

# CDS6214

# Data Science Fundamentals

Lecture 11

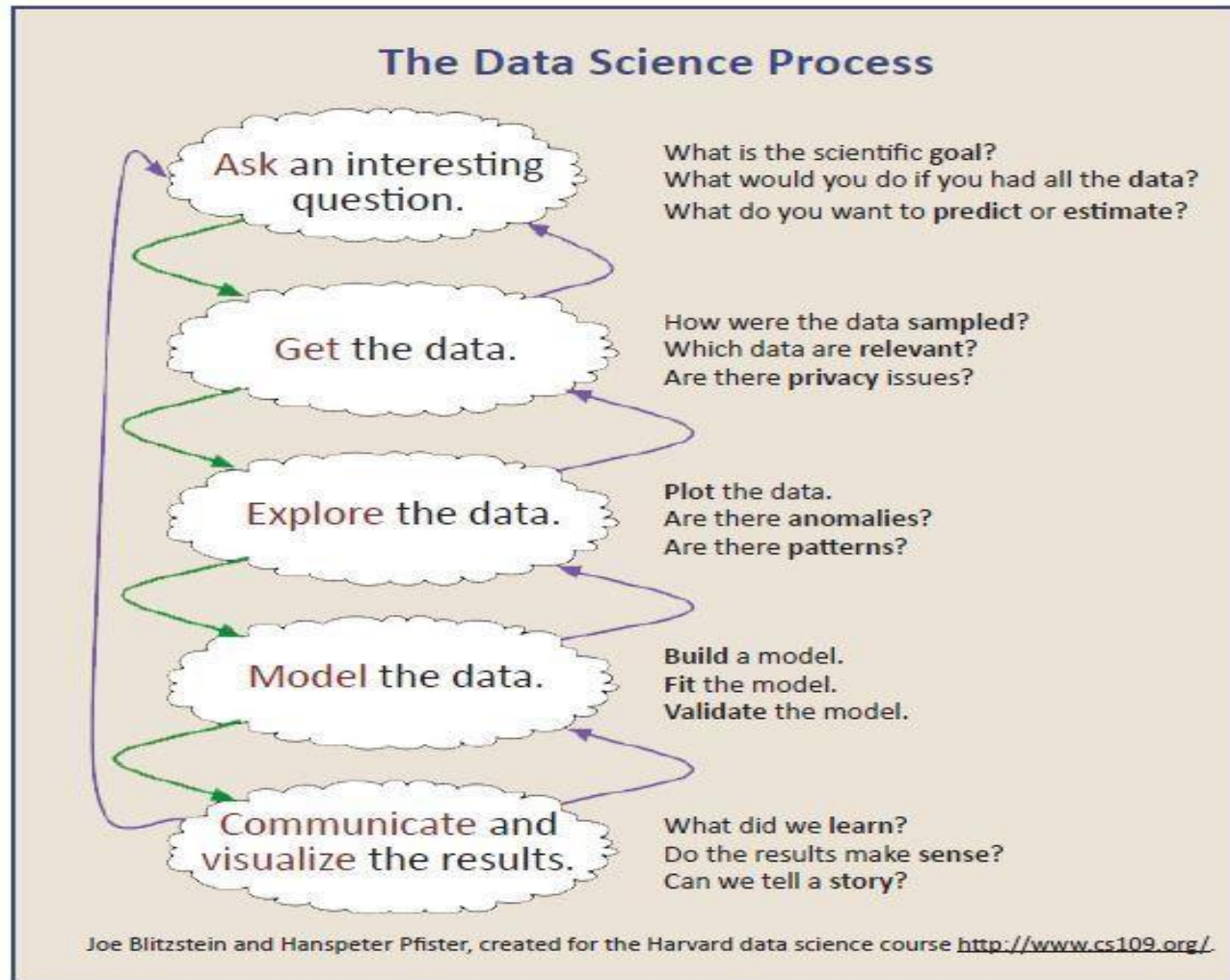
Case Study: Pima Indians Diabetes

# Outline

---

- ❖ **Case Study: Pima Indians Diabetes**
  - ❖ Questions
  - ❖ Dataset
  - ❖ Cleaning the Data
  - ❖ Describing the Data
  - ❖ Finding Associations
  - ❖ Building a Predictor
  - ❖ Reporting Results

# Data Science Process



# Data Science Pipeline: Tasks

---

- ❖ **Describe** the dataset (how many variables, which type of variable, what do the variables mean etc.)
- ❖ **Data Cleaning**: Are there missing values? Was any imputation done? Should we remove any observations or instances? Are there outliers? How do we handle it?
- ❖ **Data Transformation**: Is there a need to normalize the data? Are there any aggregated data within the dataset? Is there a need to aggregate the data?
- ❖ **Exploratory Data Analysis**: Summarize the dataset. What kind of relationships are there between the attributes? Now, are there other variables/instances to be removed?

# Data Science Pipeline: Tasks

---

- ❖ **Data Mining / Data Modelling:** Based on the exploratory analysis earlier, what are the important factors / variables that can be used for data mining and modelling? Are you interested to find patterns without the need for labels, or are you interested to perform prediction on a certain outcome?
  - ❖ If mining is the task, what are suitable algorithms to discover patterns from the data?
  - ❖ If prediction is the task, what are suitable algorithms to train a model to perform the prediction?
- ❖ **Data Visualization:** For each graphical plot you do, describe what you see and why did you do that plot.

---

# **Pima Indians Diabetes Dataset**

# Example of Questions

---

- ❖ **Descriptive Questions:**

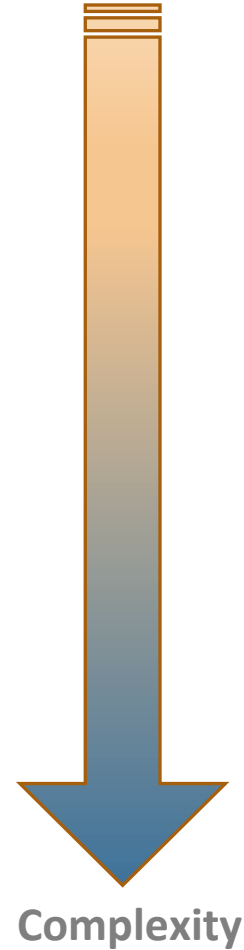
- ❖ How many instances are there in the dataset ?
- ❖ What is the age distribution of the sample population?

- ❖ **Exploratory Question:**

- ❖ For the group of people that is at risk of contracting diabetes and another group that is not at risk of contracting diabetes, which attributes are closely associated to each other (appearing together often)?

- ❖ **Predictive Question:**

- ❖ From the given the attributes that are used to describe the health of a person and the list of observations, is it possible to predict people who are at risk of contracting diabetes in the next 5 years?



# Dataset

---

- ❖ The collected samples are from the US National Institute of Diabetes and Digestive and Kidney Disease. It includes 200 women of Pima Indian heritage living near Phoenix, Arizona. This is database of patient records that records whether a patient develops diabetes in five years. It is assumed that the data represents a random and unbiased sample from the population defined.
- ❖ The attributes in the dataset:
  - Number of times pregnant
  - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Diastolic blood pressure (mm Hg)
  - Triceps skin fold thickness (mm)
  - 2-Hour serum insulin ( $\mu$ U/ml)
  - Body mass index (weight in kg/(height in m)<sup>2</sup>)
  - Diabetes pedigree function
  - Age (years)
  - Class variable (0 or 1)  
(class value 1 is interpreted as "tested positive for diabetes")



# Describing the Dataset

---

Variable	Continuous /Discrete	Anticipated role	Comments
Pregnant	Continuous	Descriptor	Number of times pregnant
Plasma-Glucose	Continuous	Descriptor	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
DiastolicBP	Continuous	Descriptor	Units: mm Hg
TricepsSFT	Continuous	Descriptor	Units: mm
2-Hour serum insulin	Continuous	Descriptor	Units: mm U/ml
BMI	Continuous	Descriptor	Body mass index
DPF	Continuous	Descriptor	Diabetes pedigree function
Age	Continuous	Descriptor	Units: years
class	Discrete	Response	0 – does not contract diabetes in five years 1 – contracts diabetes in five years

# Example of Questions

## ❖ Descriptive Questions:

- ❖ How many instances are there in the dataset ?

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- Pregnancies = Number of times pregnant
- Glucose = Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure = Diastolic blood pressure (mm Hg)
- SkinThickness = Triceps skin fold thickness (mm)
- Insulin = 2-Hour serum insulin (mu U/ml)
- BMI = Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction = Diabetes pedigree function
- Age = Age (years)
- Outcome = Class variable (0 or 1) 268 of 768 are 1, the others are 0

In [4]: `df.shape`

Out[4]: (768, 9)

There are 768 instances in the dataset.

Complexity

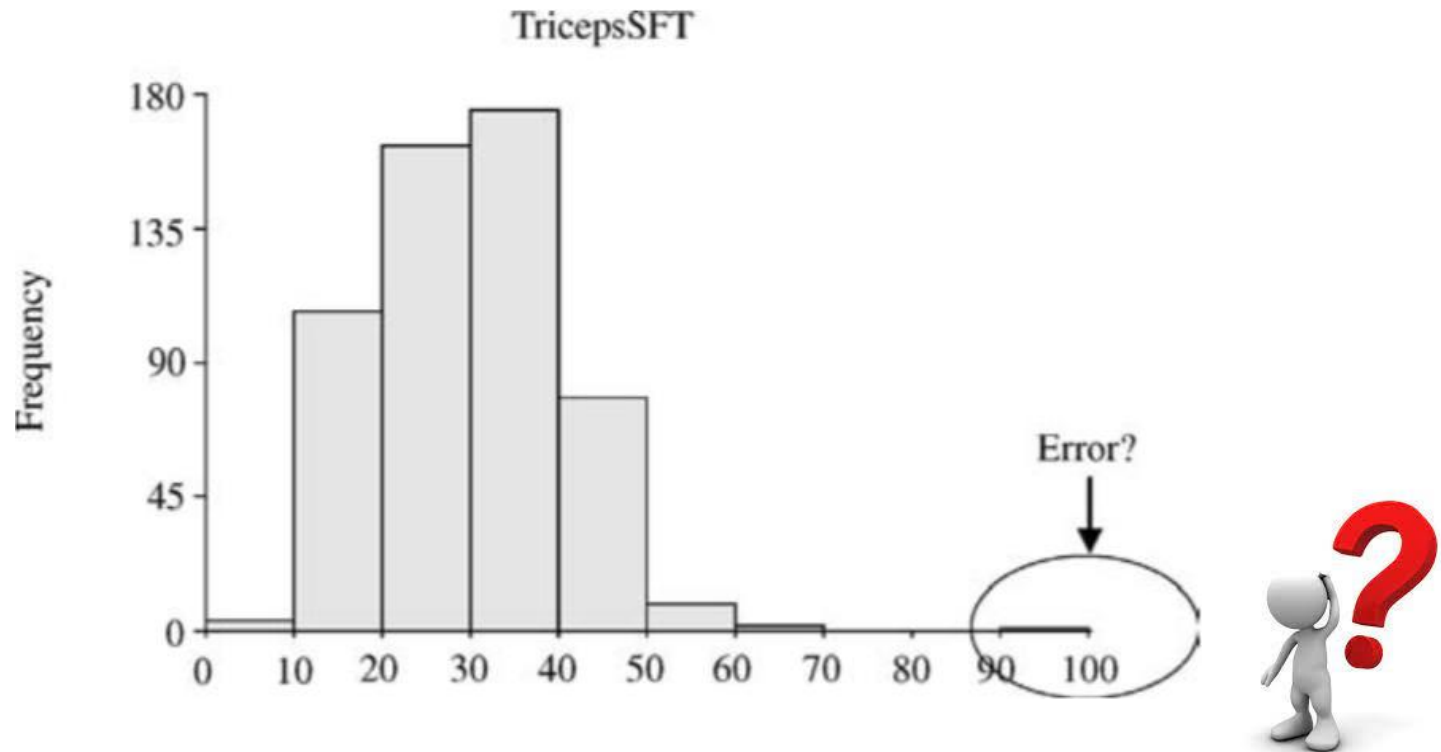
# Handling Missing Data

- ❖ A preliminary analysis of the data shows the use of zero for missing data.
- ❖ Any assumptions should be validated with those involved in collection.

Variable	Number of Zeros
Pregnant	111
Plasma-Glucose	5
DiastolicBP	35
TricepsSFT	227
2-Hour serum insulin	374
BMI	11
DPF	0
Age	0

- ❖ Observations with zero values are removed. Removing all observations with missing data would significantly decrease the size of the dataset to analyze. (This can be good or bad depending on the size of the dataset).

# Examining Outliers



# Data Transformation

---

- ❖ Determine whether any kind of transformation is required.
- ❖ The following transformations are considered within this analysis: normalization, discretization, and aggregation.

# Data Transformation

---

- ❖ One of the requirements of this analysis  $\Rightarrow$  classify general associations between classes of variables, such as high blood pressure, and diabetes.
- ❖ Bin each variable into a small number of categories. This process should be performed **in consultation with both subject matter experts and/or healthcare professionals who will use the results.**
- ❖ Ensure that any subject matter or practical considerations are taken into account prior to the analysis, since the results will be presented in terms of these categories.

# Data Transformation

---

- ❖ The following summarizes the cut-off values (shown in parentheses) along with the names of the bins for the variables:
  - ❖ **Pregnant:** low (1,2), medium (3,4,5), high (> 6)
  - ❖ **Plasma-Glucose:** low (< 90), medium (90–150), high (> 150)
  - ❖ **DiastolicBP:** normal (< 80), normal-to-high (80–90), high (> 90)
  - ❖ **BMI:** low (< 25), normal (25–30), obese (30–35), severely obese (> 35) DPF: low (< 0.4), medium (0.4–0.8), high (> 0.8)
  - ❖ **Age:** 20–39, 40–59, 60 plus
  - ❖ **Class:** yes (1), no (0)

# Normalization

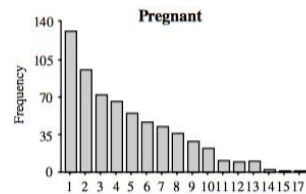
- ❖ To ensure that all variables are considered with equal weight in any further analysis, **min-max normalization** is performed

$$Value' = \frac{Value - OriginalMin}{OriginalMax - OriginalMin} (NewMax - NewMin) + NewMin$$

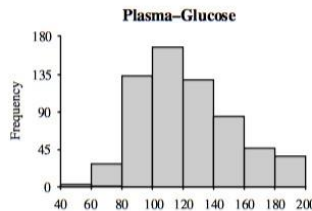
Pregnant	Pregnant (normalized)	Plasma— Glucose	Plasma— Glucose (normalized)	DiastolicBP	DiastolicBP (normalized)	BMI	BMI (normalized)	DPF	DPF (normalized)	Age	Age (normalized)
6	0.35	148	0.67	72	0.49	33.6	0.31	0.627	0.23	50	0.48
1	0.059	85	0.26	66	0.43	26.6	0.17	0.351	0.12	31	0.17
8	0.47	183	0.90	64	0.41	23.3	0.10	0.672	0.25	32	0.18
1	0.059	89	0.29	66	0.43	28.1	0.20	0.167	0.037	21	0
5	0.29	116	0.46	74	0.51	25.6	0.15	0.201	0.052	30	0.15
3	0.18	78	0.22	50	0.27	31	0.26	0.248	0.072	26	0.083
2	0.12	197	0.99	70	0.47	30.5	0.25	0.158	0.033	53	0.53
4	0.24	110	0.43	92	0.69	37.6	0.40	0.191	0.047	30	0.15
10	0.59	168	0.8	74	0.51	38	0.40	0.537	0.20	34	0.22
10	0.59	139	0.61	80	0.57	27.1	0.18	1.441	0.58	57	0.6
1	0.059	189	0.94	60	0.37	30.1	0.24	0.398	0.14	59	0.63



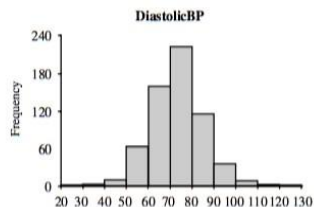
# Descriptive Statistics



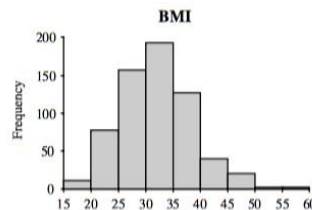
**Central tendency**  
 Mode: 1.0  
 Median: 4  
 Mean: 4.48  
**Variation**  
 Variance: 10.36  
 Standard deviation: 3.22  
**Shape**  
 Skewness: 0.89  
 Kurtosis: 0.14



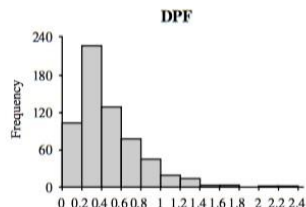
**Central tendency**  
 Mode: 99.0  
 Median: 116.5  
 Mean: 121.53  
**Variation**  
 Variance: 947.13  
 Standard deviation: 30.78  
**Shape**  
 Skewness: 0.52  
 Kurtosis: -0.33



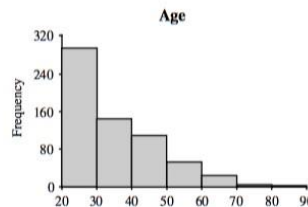
**Central tendency**  
 Mode: 70,74  
 Median: 72  
 Mean: 72.17  
**Variation**  
 Variance: 152.91  
 Standard deviation: 12.37  
**Shape**  
 Skewness: 0.13  
 Kurtosis: 1.03



**Central tendency**  
 Mode: 31.2,32.0,31.6  
 Median: 32  
 Mean: 32.01  
**Variation**  
 Variance: 41.35  
 Standard deviation: 6.43  
**Shape**  
 Skewness: 0.43  
 Kurtosis: 0.23



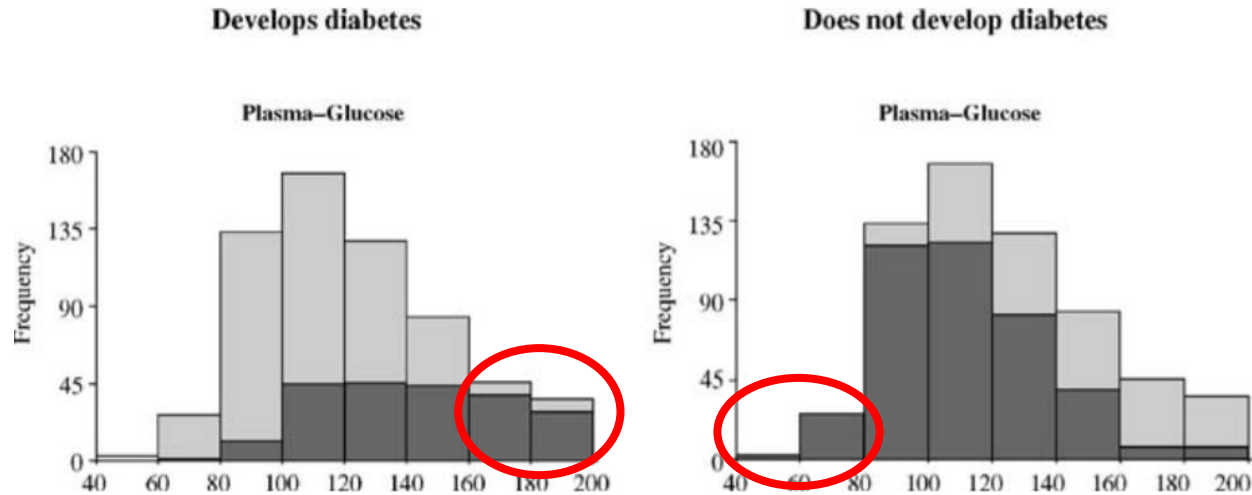
**Central tendency**  
 Mode: 0.258,0.268  
 Median: 0.38  
 Mean: 0.47  
**Variation**  
 Variance: 0.1  
 Standard deviation: 0.32  
**Shape**  
 Skewness: 1.62  
 Kurtosis: 3.74



**Central tendency**  
 Mode: 22.0  
 Median: 30.5  
 Mean: 34.2  
**Variation**  
 Variance: 137.64  
 Standard deviation: 11.73  
**Shape**  
 Skewness: 0.98  
 Kurtosis: 0.3

For each variable, a **frequency distribution** is generated and presented alongside a series of **descriptive statistics** in order to characterize the variables.

# Distribution



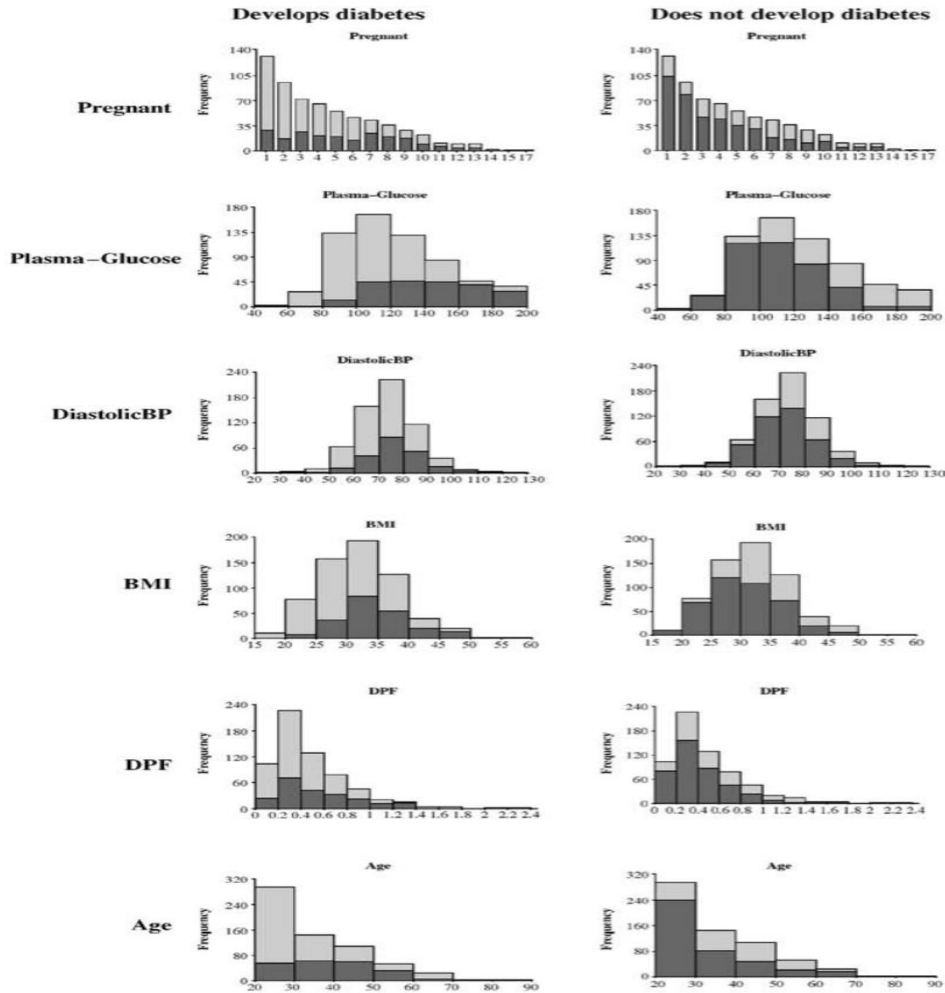
**Distribution of the Plasma-Glucose variable** (light gray). Observations belonging to the two groups are highlighted in dark gray.

- Left histogram: Belongs to patients that went on to develop diabetes.
- Right histogram: Belongs to patients that did not develop diabetes.

Significantly different distribution of Plasma-Glucose data between the groups.

- Almost all patients with the highest Plasma-Glucose values went on to develop diabetes.
- Almost all the patients with the lowest Plasma-Glucose values did not go on to develop diabetes within five years.

# Distribution



What else can you observe?

# Summary Statistics

Diabetes	Patient count	Mean (Pregnant)	Mean (Plasma–Glucose)	Mean (DiastolicBP)	Mean (BMI)	Mean (DPF)	Mean (Age)
no	408	3.87	110.7	70.51	30.58	0.43	31.93
yes	216	5.65	142	75.3	34.72	0.54	38.5

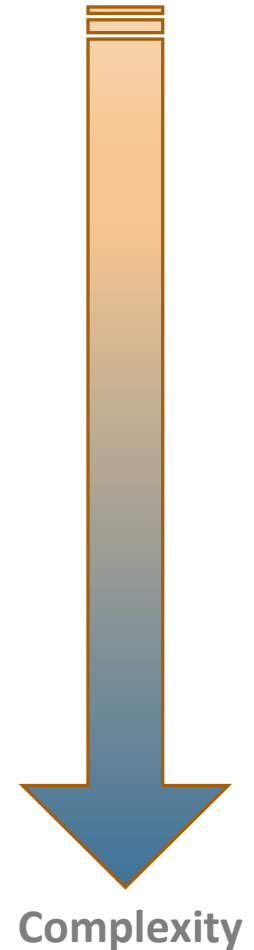
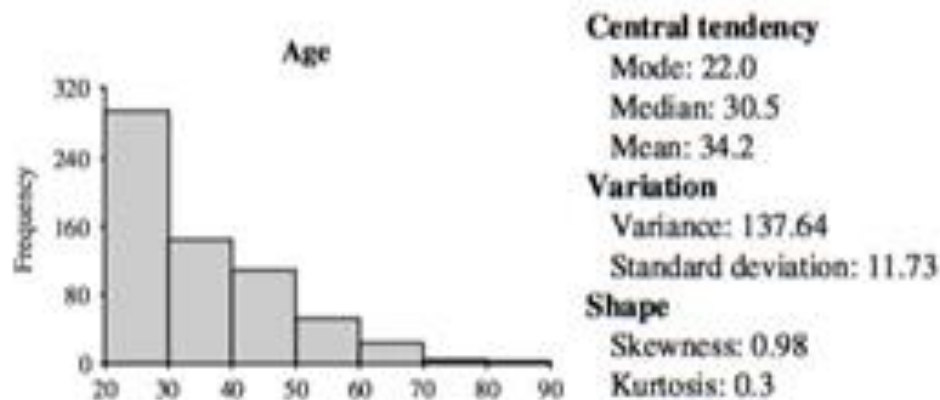
- ❖ At this point, we may want to conduct **hypothesis testing** (null and alternate hypothesis), and based on the **p-value** ( $p < 0.05$ ), conclude if there is a significant difference between the two groups on selected variables.

# Example of Questions

## ❖ Descriptive Questions:

- ❖ What is the age distribution of the sample population?

To answer the above question, **visualize and then summarize** based on visualization.



# Example of Questions

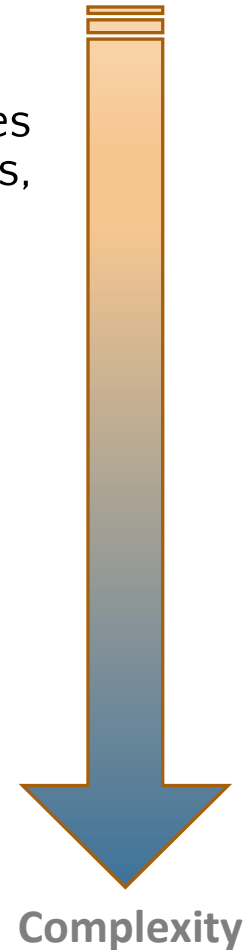
---

## ❖ Exploratory Question:

- ❖ For the group of people that is at risk of contracting diabetes and another group that is not at risk of contracting diabetes, which attributes are closely associated to each other (appearing together often)?

### Note

- The objective of this question was to identify general **associations between the different factors** and the **development of diabetes** that could be used for education and intervention purposes.
- Since the analysis makes use of categorical data, which requires the identification of associations, and must be easy to interpret, the associative rule mining (ARM) approach was selected



# ARM – Diabetes Group

If	Then	Support	Confidence	Lift
Plasma–Glucose (grouped) = high and Age (grouped) = 40–59	Diabetes = yes	6%	0.84	2.44
Plasma–Glucose (grouped) = high and BMI (grouped) = severely obese	Diabetes = yes	6.6%	0.82	2.37
Plasma–Glucose (grouped) = high and BMI (grouped) = obese	Diabetes = yes	5.6%	0.78	2.25
Pregnant (grouped) = high and Plasma–Glucose (grouped) = high	Diabetes = yes	7.5%	0.77	2.23

What can we conclude from the association rules and the measurement values ?

# ARM – No Diabetes Group

If	Then	Support	Confidence	Lift
BMI (grouped) = low and DPF (grouped) = low and Age (grouped) = 20–39	Diabetes = no	6%	1	1.53
Pregnant (grouped) = low and Plasma– Glucose (grouped) = medium and BMI (grouped) = low	Diabetes = no	5%	1	1.53
DiastolicBP (grouped) = normal and BMI (grouped) = low and DPF (grouped) = low and Age (grouped) = 20–39	Diabetes = no	5%	1	1.53
Pregnant (grouped) = low and Plasma– Glucose(grouped) = medium and DiastolicBP (grouped) = normal and DPF (grouped) = low and Age (grouped) = 20–39	Diabetes = no	8.7%	0.98	1.5
Plasma–Glucose (grouped) = medium and BMI (grouped) = low and Age (grouped) = 20–39	Diabetes = no	7.9%	0.98	1.5
Pregnant (grouped) = low and Plasma–Glucose (grouped) = low	Diabetes = no	6.6%	0.98	1.49
Pregnant (grouped) = low and BMI (grouped) = low	Diabetes = no	6.4%	0.98	1.49
Pregnant (grouped) = low and Plasma–Glucose(grouped) = low and Age (grouped) = 20–39	Diabetes = no	6.4%	0.98	1.49
Plasma–Glucose (grouped) = medium and DiastolicBP (grouped) = normal and BMI(grouped) = low and Age (grouped) = 20–39	Diabetes = no	6.4%	0.98	1.49
Pregnant (grouped) = low and Plasma–Glucose (grouped) = low and DiastolicBP (grouped) = normal	Diabetes = no	6.3%	0.98	1.49
Pregnant (grouped) = low and Plasma–Glucose (grouped) = low and DiastolicBP (grouped) = normal and Age (grouped) = 20–39	Diabetes = no	6.3%	0.98	1.49
Pregnant (grouped) = low and BMI (grouped) = low and Age (grouped) = 20–39	Diabetes = no	6.3%	0.98	1.49
Plasma–Glucose (grouped) = low and DPF (grouped) = low and Age (grouped) = 20–39	Diabetes = no	6.3%	0.98	1.49



# Example of Questions

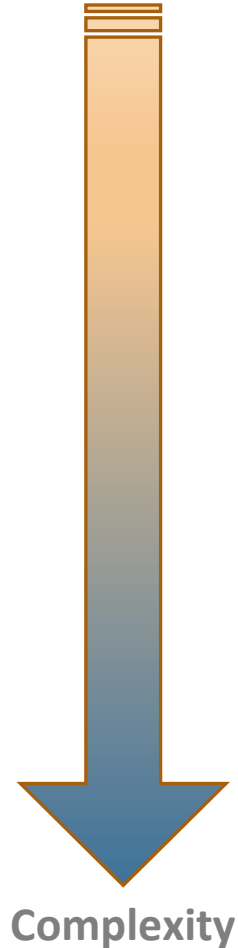
---

- ❖ **Predictive Question:**

- ❖ From the given the attributes that are used to describe the health of a person and the list of observations, is it possible to predict people who are at risk of contracting diabetes in the next 5 years?



Since the response variable (class) **is categorical**, we must develop a **classification** model.



# Choosing a Model

---

- ❖ There are many alternative classification modeling approaches that we can consider. Since there is no need to explain how these results were calculated, selecting a method that generates explanations or confidence values is not necessary.
- ❖ **k-Nearest Neighbors** and **Neural Network** approaches were selected to build the models. For both types of models, an experiment was designed to optimize the parameters used in generating the models.
- ❖ The analysis performed so far is critical to the process of developing prediction models. It helps us understand which variables are most influential, as well as helping us to interpret the results.

# Measure of Sensitivity/Specificity

		Predicted Response	
		True (1)	False (0)
Actual Response	True (1)	$Count_{11}$	$Count_{01}$
	False (0)	$Count_{10}$	$Count_{00}$

- **Count:**<sub>01</sub> The number of observations that were true and predicted to be false (false positives).
- **Count:**<sub>00</sub> The number of observations that were false and predicted to be false (true negatives).
- **Sensitivity:** This is an assessment of how well the model is able to predict 'true' values and the formula is:

$$Sensitivity = \frac{Count_{11}}{(Count_{11} + Count_{01})}$$

- **Specificity:** This is an assessment of how well the model is able to predict 'false' values and the formula is:

$$Specificity = \frac{Count_{00}}{(Count_{10} + Count_{00})}$$

# Optimization of kNN model

---

- ❖ Table in the next slide illustrates the optimization of the kNN (k-Nearest Neighbors) model using different descriptor variables with an optimal value for k. The Euclidean distance is chosen as the distance measure.
- ❖ From clinical view, the **risk associated with failure to diagnose diabetes is high**. Think of what will happen when a patient with higher risk of contracting diabetes is not identified by a model.

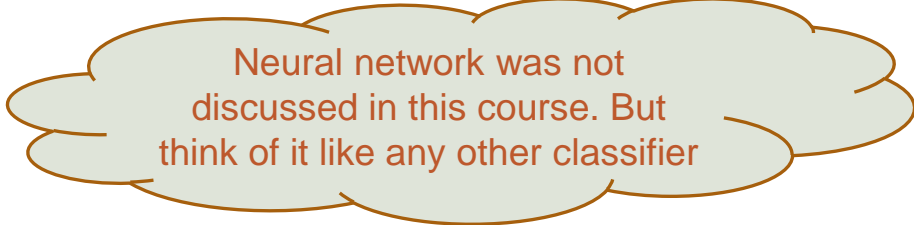
# Optimization of kNN parameters

Variable	Abbreviation
Pregnant	PRE
Plasma-Glucose	PG
DiastolicBP	DBP
BMI	BMI
DPF	DBF
Age	AGE

						k	Sensitivity/Specificity
PRE	PG	DBP	BMI	DPF	AGE	21	0.54/0.87
-	PG	DBP	BMI	DPF	AGE	29	0.6/0.88
PRE	-	DBP	BMI	DPF	AGE	29	0.41/0.85
PRE	PG	-	BMI	DPF	AGE	21	0.56/0.88
PRE	PG	DBP	-	DPF	AGE	22	0.53/0.89
PRE	PG	DBP	BMI	-	AGE	18	0.56/0.88
PRE	PG	DBP	BMI	DPF	-	29	0.51/0.9
-	-	DBP	BMI	DPF	AGE	28	0.41/0.86
-	PG	-	BMI	DPF	AGE	27	0.62/0.87
-	PG	DBP	-	DPF	AGE	29	0.58/0.88
-	PG	DBP	BMI	-	AGE	23	0.6/0.86
-	PG	DBP	BMI	DPF	-	29	0.5/0.88
PRE	-	-	BMI	DPF	AGE	16	0.38/0.88
PRE	-	DBP	-	DPF	AGE	28	0.28/0.91
PRE	-	DBP	BMI	-	AGE	27	0.41/0.84
PRE	-	DBP	BMI	DPF	-	25	0.33/0.87
PRE	PG	-	-	DPF	AGE	28	0.51/0.89
PRE	PG	-	BMI	-	AGE	29	0.53/0.87
PRE	PG	-	BMI	DPF	-	29	0.51/0.89
PRE	PG	DBP	-	-	AGE	29	0.54/0.84
PRE	PG	DBP	-	DPF	-	28	0.49/0.9
PRE	PG	DBP	BMI	-	-	29	0.52/0.87
-	-	-	BMI	DPF	AGE	23	0.5/0.86
-	-	DBP	-	DPF	AGE	23	0.39/0.85
-	-	DBP	BMI	-	AGE	27	0.46/0.81
-	-	DBP	BMI	DPF	-	29	0.35/0.89
-	PG	-	-	DPF	AGE	23	0.58/0.86
-	PG	-	BMI	-	AGE	28	0.6/0.87
-	PG	-	BMI	DPF	-	26	0.49/0.9
-	PG	DBP	-	-	AGE	25	0.56/0.88
-	PG	DBP	-	DPF	-	29	0.51/0.88
-	PG	DBP	BMI	-	-	28	0.46/0.89
PRE	-	-	-	DPF	AGE	29	0.37/0.85
PRE	-	-	BMI	-	AGE	24	0.42/0.86
PRE	-	-	BMI	DPF	-	27	0.36/0.88
PRE	-	DBP	-	-	AGE	28	0.34/0.85
PRE	-	DBP	-	DPF	-	29	0.29/0.88
PRE	-	DBP	BMI	-	-	29	0.31/0.88
PRE	PG	-	-	-	AGE	22	0.53/0.86
PRE	PG	-	BMI	-	-	28	0.54/0.89
PRE	PG	DBP	-	-	-	29	0.48/0.87
PRE	PG	-	-	-	-	29	0.48/0.88
PRE	-	DBP	-	-	-	29	0.2/0.88
PRE	-	-	BMI	-	-	20	0.31/0.87
PRE	-	-	-	DPF	-	28	0.24/0.91

# Optimization of Neural Network Models

---



Neural network was not discussed in this course. But think of it like any other classifier

- ❖ Next table shows a section of the optimization of the neural network models, using different input variables, different numbers of iterations, and different numbers of hidden layers. Again, the resulting model accuracy is displayed using the format “sensitivity/specificity”.

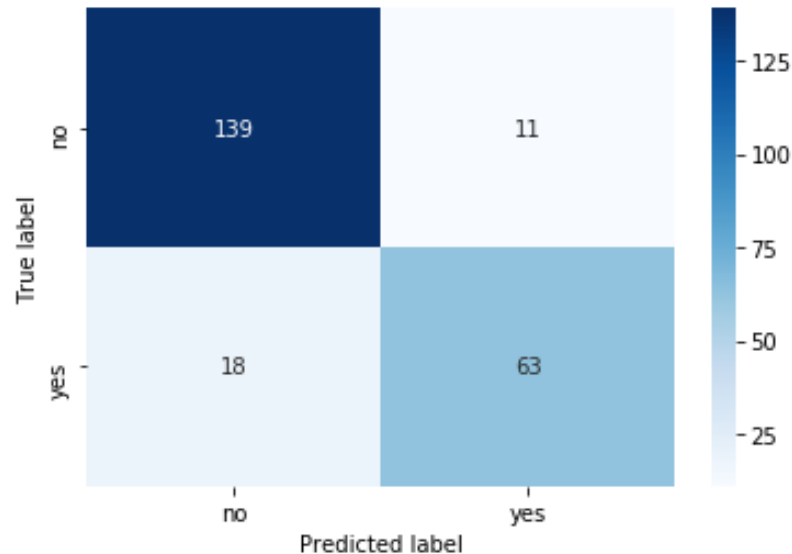
# Optimization of Neural Network Models

						1 hidden layer				2 hidden layers				3 hidden layers			
						5 K	20k	50 K	100 k	5 k	20 K	50 k	100 k	5 k	20 K	50 k	100 k
PRE	PG	DBP	BMI	DPF	AGE	0.55/0.89	0.62/0.85	0.63/0.80	0.65/0.84	0.58/0.85	0.59/0.84	0.66/0.86	0.64/0.83	0.23/0.89	0.52/0.84	0.58/0.84	0.65/0.83
-	PG	DBP	BMI	DPF	AGE	0.62/0.83	0.57/0.87	0.63/0.84	0.64/0.83	0.56/0.83	0.55/0.87	0.50/0.85	0.62/0.83	0.00/1.00	0.45/0.91	0.60/0.82	0.55/0.88
PRE	-	DBP	BMI	DPF	AGE	0.28/0.89	0.45/0.83	0.45/0.83	0.48/0.79	0.23/0.91	0.33/0.87	0.40/0.83	0.46/0.81	0.00/1.00	0.49/0.79	0.49/0.77	0.44/0.79
PRE	PG	-	BMI	DPF	AGE	0.57/0.83	0.64/0.84	0.57/0.88	0.61/0.83	0.35/0.90	0.62/0.83	0.60/0.85	0.60/0.83	0.00/1.00	0.41/0.92	0.61/0.89	0.61/0.85
PRE	PG	DBP	-	DPF	AGE	0.47/0.90	0.63/0.82	0.60/0.85	0.65/0.82	0.36/0.89	0.60/0.84	0.58/0.82	0.64/0.82	0.00/1.00	0.57/0.84	0.60/0.85	0.58/0.85
PRE	PG	DBP	BMI	-	AGE	0.52/0.90	0.60/0.87	0.62/0.82	0.56/0.85	0.48/0.81	0.59/0.87	0.70/0.77	0.57/0.87	0.10/0.97	0.47/0.85	0.64/0.83	0.63/0.86

- ❖ The following model gave the good overall performance (both sensitivity and specificity) and was selected:
  - ❖ A neural network with 2 hidden layers, 50,000 epochs, and a learning rate of 0.5 using 5 descriptors as inputs. The overall sensitivity for this model was 0.70 (70%) while specificity was 0.77 (77%).

# Other Measures

- ❖ In most other classification cases, the **Accuracy** or **F1-Score** (if classes are imbalanced) are popular choices.
- ❖ Often, a **Confusion matrix** gives a good overview of the performances in each class.



- ❖ Once the final model has been built, it is often a valuable exercise to look at observations that were not correctly predicted and analyse them.



# Reading Material

---

- Kaggle, “Pima Indians Diabetes Database”.
- Piotr Tynecki, “Predict diabetes diagnosis for Pima Female Indians with Logistic Regression”, 2018.
- Lahiru Liyanapathirana, “Machine Learning Workflow on Diabetes Data”, 2018.
- Andrea Grandi, “Machine Learning: Pima Indians Diabetes”, 2018.
- Love for Data Science, “Data Analysis and Visualization in Python (Pima Indians Diabetes Dataset)”, 2017.