AI SUMMIT

CINCINNATI, OH • NOVEMBER 14–16, 2023

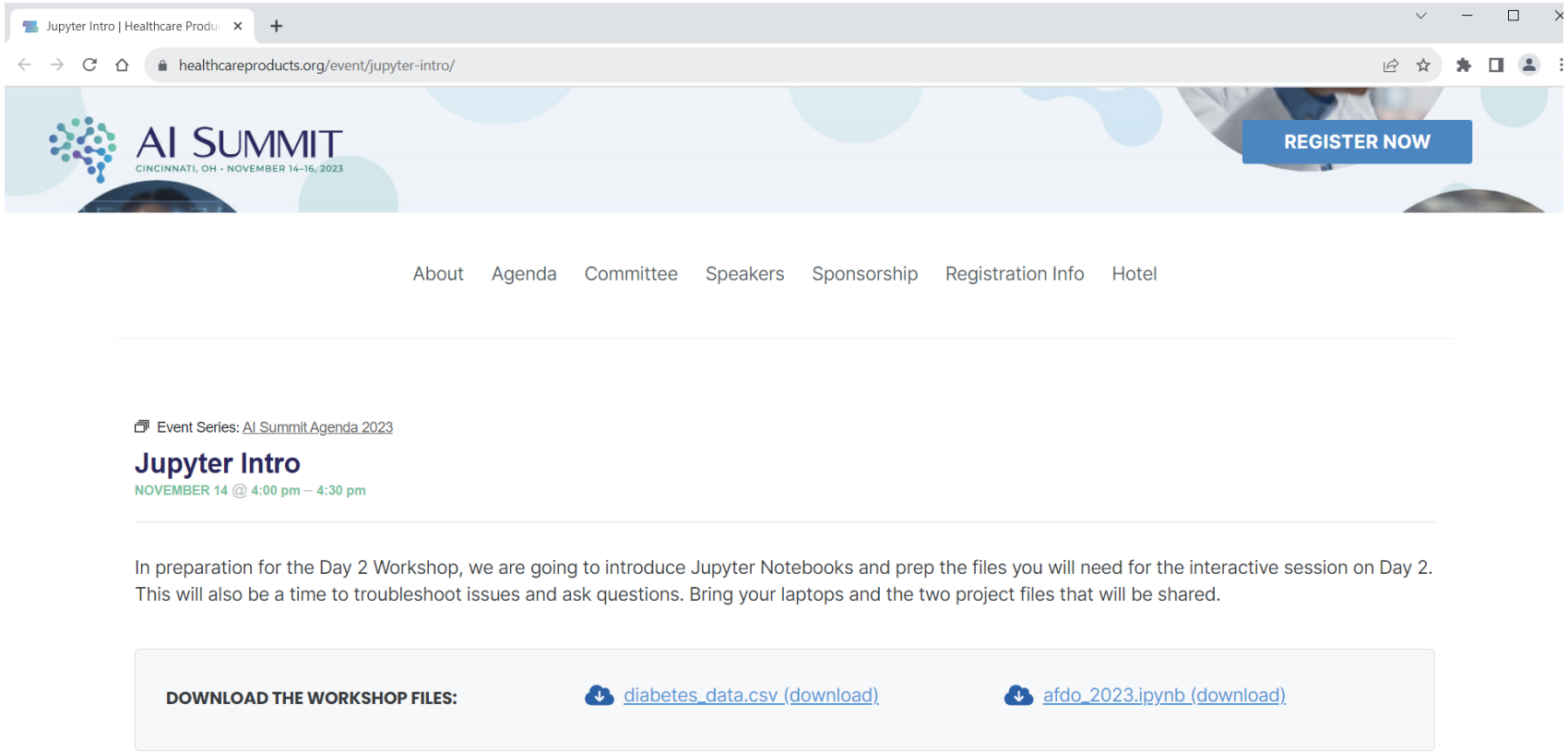
Intro to Jupyter

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

Jupyter

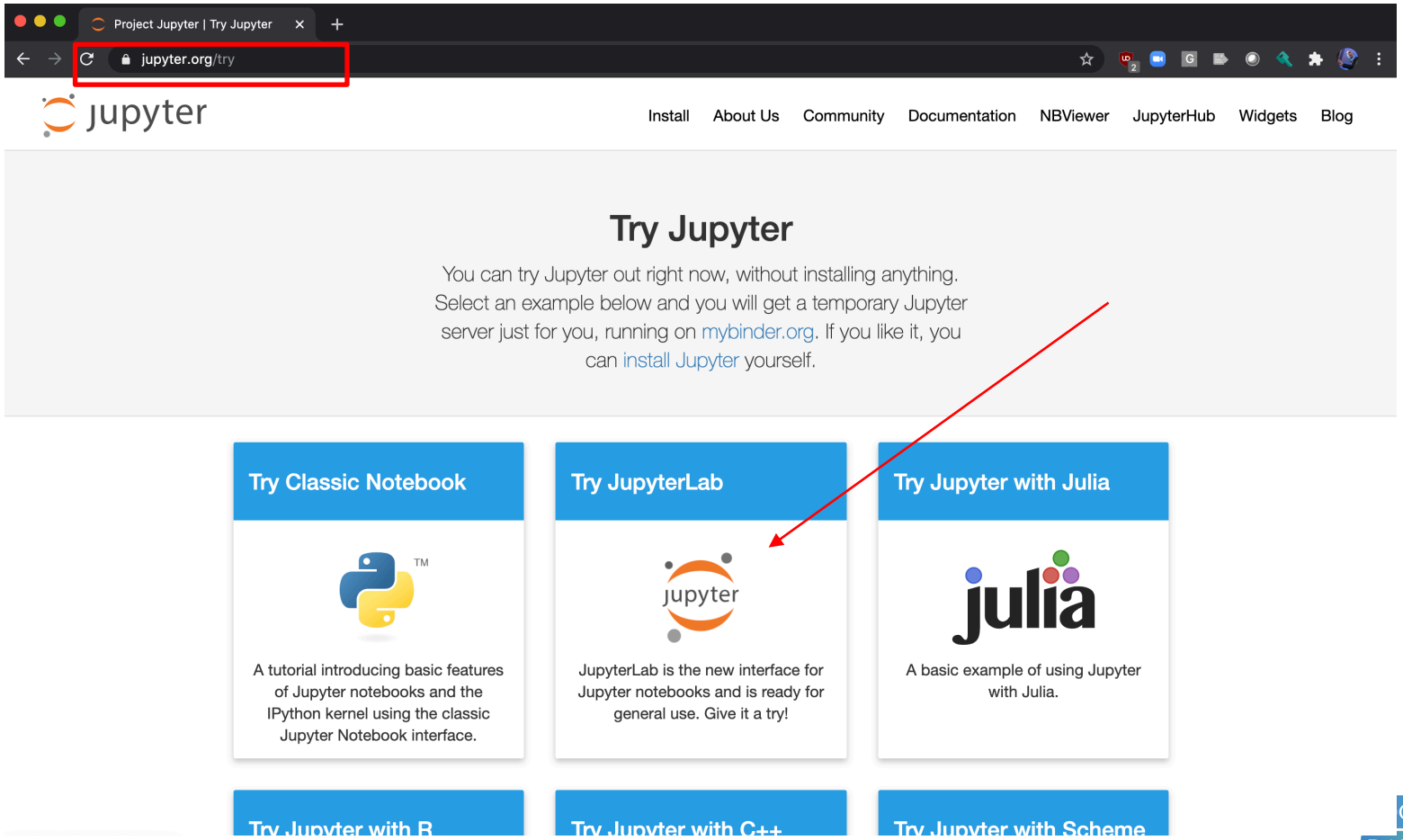
- Web-based development environment for notebooks, code, and data
- Can run Python

Download files from healthcareproducts.org/events/jupyter-intro

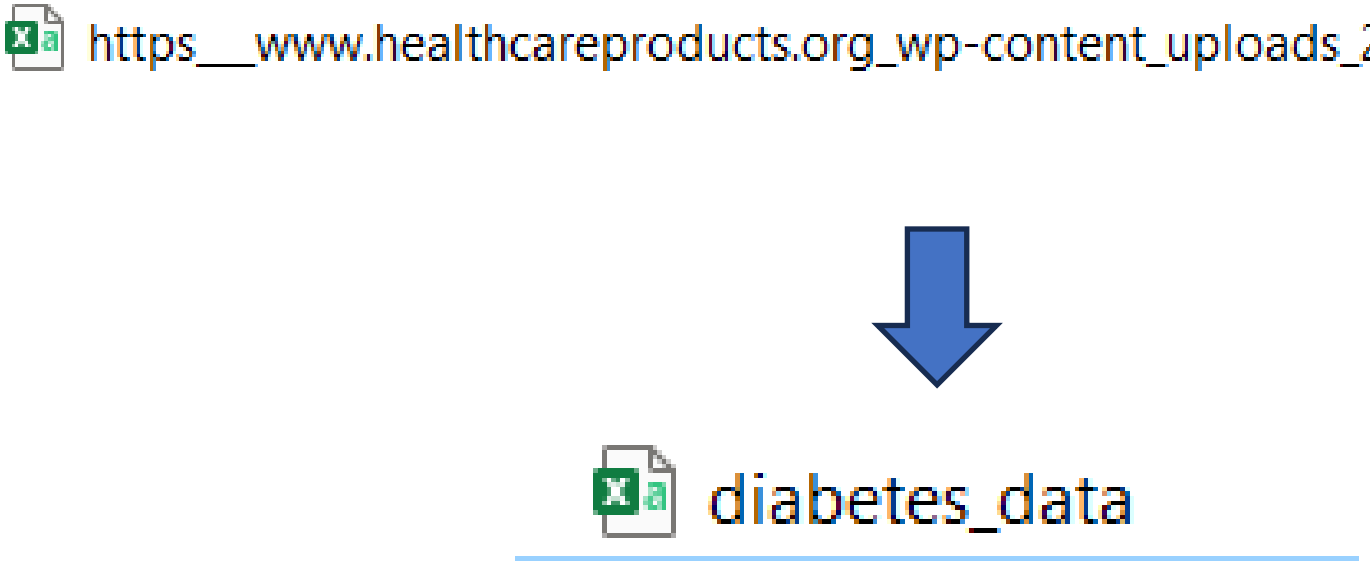


Preparing the environment

Go to jupyter.org and click ‘Try’ JupyterLab

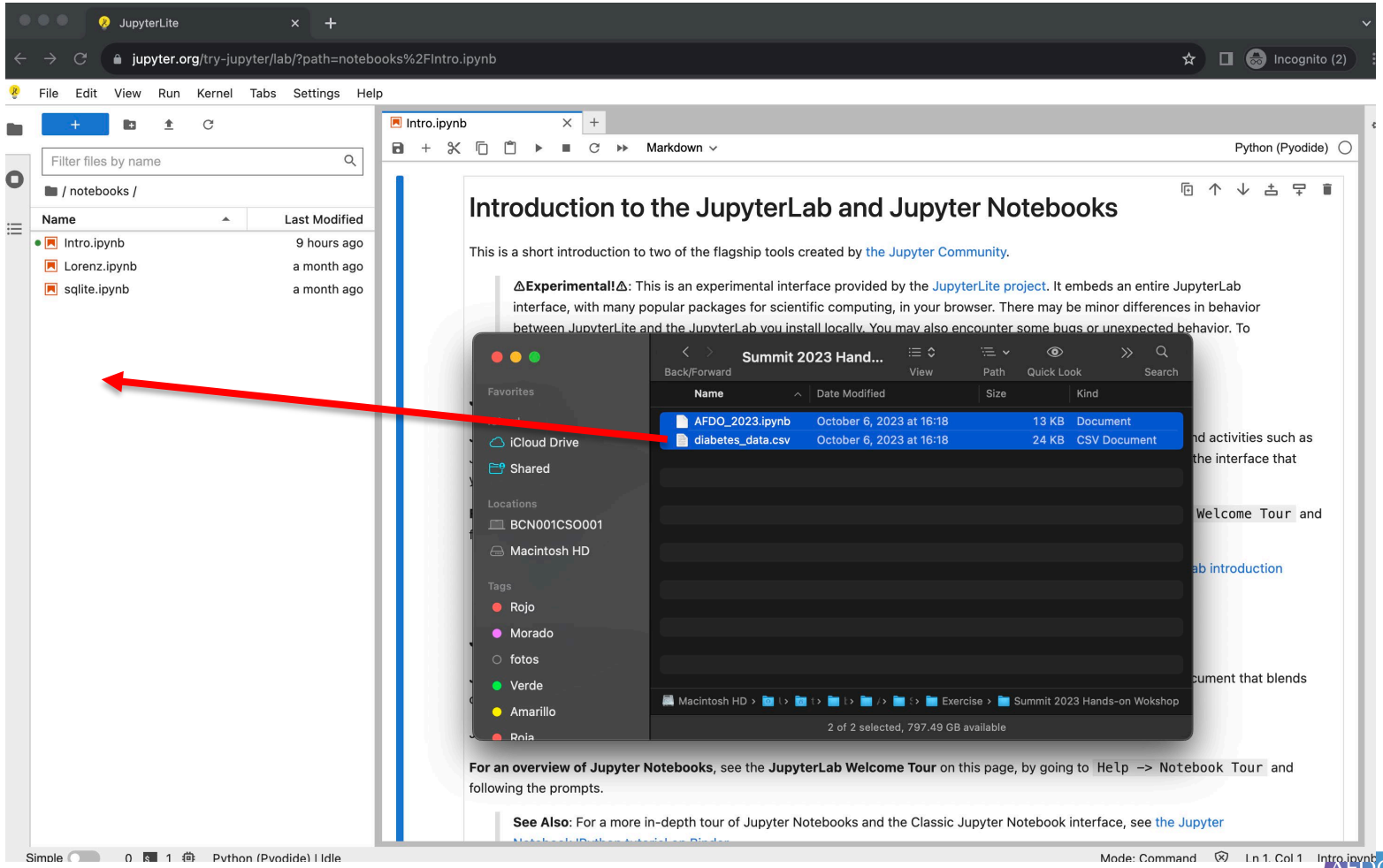


Rename the download



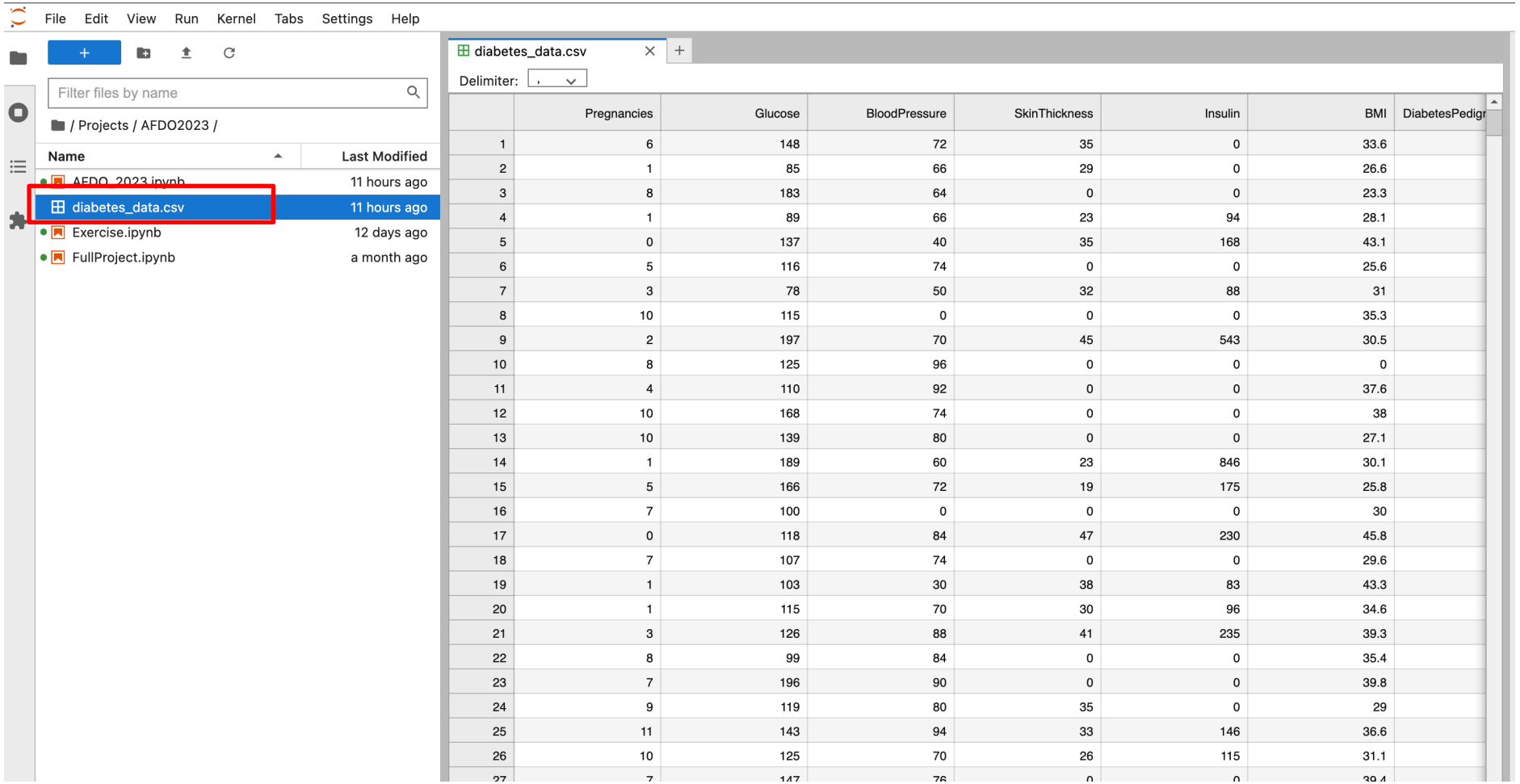
Preparing the environment

Drag the two files to the data folder



Preparing the environment

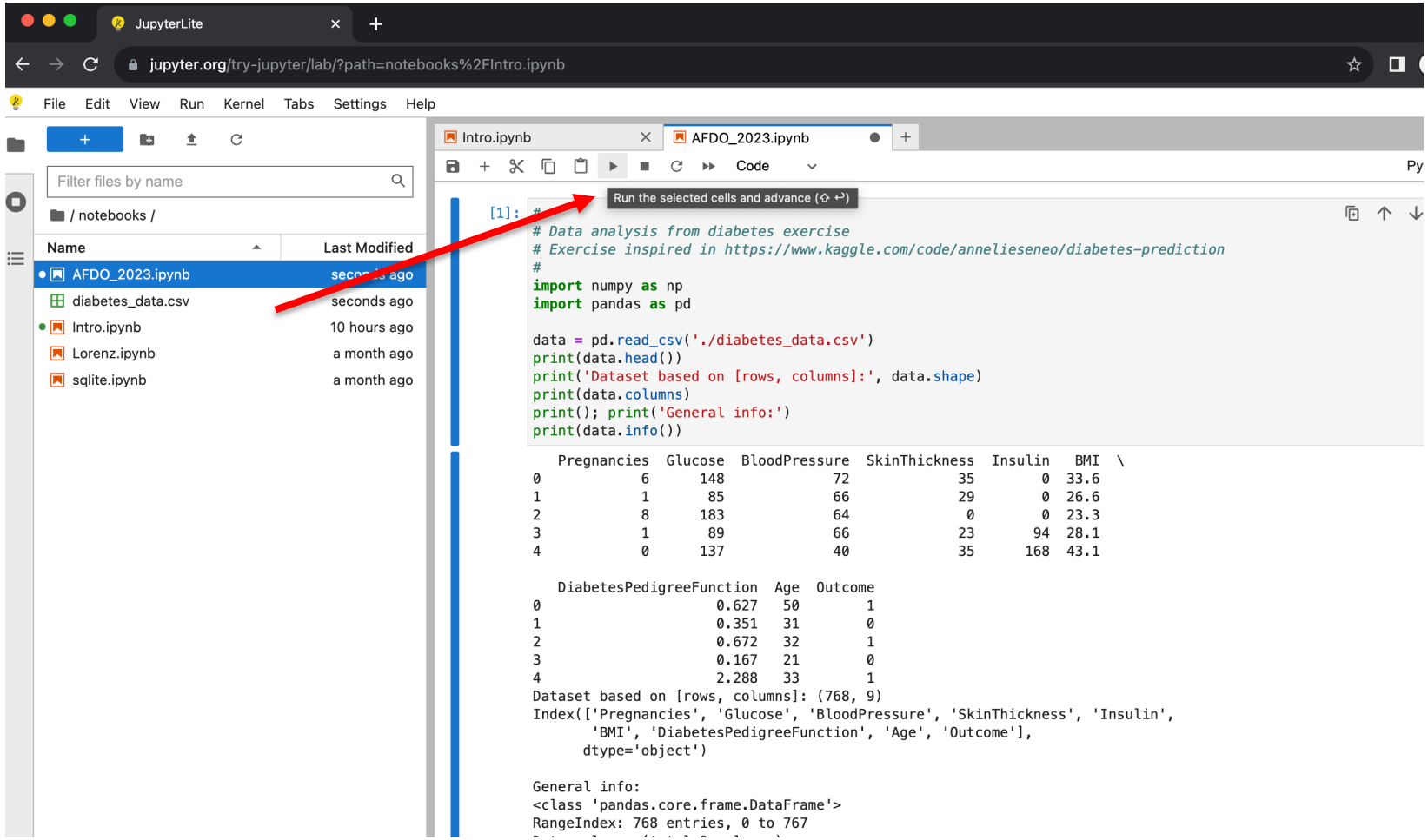
Confirm the data upload



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigr
1	6	148	72	35	0	33.6	
2	1	85	66	29	0	26.6	
3	8	183	64	0	0	23.3	
4	1	89	66	23	94	28.1	
5	0	137	40	35	168	43.1	
6	5	116	74	0	0	25.6	
7	3	78	50	32	88	31	
8	10	115	0	0	0	35.3	
9	2	197	70	45	543	30.5	
10	8	125	96	0	0	0	
11	4	110	92	0	0	37.6	
12	10	168	74	0	0	38	
13	10	139	80	0	0	27.1	
14	1	189	60	23	846	30.1	
15	5	166	72	19	175	25.8	
16	7	100	0	0	0	30	
17	0	118	84	47	230	45.8	
18	7	107	74	0	0	29.6	
19	1	103	30	38	83	43.3	
20	1	115	70	30	96	34.6	
21	3	126	88	41	235	39.3	
22	8	99	84	0	0	35.4	
23	7	196	90	0	0	39.8	
24	9	119	80	35	0	29	
25	11	143	94	33	146	36.6	
26	10	125	70	26	115	31.1	

Preparing the environment

Confirm the code upload



Questions?

We will be here if you need help.

Networking reception will start at 5:00 in the Fountain Room (where lunch was) and end at 6:30.

AI SUMMIT

CINCINNATI, OH • NOVEMBER 14–16, 2023

Workshop

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

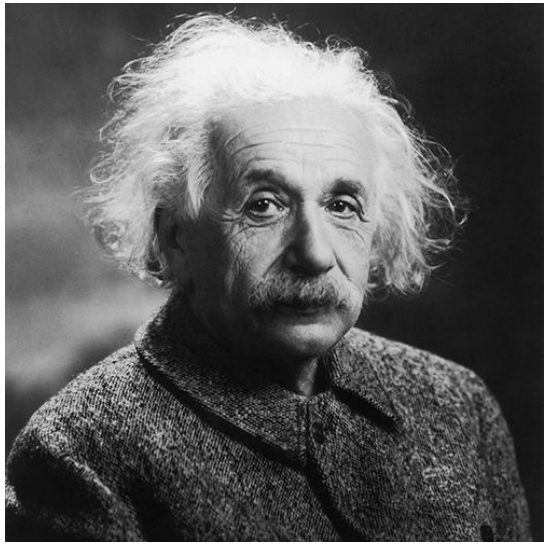
How do I use AI?

- Avoid the downfalls:

1. Executives lack a clear vision for advanced analytics
2. Ignore Lean Principles
3. Goals are driven by tools and not business problems

Defining Problems

"If I were given an hour in which to do a problem upon which my life depended, I would spend 40 minutes studying it, 15 minutes reviewing it, and 5 minutes solving it."



Defining Problems

1. Establish the need for a solution
   - What is the basic need?
   - What is the desired outcome?
2. Justify the need
   - Does this follow the company's strategy?
   - What are the benefits, and how will they be measured?
3. Contextualize the problem
   - What approaches have been tried?
   - Have other organizations tried?
4. Write the problem statement

Good vs Bad

We are behind schedule.

The lack of clean water in developing countries is leading to increased incidence of water-borne diseases and limiting socio-economic development.

We need to use ChatGPT.

Employee turnover of 25% in Company X is negatively impacting productivity, morale, and profitability.

Using AI

- Is there a clear objective?

- Is good quality data available?
- Have traditional statistics failed?
- Do you have a skilled team?
- Are your operations ready for AI?

Business Problem

1. Establish the need for a solution
   - The disease progression for diabetic patients on a certain treatment cannot be determined until 1 year after the treatment started. Knowing the progression sooner will allow earlier adjustments to the treatment.
2. Justify the need
   - Adjusting the treatments earlier will improve the chances of a positive progression improving the lives of our patients. The benefits will be a savings in unnecessary treatment and extended life for patients.
3. Contextualize the problem
   - The dataset was reviewed, but there was no apparent correlation between the measurable attributes and the outcome. Other organizations have used AI models to predict the outcomes.

Business Problem

4. Write the problem statement.
   - The disease progression for diabetic patients on a certain treatment cannot be determined until one year after the treatment started. Knowing the progression sooner will allow earlier adjustments to the treatment, improving the chances of a positive. The benefits will be savings in unnecessary treatment and extended life for patients. The dataset was reviewed, but there was no apparent correlation between the measurable attributes and the outcome. Other organizations have used AI models to predict the outcomes.

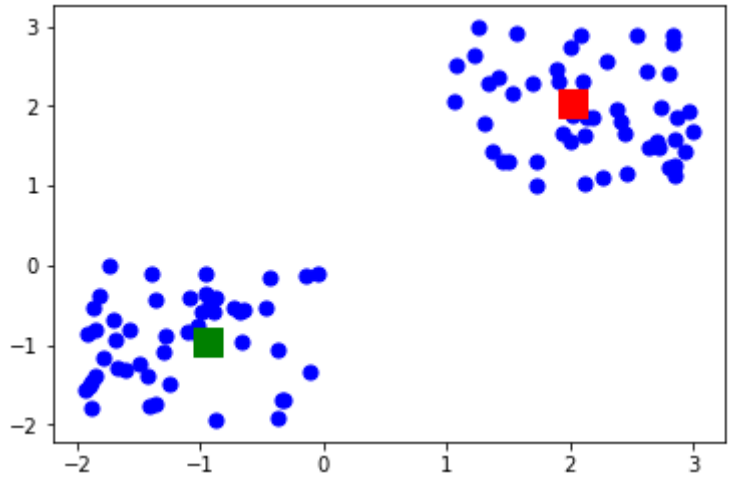
K-Means Clustering

Unsupervised approach used to group similar objects into clusters

Clusters so that the sum of the squared distances between the objects and their cluster mean is minimized.

Input: Columns of features with numerical values

Output: Objects placed into similar clusters

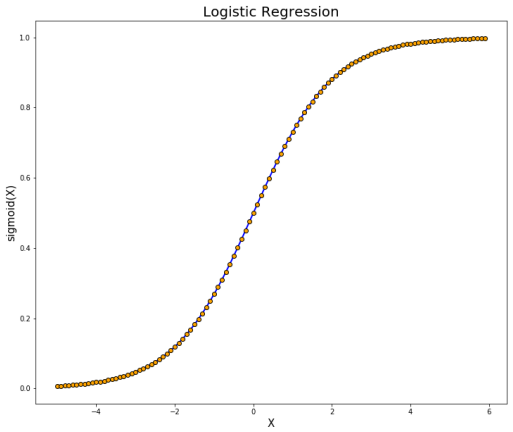


Logistic Regression

Supervised approach used to determine the probability of a discrete outcome (y) give input variables (x)

Input: Columns of numerical inputs, column of categorical output

Output: Predictive model predicting the outcome of future objects

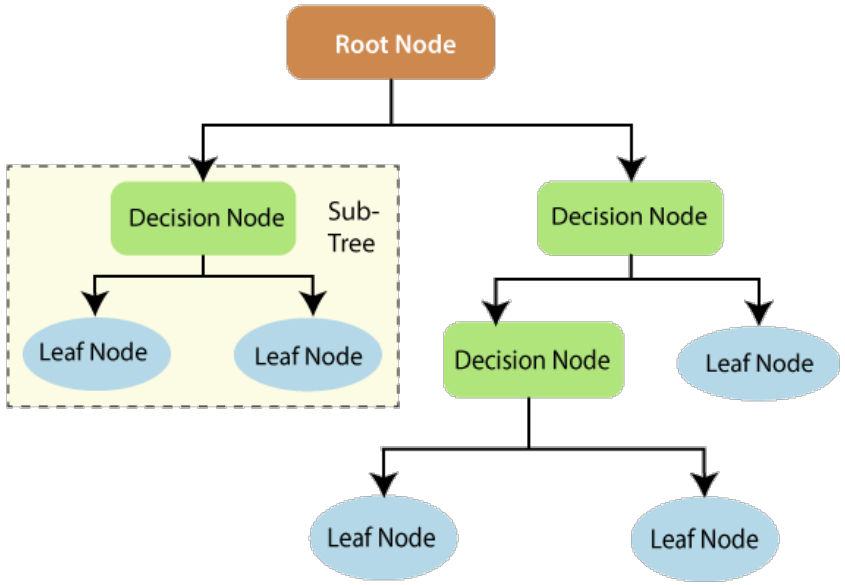


Random Forest

Supervised approach using decision trees to solve regression and classification problems

Input: Columns of numerical inputs, column of numerical output

Output: Predictive model to categorize future objects



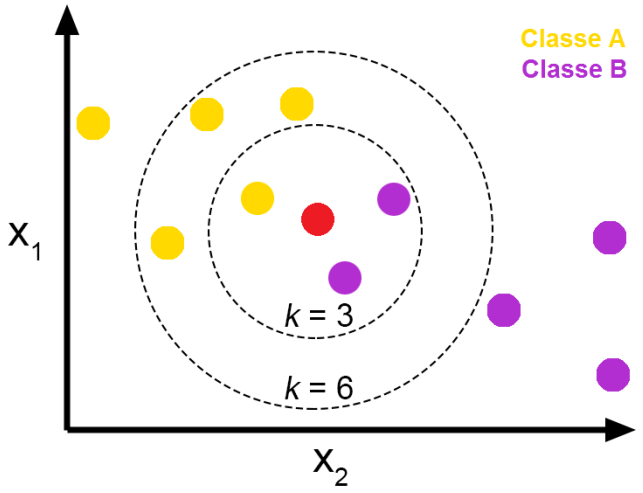
K-Nearest Neighbors Algorithm

Supervised approach used to group similar objects into clusters

Uses distance to locate the closest neighbors of an object

Input: Columns of features with numerical values, column of numerical categorization

Output: Predictive model to categorize future objects



AI SUMMIT
CINCINNATI, OH · NOVEMBER 14–16, 2023

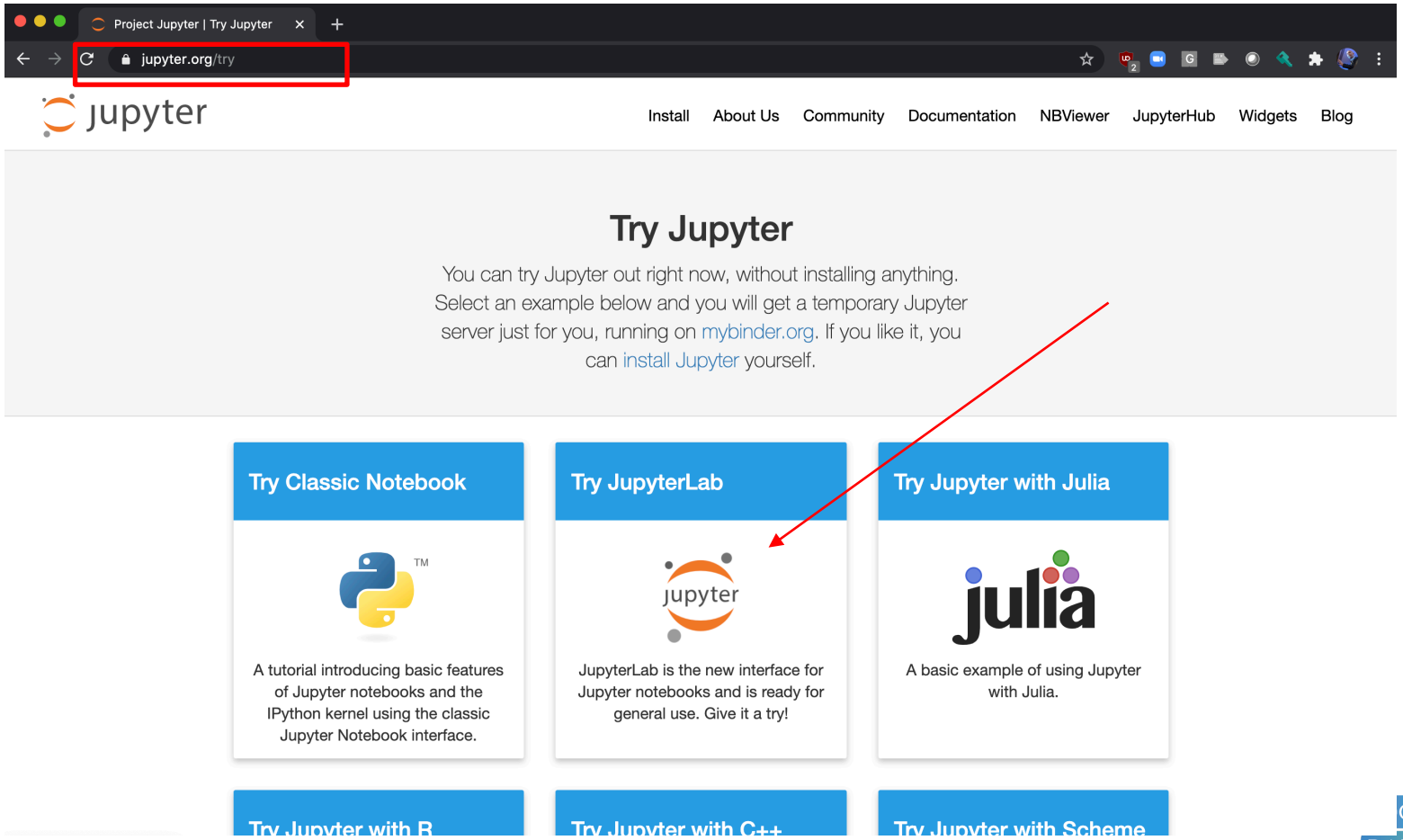
Inspiring Collaboration. Leading Innovation. Making a difference.

Hand—on application

Analyzing a disease

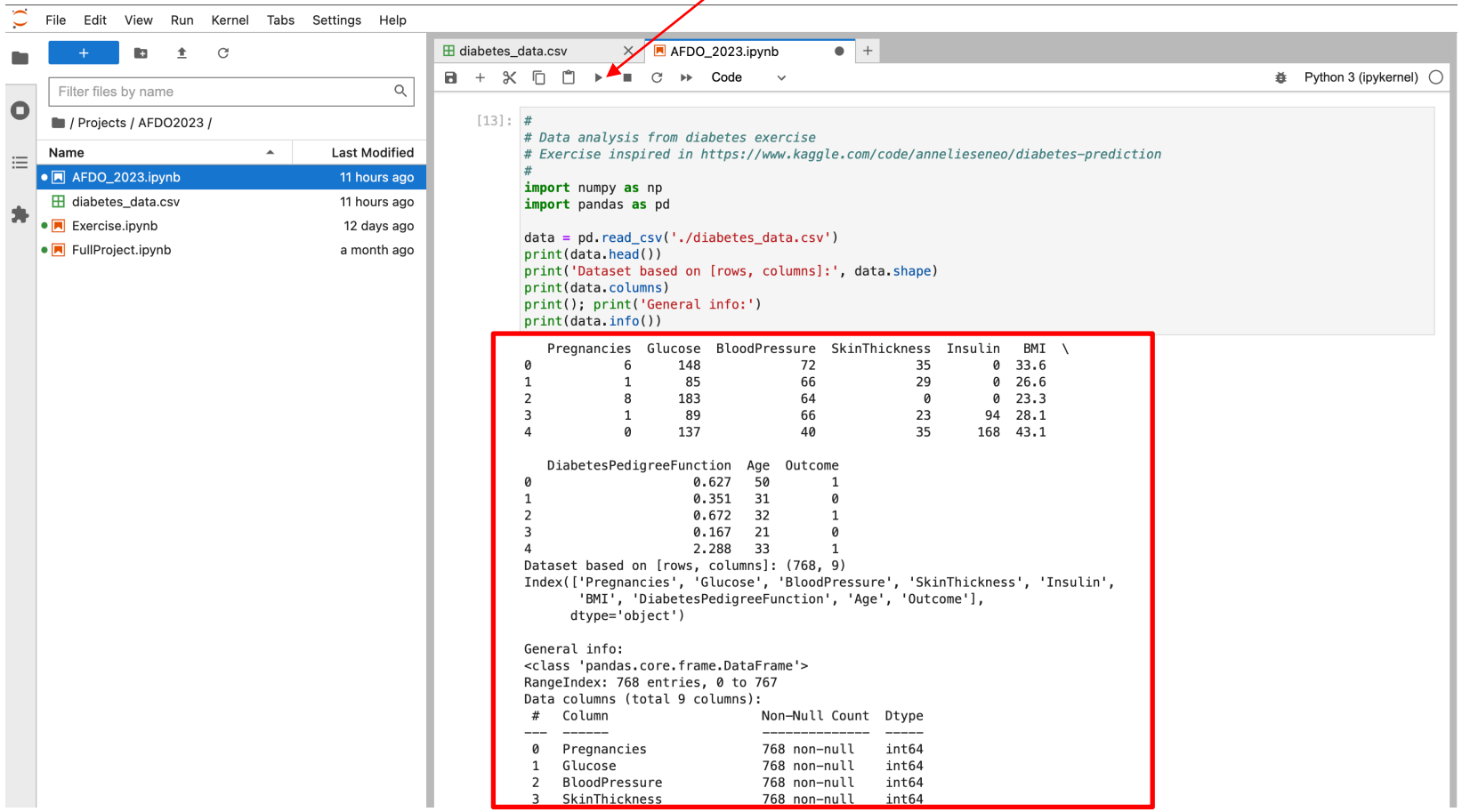
Preparing the environment

Go to jupyter.org and click ‘Try’ JupyterLab



Starting with the data

Execute the code and look at the data



Looking at the data

Making decisions around data

```
#
# Values ••like glucose,bloodpressure or BMI can not be 0. We have to fix the problem.

# looking for empty data, lost data.

print(); print('Null values:')
data.isna().sum()
print(); print('Empty values:')
data.eq(0).sum()
```

```
Null values:

Empty values:
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

Looking at the data

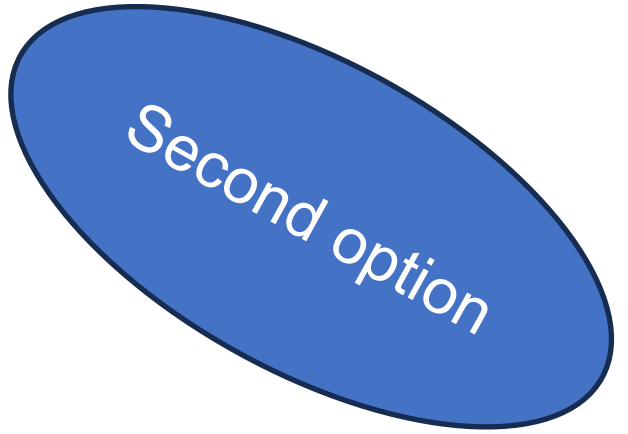
First option

Interpolating values to not lose data for columns with value = 0

```
#Missing Data Imputation Using Regression


def ImputeZeroValuesWithRegression(dataset):

  columnsToBeImputed = ['BloodPressure','Glucose','Insulin','SkinThickness','BMI']
  for column in columnsToBeImputed:

    test_df = dataset[dataset[column]==0]


    y_train= dataset[column]
    x_train= dataset.drop(column,axis=1)

    X_test = test_df.drop(column, axis=1)

    lr=LinearRegression()
    lr.fit(x_train,y_train)
    y_pred=lr.predict(X_test)


    dataset.loc[dataset[column]==0,column] = y_pred

  return dataset

# Interppolating blank values for columns that do not make sense to have empty values
df=ImputeZeroValuesWithRegression(dataset=data)
print(); print('Empty values:')
df.eq(0).sum()
```

```
Empty values:

Pregnancies                 111
Glucose                       0
BloodPressure                 0
SkinThickness                 0
Insulin                       0
BMI                           0
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

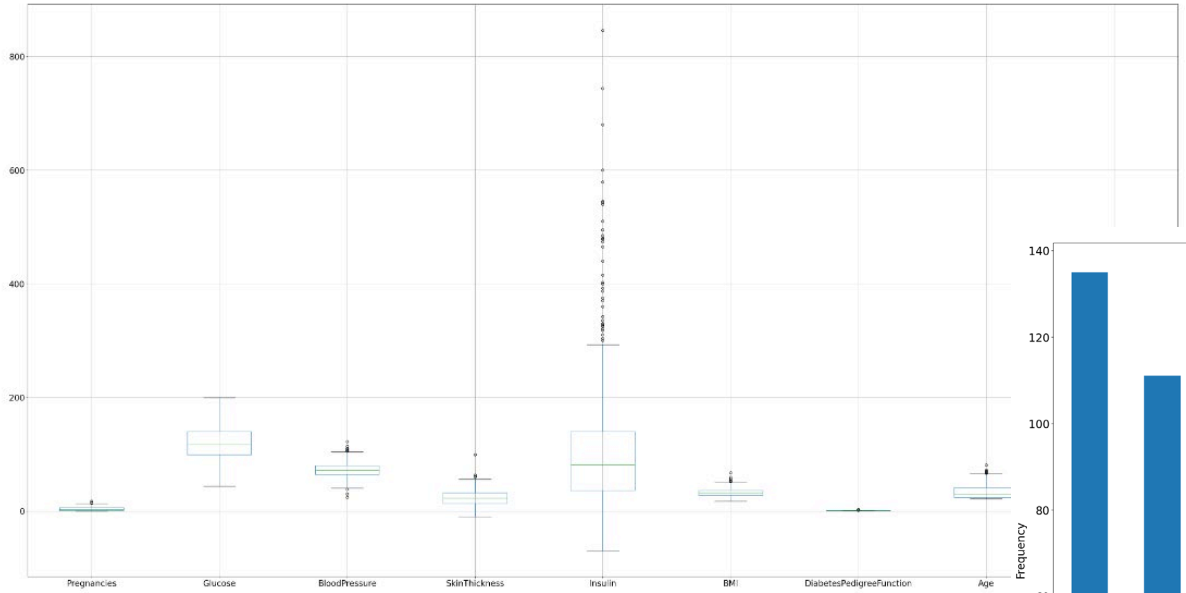
Looking at the data

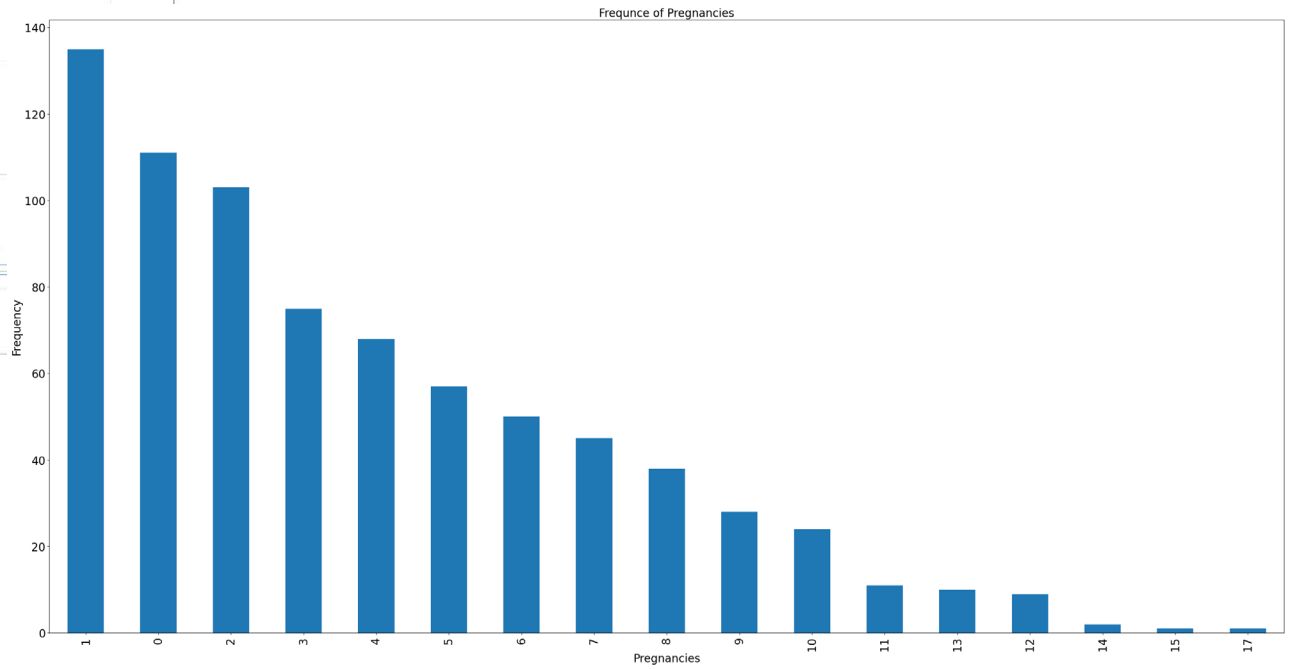


Changing values = 0 with the mean

```
# Changing first the emptiy values of the table.
# NaN instead of 0
data[["Pregnancies","Glucose","BloodPressure","SkinThickness","Insulin","BMI","DiabetesPedigreeFunction","Age"
    ]]=data[[
    "Pregnancies","Glucose","BloodPressure","SkinThickness","Insulin","BMI","DiabetesPedigreeFunction","Age"]].replace(0,np.NaN)

#filling in the missing values
data.fillna(data.mean(),inplace=True)

#I filled the empty spaces with average.
data.head()
```

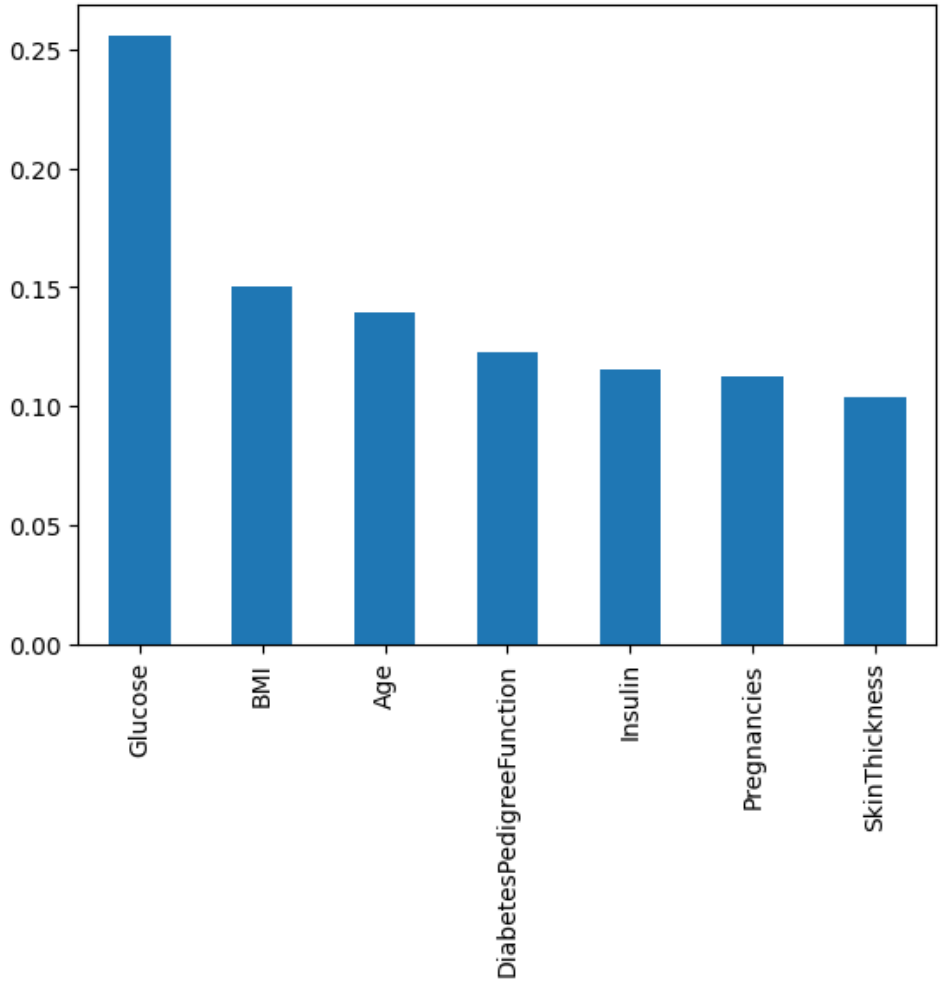
Exploring the Data





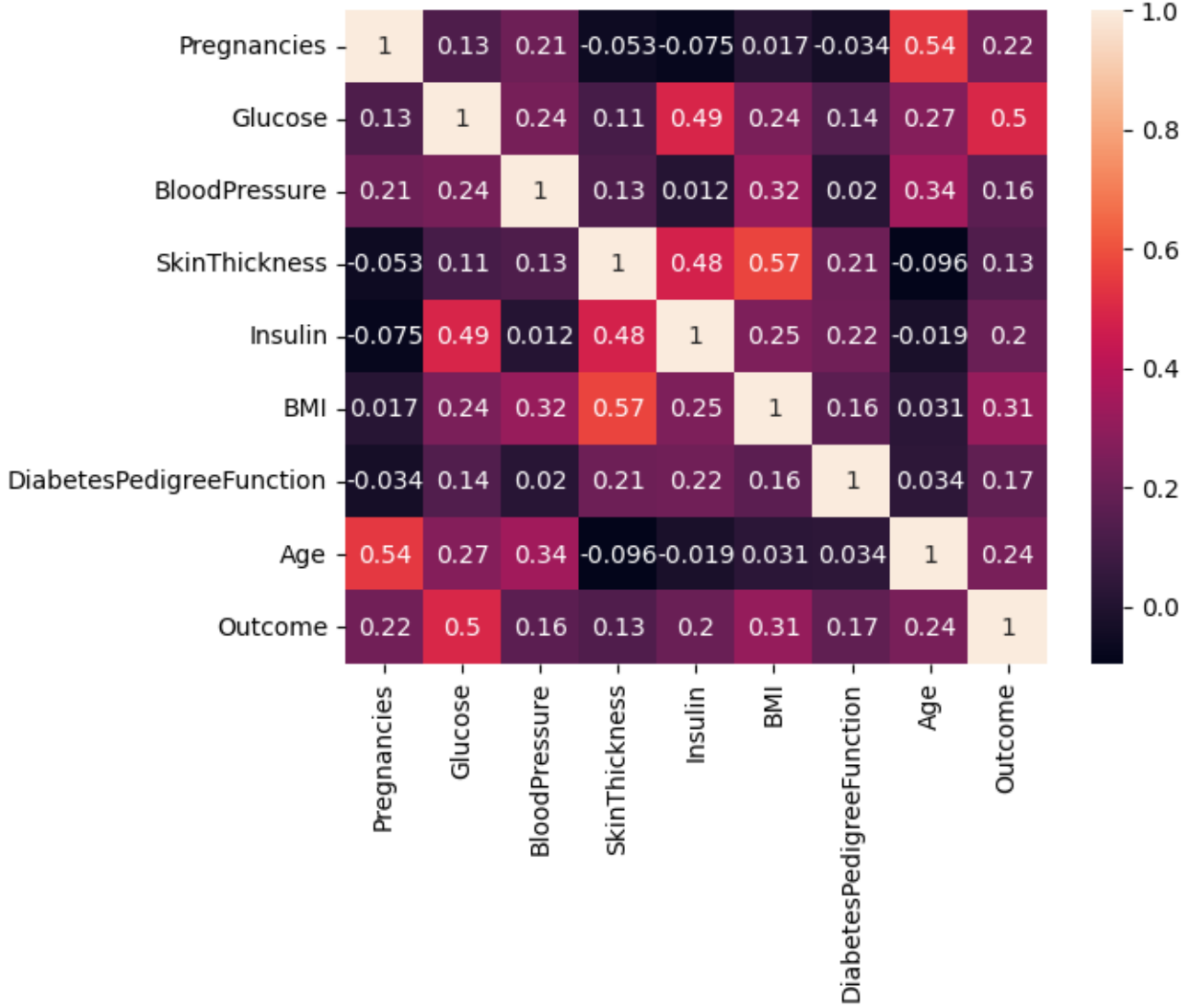
Order of Importance

```
#
# Order of importance
#
x=data[['Glucose', 'BMI', 'Age', 'Pregnancies', 'SkinThickness',
       'Insulin', 'DiabetesPedigreeFunction']]
y=data.iloc[:,8]

model = ExtraTreesClassifier()
model.fit(x,y)
print(model.feature_importances_)
#plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=x.columns)
feat_importances.nlargest(20).plot(kind='bar')
plt.show()
```



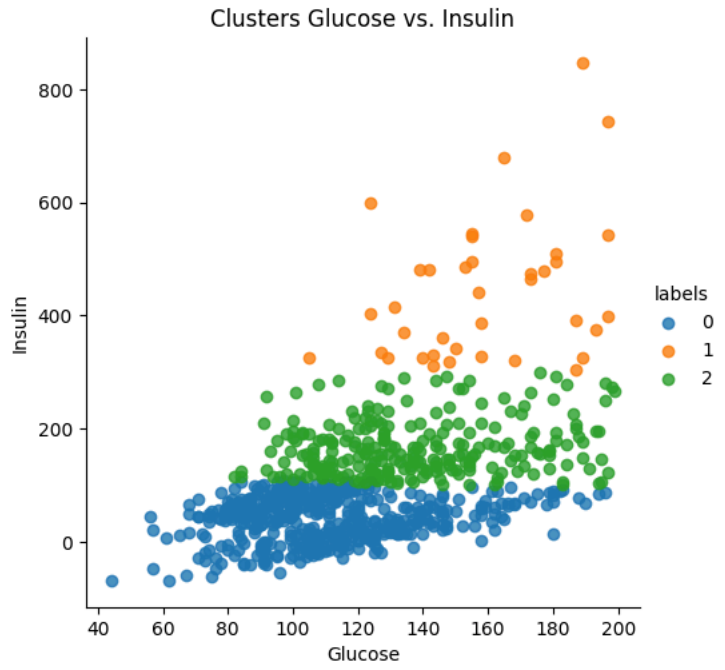
Classic Correlations

```
#
# Classic Correlation
#
#The measure of the relationship between variables.
print(data.corr())
sns.heatmap(data.corr(),annot=True)
```



Creating AI Models

- Clustering data by means k-means
- Classification Task:Logistic Regression
- Classification Task:Random Forest
- Classification Task:KNN Classifier



Solve the problem

Which model would you implement? Why?

Explore the outputs and discuss the options

Lessons Learned

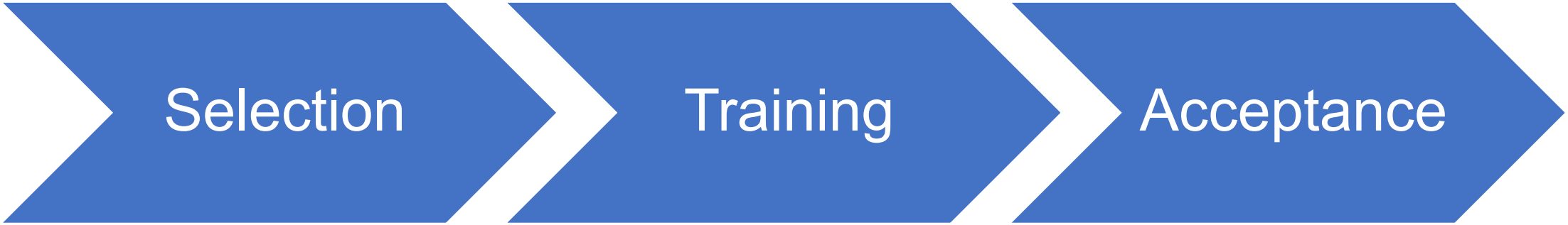
- Did you trust the…
  - Data
  - Algorithms
  - Models
- Next Steps:
  - How to measure the model drifting?
  - What causes drift? Can it be documented in a risk assessment?
  - What are the dangers of model drift?

What if…

- The average BMI of the patients has decreased?
- The starting age of participants all increased?
- The precision of an instrument has increased? Decreased?
- Blanks are being introduced due to a change in the process?
- The disease mutates or becomes less understood?

Validation – Data

Validation – Algorithm



Validation – Model

Q&A

AI SUMMIT
CINCINNATI, OH · NOVEMBER 14–16, 2023

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

Inspiring Collaboration. Leading Innovation. Making a difference