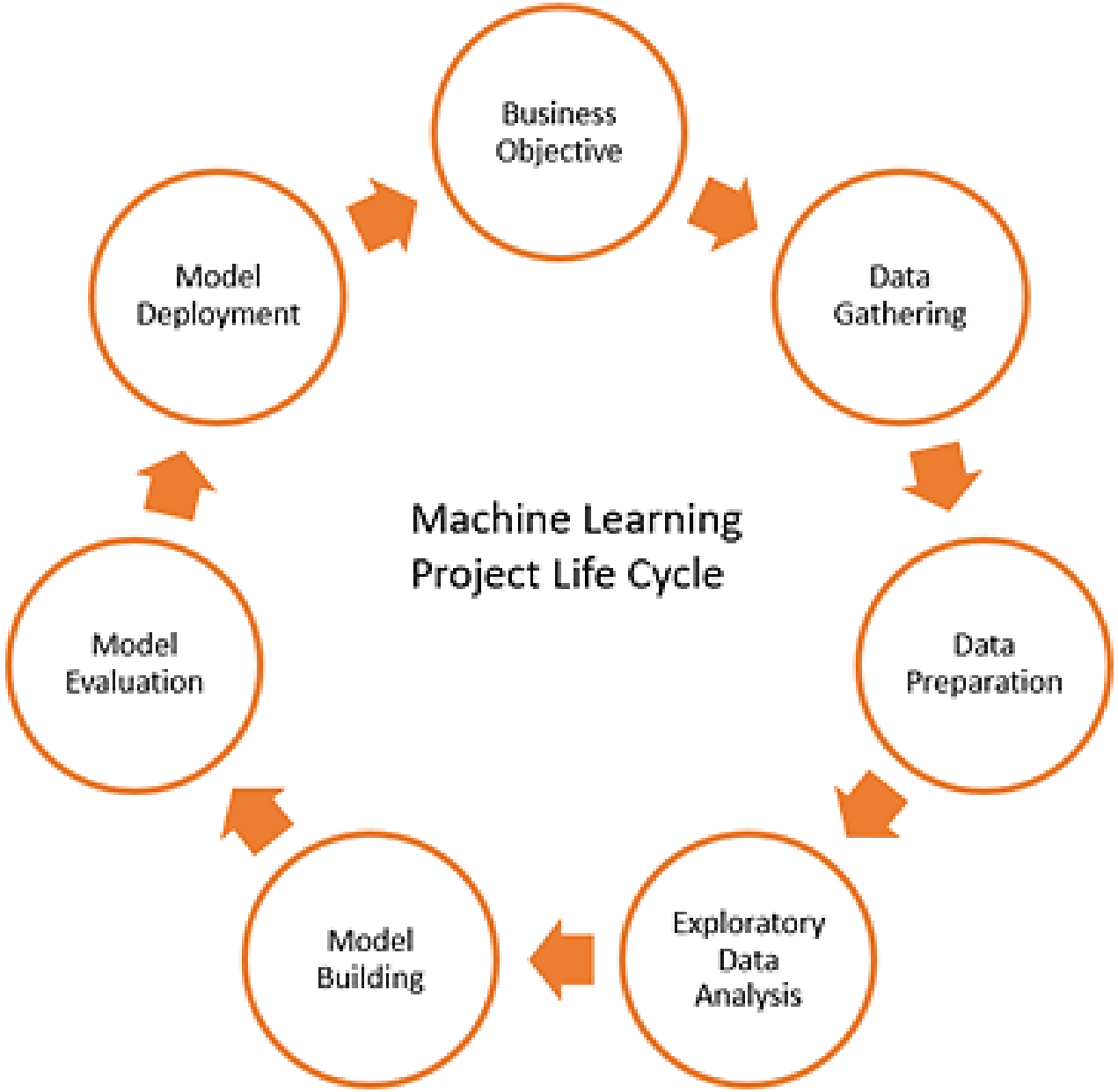
1. Модель машинного обучения



Виды машинного обучения



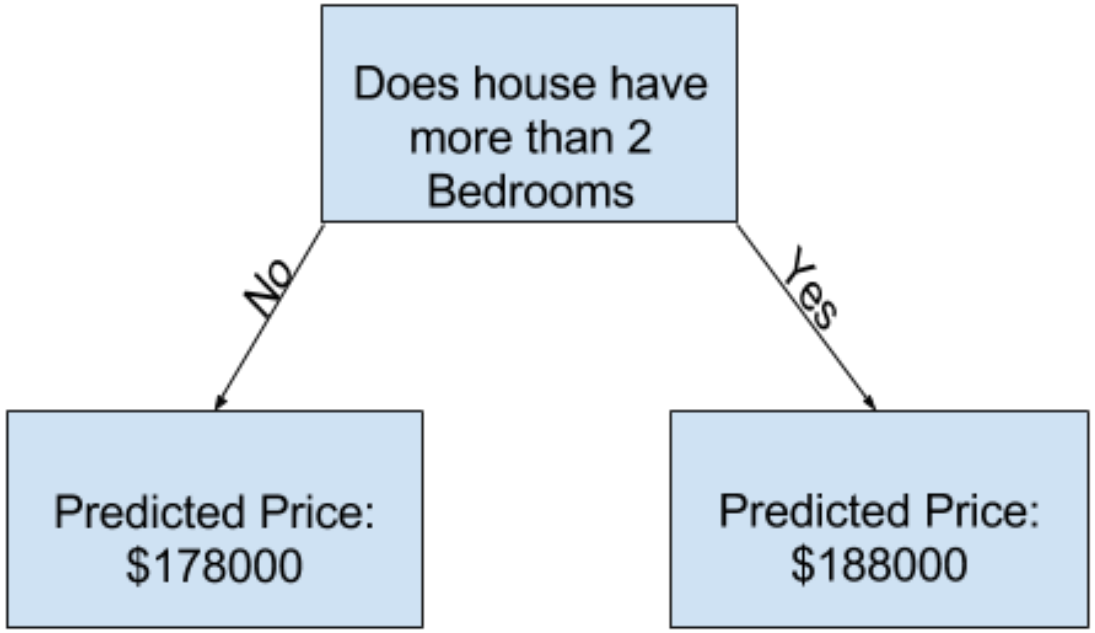
Проект машинного обучения



Сценарий проекта:

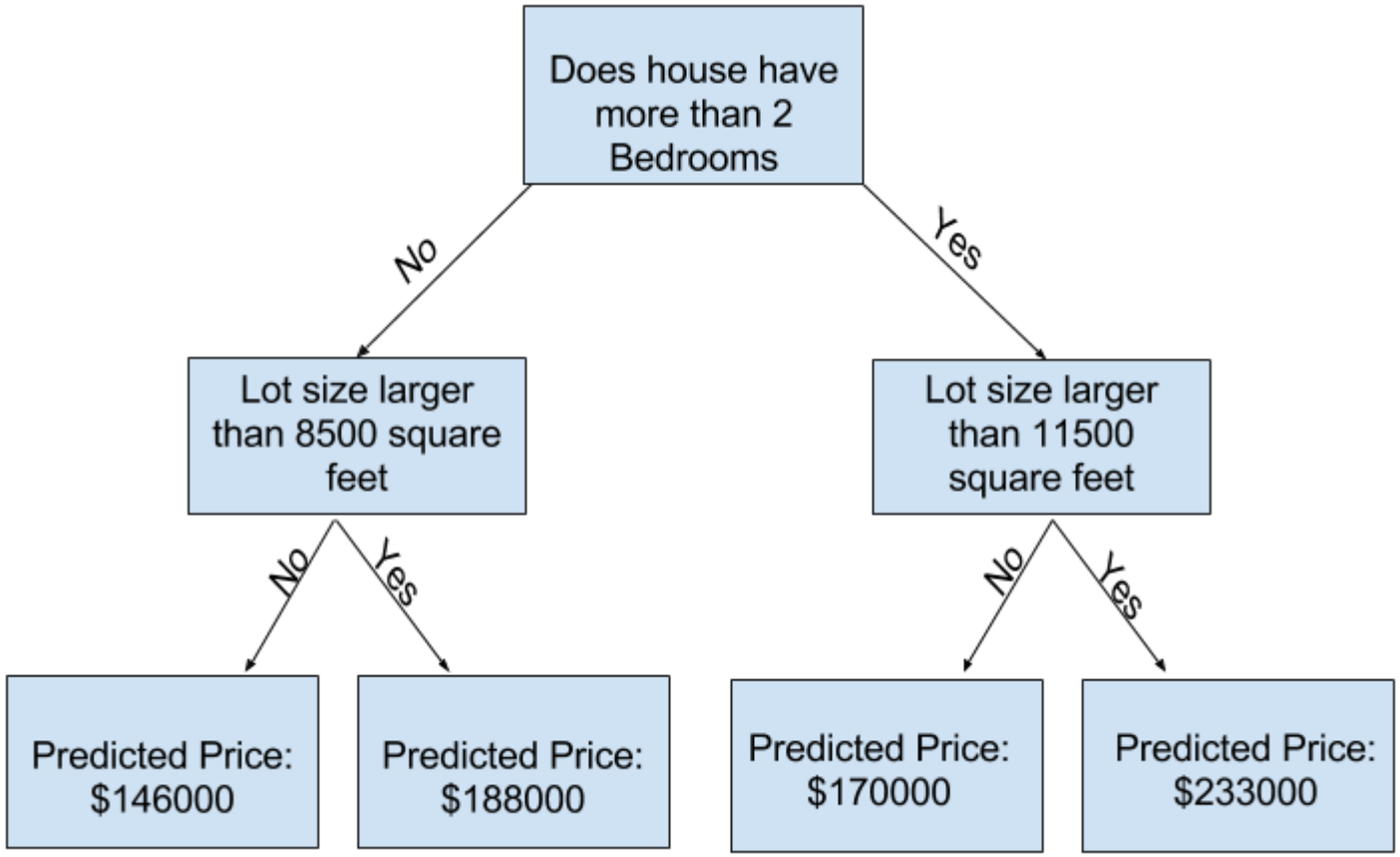
- Разработать модель для предсказания цен на недвижимость по имеющимся данным.
- Мы начнем с модели под названием «Дерево решений»
  - Есть более сложные модели, которые дают более точные прогнозы. Но деревья решений легко понять, и они дают хорошее представление о факторах, влияющих.

Простое дерево решений



- Прогнозируемая цена дома представляет собой среднюю цену домов той же категории.
- Подгонка или обучение модели (*fitting or training the model*) – использование **тренировочных данных** для создания модели.

Улучшение дерева решений (Improving the Decision Tree)



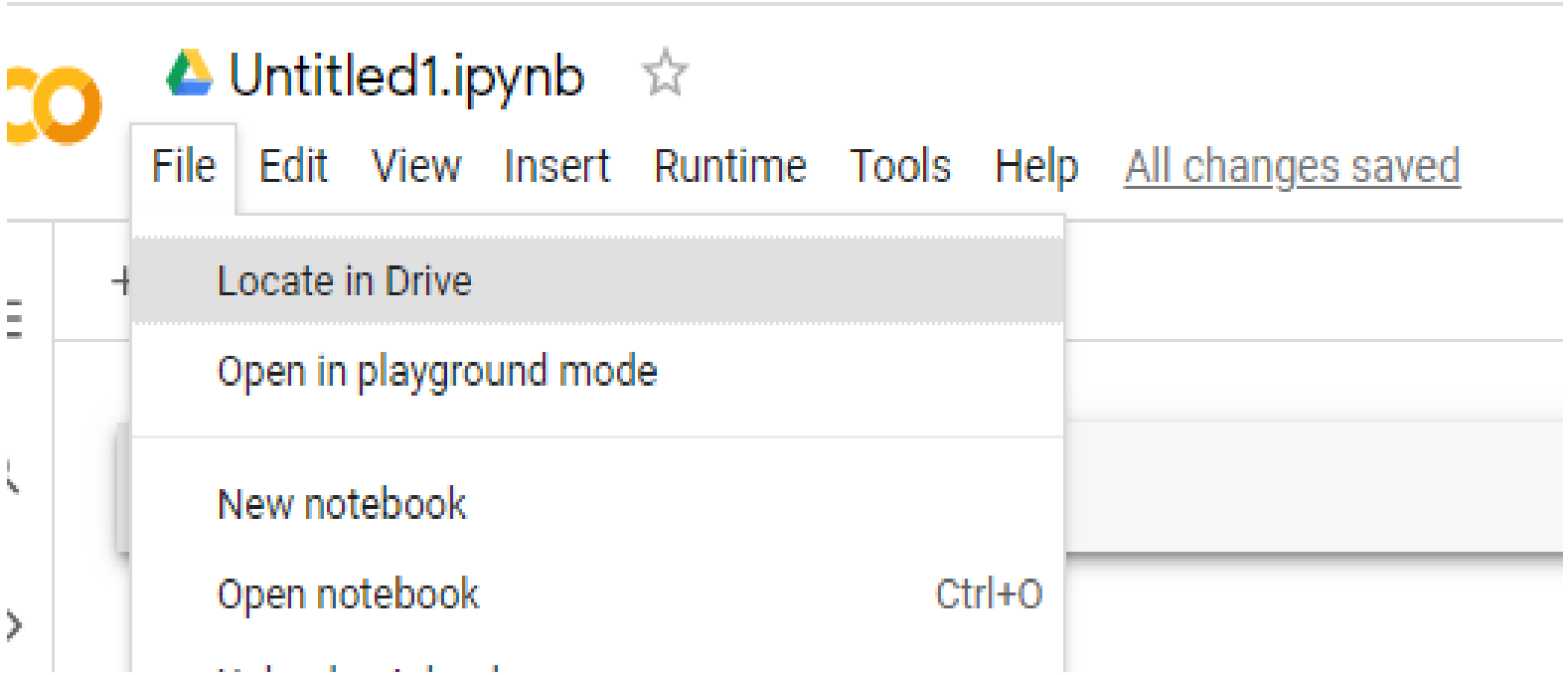
- Улучшить модель можно учитывая больше факторов, влияющих на цену дома (количество ванных комнат, размер участка, местоположение, …).
- Цена дома прогнозируется, просматривая дерево решений, выбирая путь, соответствующий характеристикам этого дома. Прогнозируемая цена дома находится в нижней части дерева. Точка прогноза называется листом.

2. Среда Google Colab. Загрузка и обработка данных

- Google Colab (Colaboratory) — это облачная среда, которая позволяет писать и выполнять код Python в браузере.
- Преимущества Google Colab:
  - бесплатный облачный сервис
  - не требует настройки
  - бесплатный доступ к графическим процессорам
  - легкий доступ к популярным библиотекам Python.

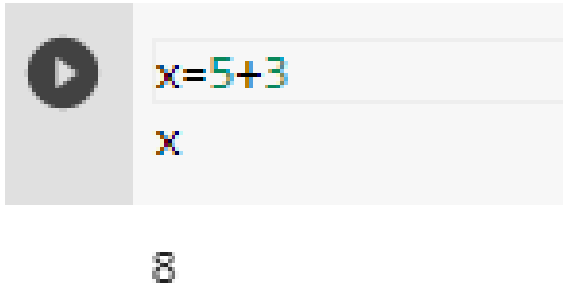
Блокнот (notebook) в Google Colab

- Откройте Google Colab https://colab.research.google.com/
- Создайте ‘New Notebook’
- Переименуйте и сохраните на Google Диске



Colab: написание и документирование кода

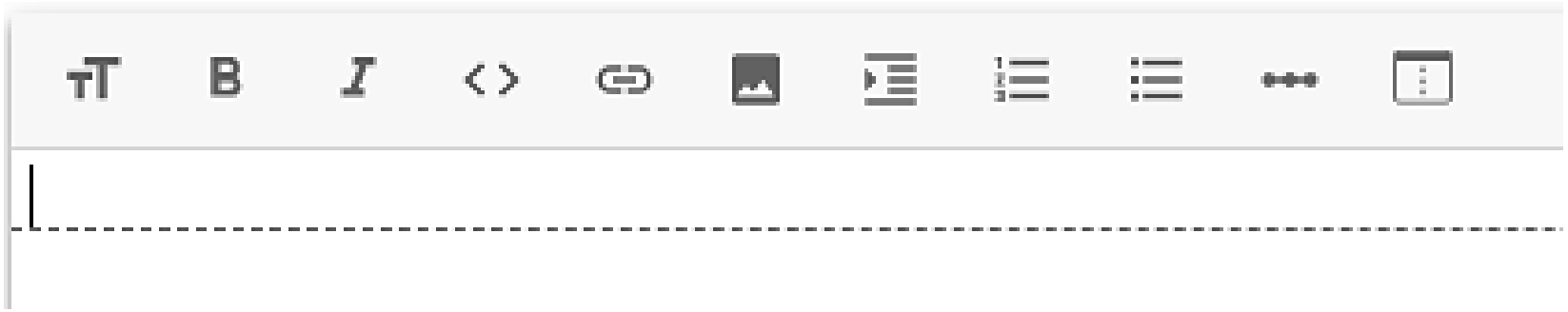
- Ячейка с кодом



- Чтобы выполнить код в ячейке, нажмите кнопку воспроизведения слева от кода или используйте сочетание клавиш Ctrl + Enter.

Colab: написание и документирование кода

- Ячейка с текстом



- Для редактирования – двойной щелчок на ячейке
- Используется язык разметки Markdown

  https://colab.research.google.com/notebooks/markdown_guide.ipynb

Загрузка данных

- Melbourne Housing Snapshot

  https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot данные получены с сайта Domain.com.au:

  - Rooms: Number of rooms
  - Price: Price in dollars
  - Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
  - Type: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
  - SellerG: Real Estate Agent
  - Date: Date sold
  - Distance: Distance from CBD
  - Regionname: General Region (West, North West, North, North east …etc)
  - Propertycount: Number of properties that exist in the suburb.
  - Bedroom2 : Scraped # of Bedrooms (from different source)
  - Bathroom: Number of Bathrooms
  - Car: Number of carspots
  - Landsize: Land Size
  - BuildingArea: Building Size
  - CouncilArea: Governing council for the area.

Загрузка данных

- Файл с данными melb_data.csv находится на сервере колледжа https://data.vk.edu.ee/Andmetootlus/
- Загрузите файл melb_data.csv в Google Colab