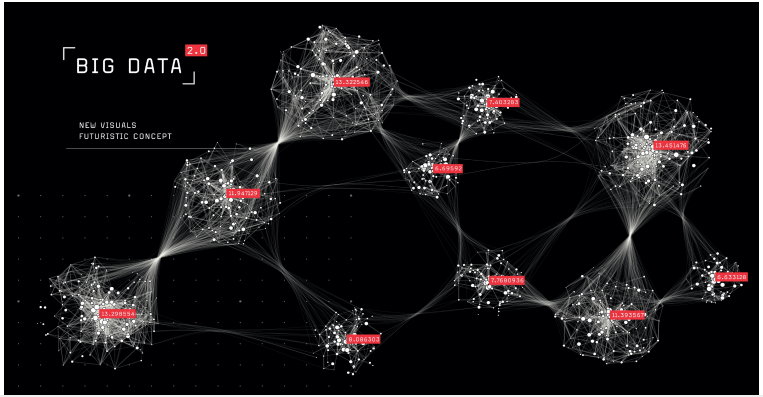
Data Science in Action using Python

An AAII Artificial Intelligence– Technical Track Course



Course Outline

Copyright © Applied AI Institute 2021

Objectives

- ✓ Exposure to proven methodology
- ✓ Hands-on course, a solution you can use
- ✓ Program for Coders

Copyright © Applied AI Institute 2021

Prerequisites

- Knowledge of Python desired but not required
- We will gradually introduce 7 major Python libraries over the duration of the course.
- We will also provide recommendations for advanced Python learning

Copyright © Applied AI Institute 2021

Audience

This course is for anyone interested in becoming a data scientist

- Students
- Business Analysts
- Developers
- Testing professionals

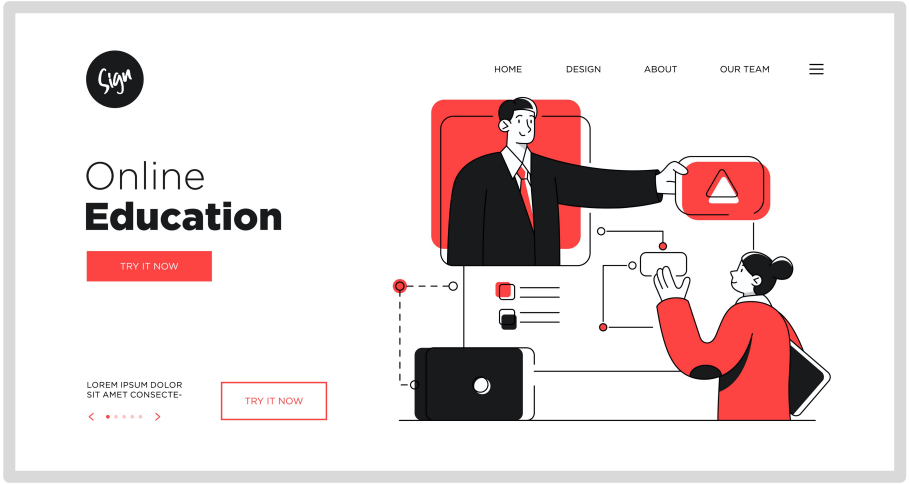
Related careers:

- Data Scientists
- Data / AI / Automation Engineers
- Test Engineers
- Knowledge Engineers

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox



Copyright © Applied AI Institute 2021

Data Science in Action using Python

	Pandas	Matplotlib	Numpy	Stats Models	Scikit Learn	Networkx	ONNX
Step 1: Describe Use Case							
Step 2: Describe Data	√						
Step 3: Prepare Data	√						
Step 4: Develop Model	√	√	√	√	√	√	
Step 5: Evaluate Model	√	√	√	√	√		
Step 6: Deploy Model	√		√		√		√
Step 7: Monitor Model							

Copyright © Applied AI Institute 2021

Python Basics

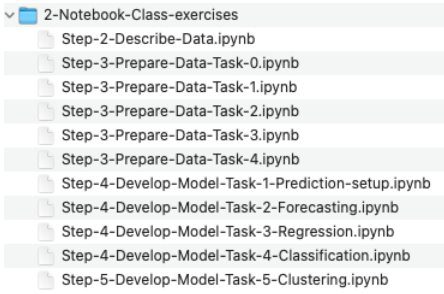
1. Python variables assignment
2. Python and user-defined functions
3. Common data types:

- Numeric Types: int, float
- Text Type: str
- Sequence Type: list
- Mapping Type (key-value): dict
- Boolean Type: bool
- Date type: date

```
In [1]: import pandas as pd
        from datetime import date

In [21]: population =50
         type(population)
Out[21]: int

In [14]: average_cases = 31.7
         type(average_cases)
Out[14]: float

In [15]: countyName = "Orange County"
         type(countyName)
Out[15]: str

In [16]: days_of_week = ["Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"]
         type(days_of_week)
Out[16]: list

In [17]: column_list = {"column 1": "Date", "Column 2": "Incremental Cases"}
         type(column_list)
Out[17]: dict

In [18]: mask_policy = False
         type(mask_policy)
Out[18]: bool

In [20]: from datetime import date
         today = date(2020, 12, 29).isoformat()
         print(today)
         2020-12-29
```

Copyright © Applied AI Institute 2021

Notebook-Class-exercises



Contains multiple Notebook files to support COVID use case for each data science step

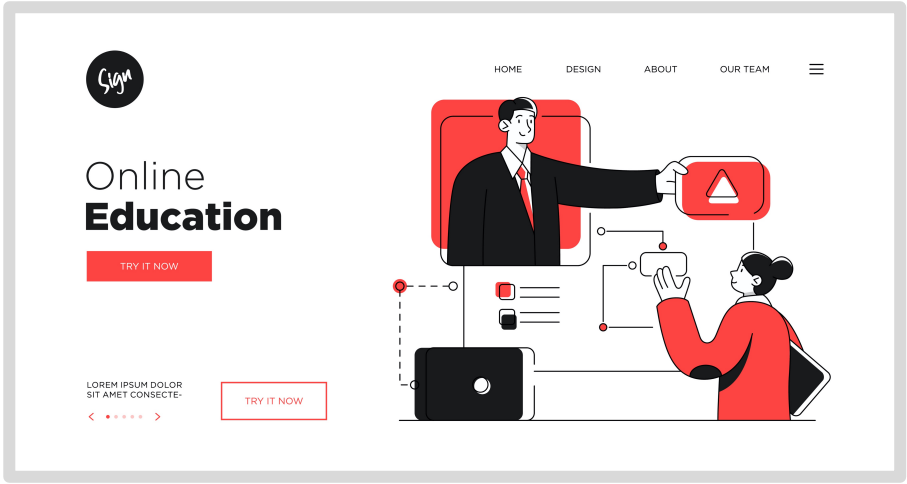
- Step-2-Describe-Data.ipynb → Notebook for Step 2 - Describe Data
- Step-3-Prepare-Data-Task-1.ipynb
- Step-3-Prepare-Data-Task-2.ipynb
- Step-3-Prepare-Data-Task-3.ipynb
- Step-3-Prepare-Data-Task-4.ipynb → Notebook for Step 3 – Prepare Data - Tasks 1 through 4
- Step-4-Develop-Model-Task-1-Prediction-setup.ipynb → Notebook for Step 4 – Develop Model - Tasks 1
- Step-4-Develop-Model-Task-2-Forecasting.ipynb → Notebook for Step 4 – Develop Model - Tasks 2
- Step-4-Develop-Model-Task-3-Regression.ipynb → Notebook for Step 4 – Develop Model - Tasks 3
- Step-4-Develop-Model-Task-4-Classification.ipynb → Notebook for Step 4 – Develop Model - Tasks 4
- Step-4-Develop-Model-Task-5-Clustering.ipynb → . Notebook for Step 4 – Develop Model - Tasks 5

And so on

Copyright © Applied AI Institute 2021

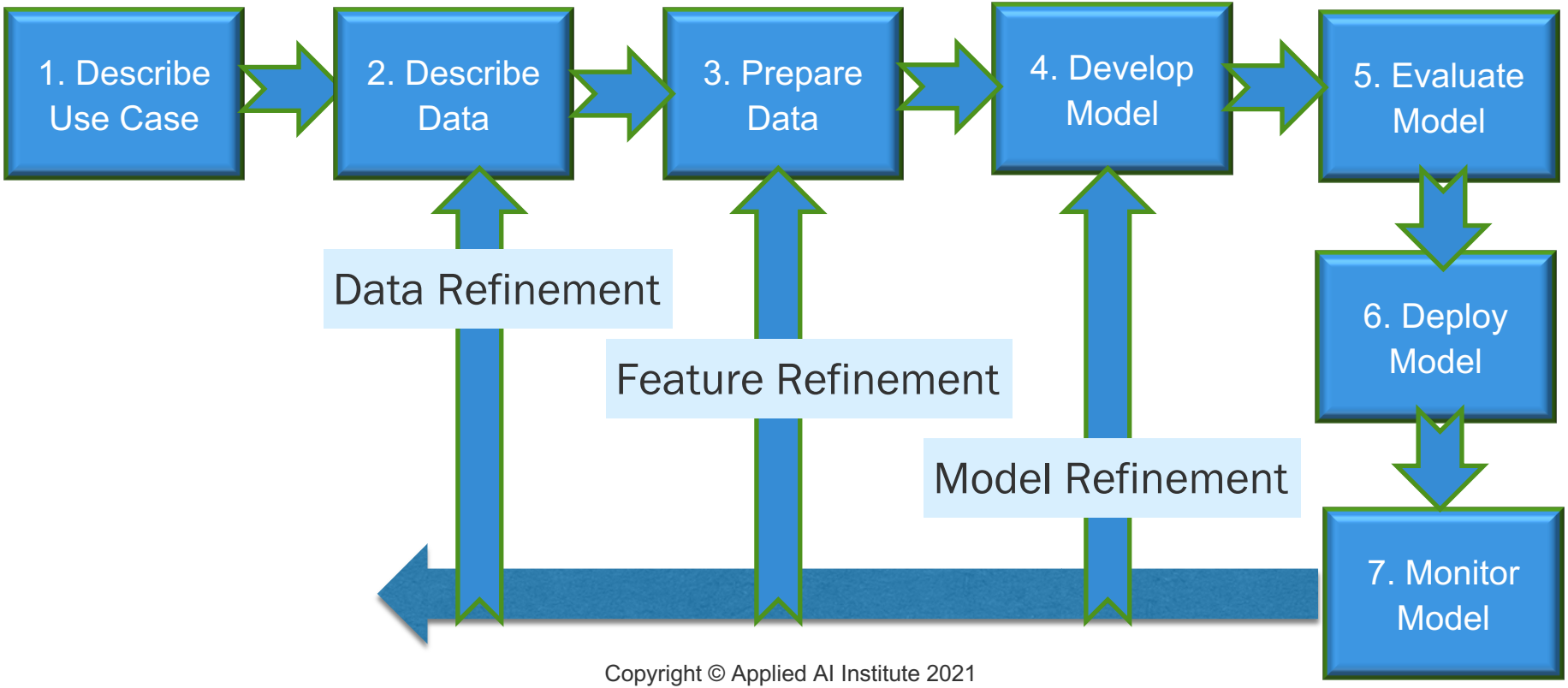
Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology



Copyright © Applied AI Institute 2021

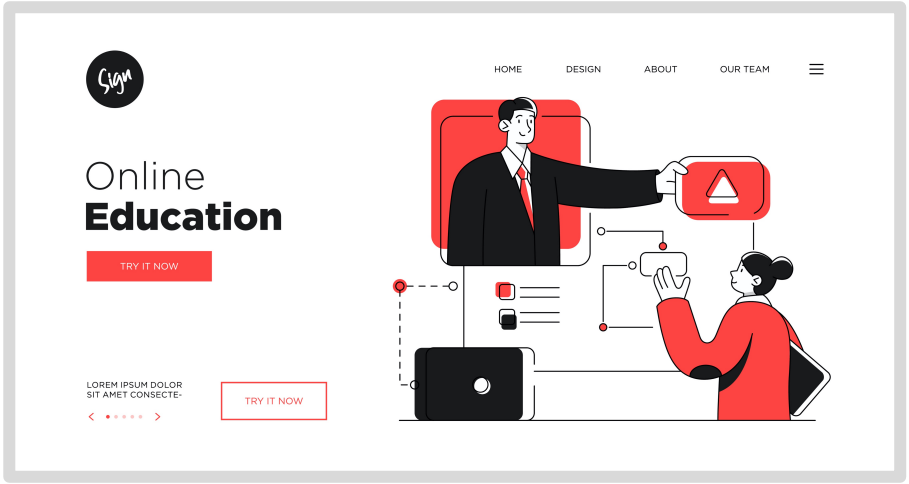
Data Science Methodology



Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology
4. Step 1 - Define Use Case



Copyright © Applied AI Institute 2021

Step 1 – Describe Use Case

Use Case Name: Coronavirus (COVID-19) Outbreak Forecast

Key Objectives

Develop a tool to predict rate of Coronavirus spread in a given region

Problem Overview

Analyze the COVID-19 data sources and forecast 2 weeks outbreak insights on projected cases by various factors such as Location, Mobility Population and other factors as needed

User Persona

- Medical Professionals of various countries trying to plan out the Coronavirus outbreak spread
- Families trying to plan out family vacation

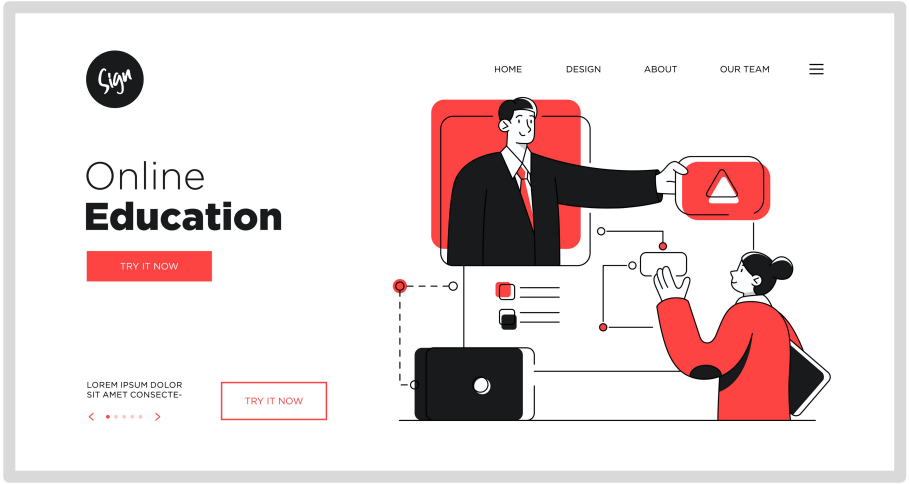
Business Benefits

- Build awareness of coronavirus spread among medical professional to help track the spread
- Manage hospital staff capacity adequately
- Provide proper care and support to patients
- Plan out family vacation to spots which are COVID safe and Maintain normal social life during the pandemic

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology
4. Step 1 - Define Use Case
5. Step 2 – Describe Data



Copyright © Applied AI Institute 2021

Step 2 - Describe data

Class Assignment – Instructions (Cont.)

DU2.2 Review data types included in this Data Set - “covid_confirmed_usafacts”

We are using column function to read columns included in pd data frame – df_covid_confirmed

```
In [ ]: # DU2.2 Review data types included in this  Data Set – "covid
```

```
In [23]: df_covid_confirmed.columns
```

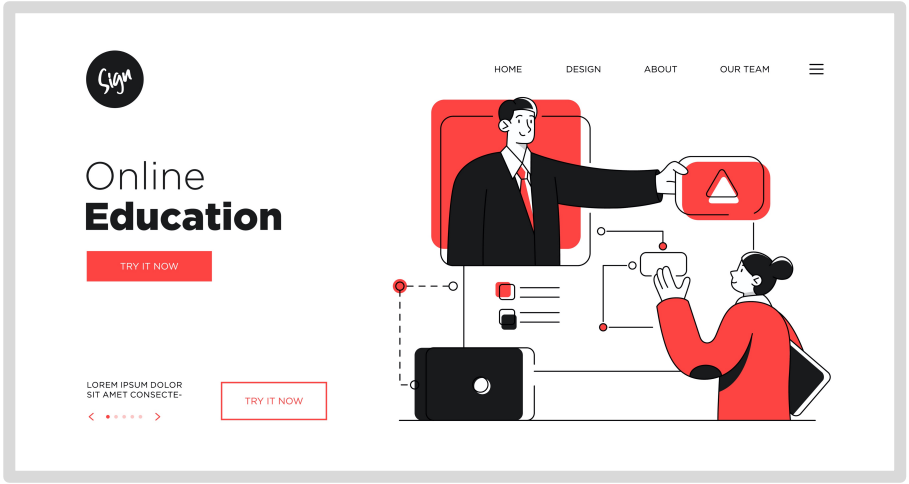
```
Out[23]: Index(['countyFIPS', 'County Name', 'State', 'stateFIPS', '1/22/20', '1/23/20',
       '1/24/20', '1/25/20', '1/26/20', '1/27/20',
       ...
       '12/1/20', '12/2/20', '12/3/20', '12/4/20', '12/5/20', '12/6/20',
       '12/7/20', '12/8/20', '12/9/20', '12/10/20'],
      dtype='object', length=328)
```

You will see that this data set includes fields CountyFIPS, County Name, State, StateFIPS and various date fields ranging from 1/22/20 through 12/14/20 for a total of 328 fields

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology
4. Step 1 - Define Use Case
5. Step 2 - Describe Data
6. Step 3 - Prepare Data



Copyright © Applied AI Institute 2021

Step 3 – Prepare Data

Original Format

```
In [3]: df_confirmed_cases = pd.read_csv('../input/covid_confirmed_usafacts.csv')
        df_confirmed_cases = df_confirmed_cases.astype({'countyFIPS': str}).astype({'stateFIPS': str})
        df_confirmed_cases
```

Out[3]:

	countyFIPS	County Name	State	stateFIPS	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/11/20	12/12/20	12/13/20	12/14/20	12/15/20	12/1(
0	0	Statewide Unallocated	AL	1	0	0	0	0	0	0	...	0	0	0	0	0	
1	1001	Autauga County	AL	1	0	0	0	0	0	0	...	3233	3233	3233	3329	3426	3
2	1003	Baldwin County	AL	1	0	0	0	0	0	0	...	10489	10489	10489	10898	11061	11
3	1005	Barbour County	AL	1	0	0	0	0	0	0	...	1264	1264	1264	1275	1292	1
4	1007	Bibb County	AL	1	0	0	0	0	0	0	...	1398	1398	1398	1455	1504	1

Transformed Format

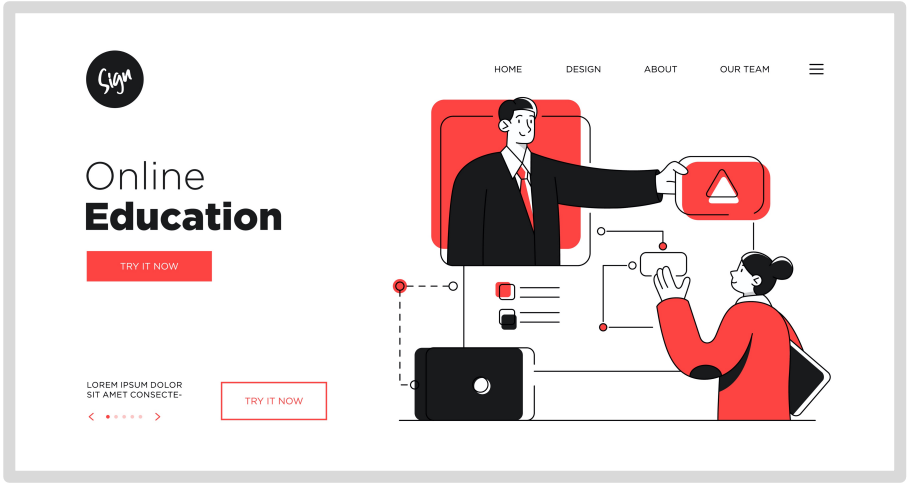
```
In [9]: df_google_mobility_data_selected = df_google_mobility_data_selected.dropna(subset=['countyFIPS'])
        df_google_mobility_data_selected = df_google_mobility_data_selected.astype({'countyFIPS': int})
        df_google_mobility_data_selected = df_google_mobility_data_selected.astype({'countyFIPS': str})
        df_google_mobility_data_selected
```

Out[9]:

	sub_region_1	countyFIPS	date	retail_and_recreation_percent_change_from_baseline	grocery_and_pharmacy_percent_change_from_baseline	parks_
630	Alabama	1001	2020-02-15	5.0	7.0	
631	Alabama	1001	2020-02-16	0.0	1.0	
632	Alabama	1001	2020-02-17	8.0	0.0	
633	Alabama	1001	2020-02-18	-2.0	0.0	
634	Alabama	1001	2020-02-19	-2.0	0.0	

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology
4. Step 1 - Define Use Case
5. Step 2 - Describe Data
6. Step 3 - Prepare Data
7. Step 4 – Develop Model

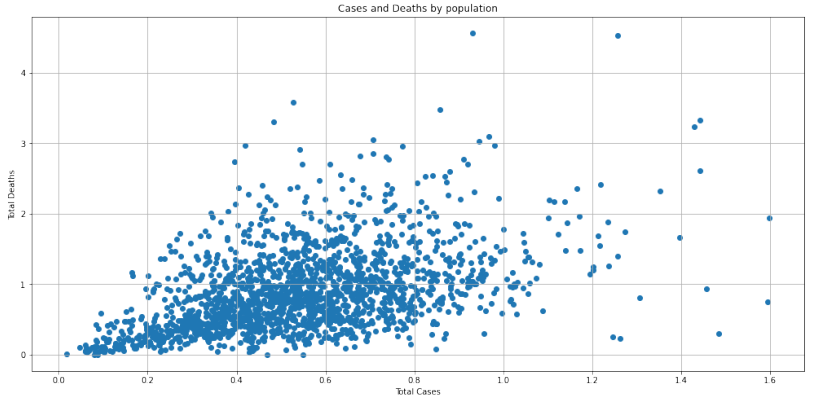


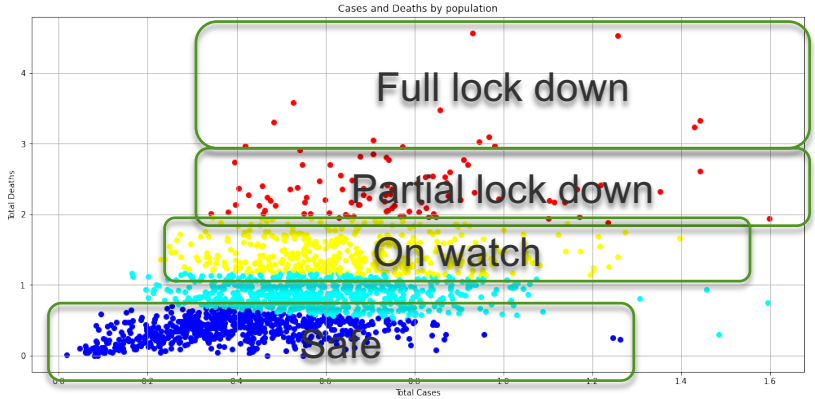
Copyright © Applied AI Institute 2021

Step 5 – Develop Model

Clustering technique – K-means

- Based on shortest distance
- Number of clusters specified as input
- Algorithm discovers cluster centers
- Optimizes on distance from center
- K-means filter uses mean distance
- Expert provides labels / attach actions





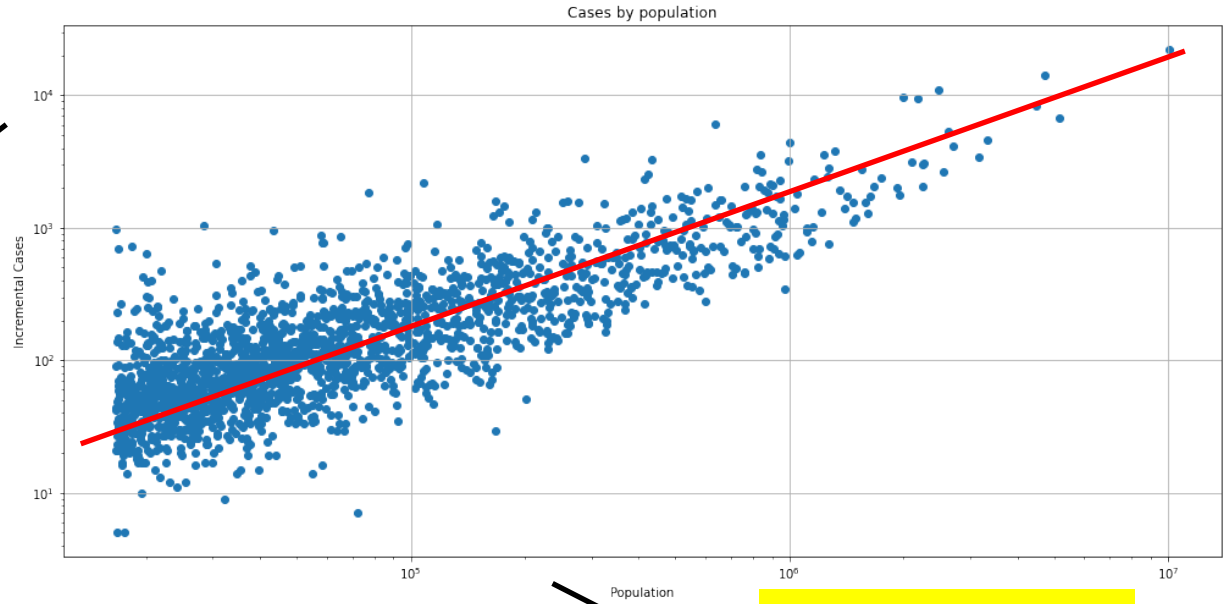
Copyright © Applied AI Institute 2021

Step 5 - Develop Model

Regression

Plot results

Plot in log scale

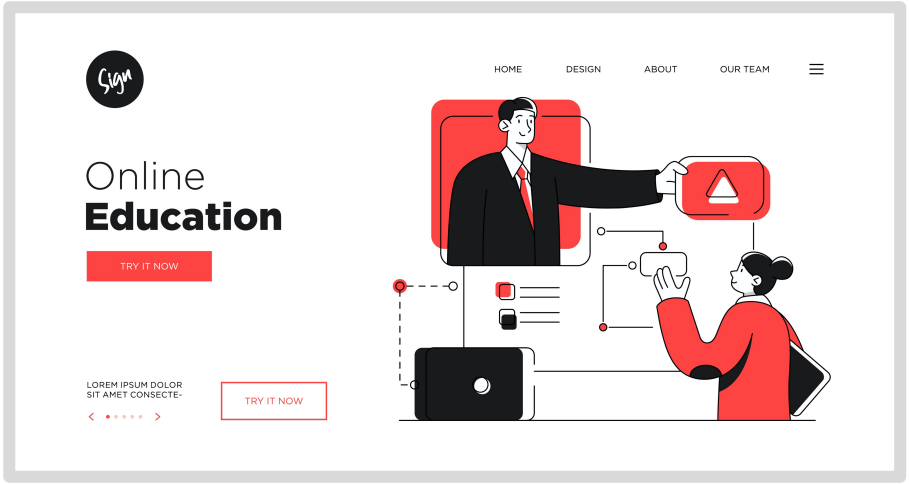


Cases increase with population on log scale

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Data Science Methodology
4. Step 1 - Define Use Case
5. Step 2 - Describe Data
6. Step 3 - Prepare Data
7. Step 4 – Develop Model
8. Step 5 – Evaluate Model



Copyright © Applied AI Institute 2021

Step 5 – Evaluate Model

EM2,4 Test the model and compute confusion matrix, precision and recall

Prediction and confusion matrix:

```
[25] y_predict_all = decision_tree_all.predict(X_test_all)
     accuracy_score(y_test_all, y_predict_all)
```

```
0.6139945062888535
```

Compute prediction

Compute Confusion_matrix

```
pd.DataFrame(
    confusion_matrix(y_test_all, y_predict_all),
    columns=['Predicted Not increase', 'Predicted Increase'],
    index=['True Not Increase', 'True Increase'])
```

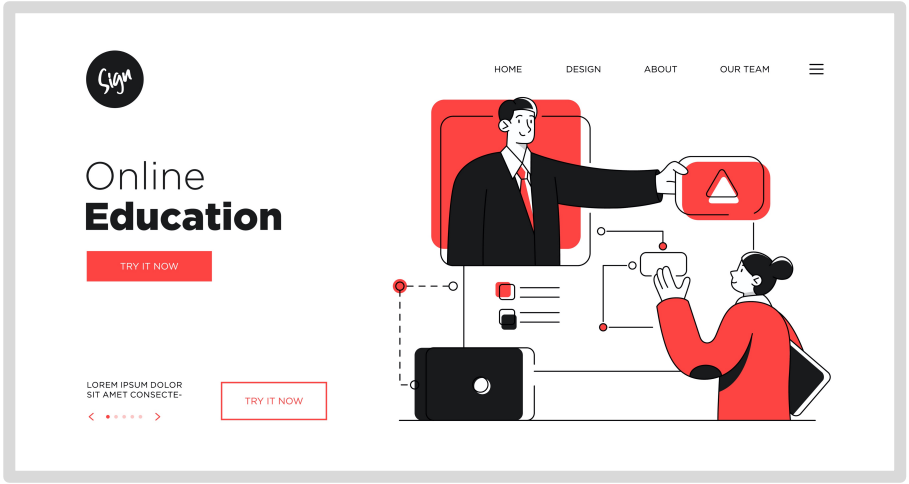
Create confusion matrix

	Predicted Not increase	Predicted Increase
True Not Increase	1688	891
True Increase	1779	2559

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Review Data Science Methodology
4. Step 1 – Describe Use Case
5. Step 2 – Describe Data
6. Step 3 - Prepare Data
7. Step 4 – Develop Model
8. Step 5 – Evaluate Model
9. Step 6 - Deploy Model



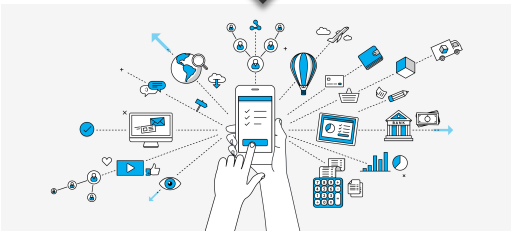
Copyright © Applied AI Institute 2021

Interoperability through Standards

Development platform

Interoperable Models

Deployment platform



Pickle

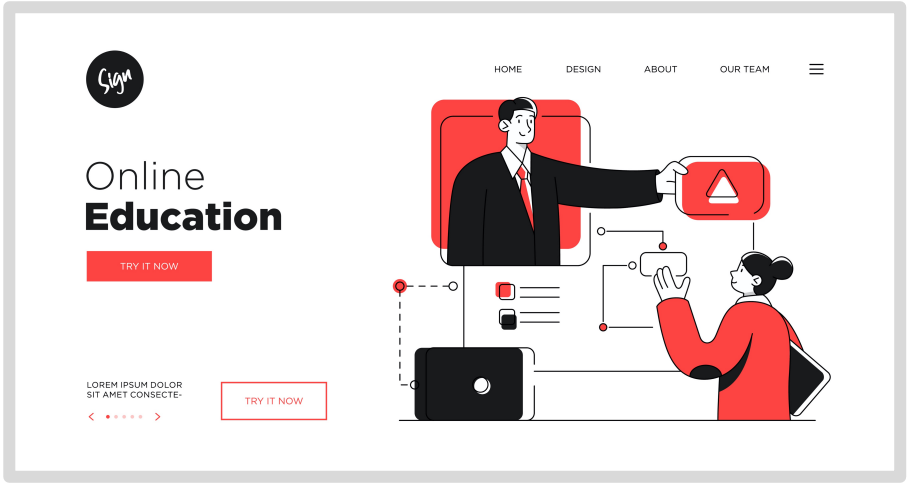
ONNX

PMML

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Review Data Science Methodology
4. Step 1 – Describe Use Case
5. Step 2 – Describe Data
6. Step 3 - Prepare Data
7. Step 4 – Develop Model
8. Step 5 – Evaluate Model
9. Step 6 - Deploy Model
10. Step 7 - Monitor Model



Copyright © Applied AI Institute 2021

Step 7 - Monitor Model

Data Scientist

Expert

1. Model Training

2. Model Deployment

3. Collect Model Results

4. Expert Reviews Results / Feedback

Librarian

3a. Switch Models

New Model >Last Model

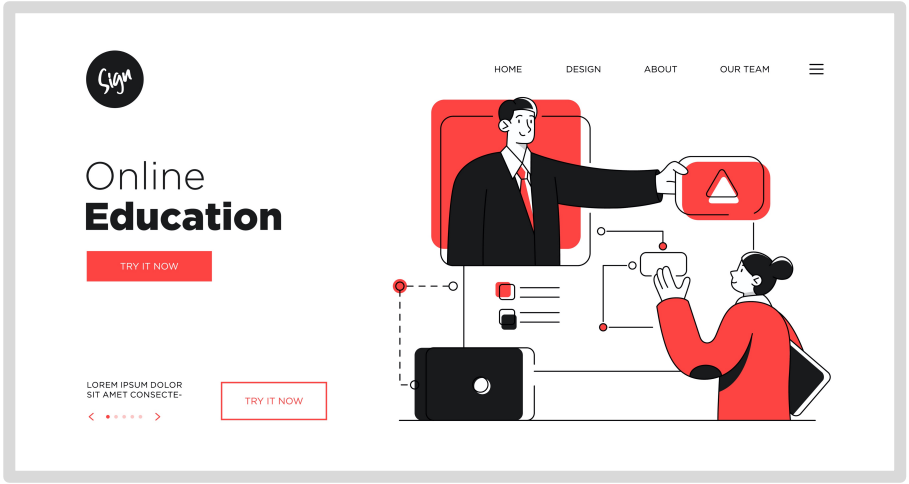
5. Review/ Enhance Training Sets

5a. Archive Training Sets

Copyright © Applied AI Institute 2021

Course Outline

1. Introduction to Course
2. Set up sandbox
3. Review Data Science Methodology
4. Step 1 – Describe Use Case
5. Step 2 – Describe Data
6. Step 3 - Prepare Data
7. Step 4 – Develop Model
8. Step 5 – Evaluate Model
9. Step 6 - Deploy Model
10. Step 7 - Monitor Model
11. Summary and Next Steps



Copyright © Applied AI Institute 2021

Course Deliverables

In this course, you will work on prototyping a data science engagement . In each section, you will develop a component of your COVID use case.

By the time, you end the course, you will have a working prototype of Data Science engagement in Python for COVID Use case containing

i. Refined Data Sets
ii. Python Notebooks for each step of our data science methodology

Copyright © Applied AI Institute 2021

Course Certificate

To successfully complete and receive certification for the course:

1. **Complete** all interactive **quizzes** after each section

2. Download all data sets and sample python code. Complete all assignment sections and submit your final notebook using instructions provided.

Copyright © Applied