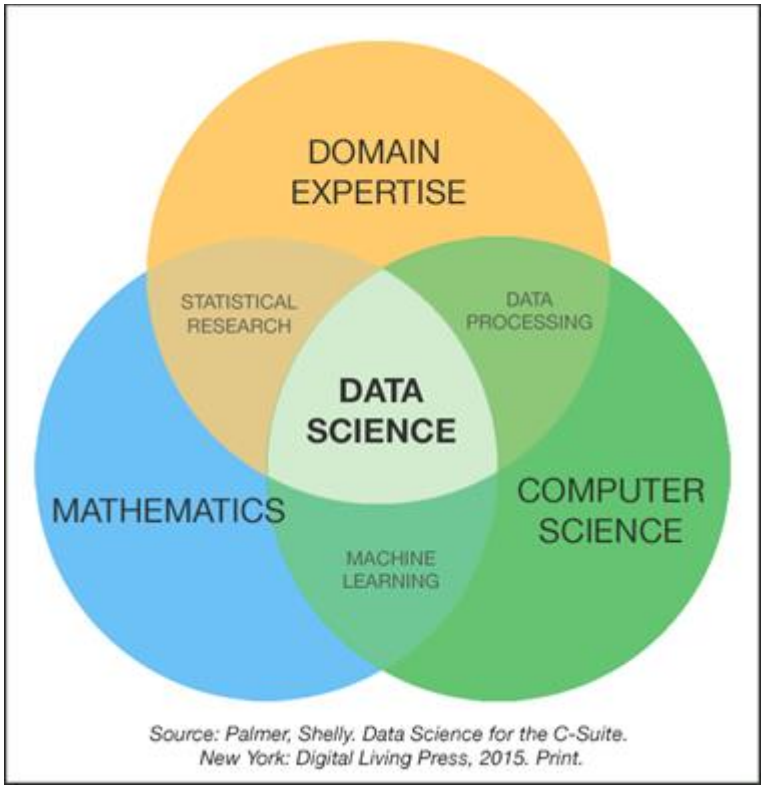
CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

CIENCIA DE DATOS:

APRENDE LOS FUNDAMENTOS DE MANERA PRÁCTICA



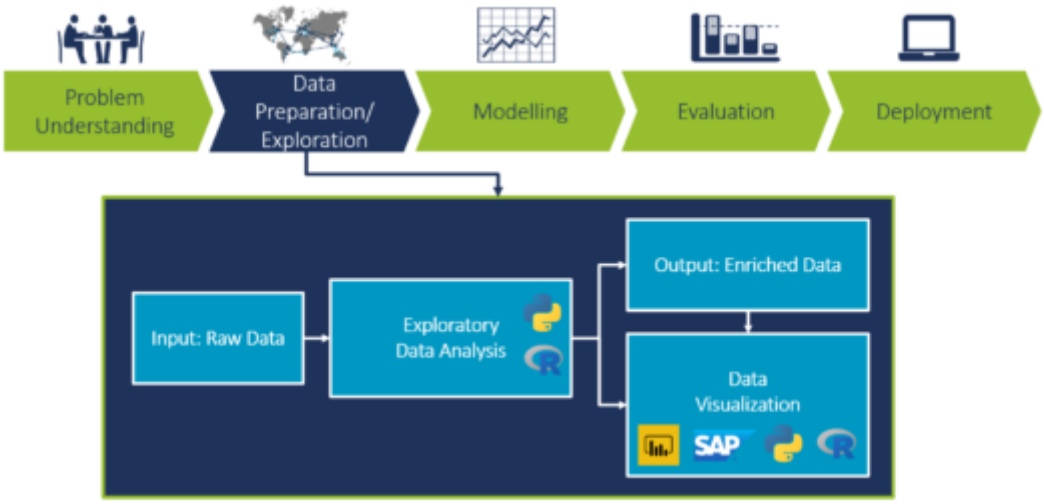
SESION 02

ESTADISTICA DESCRIPTIVA EXPLORATORY DATA ANALYSIS

Juan Antonio Chipoco Vidal
jchipoco@gmail.com

ÍNDICE

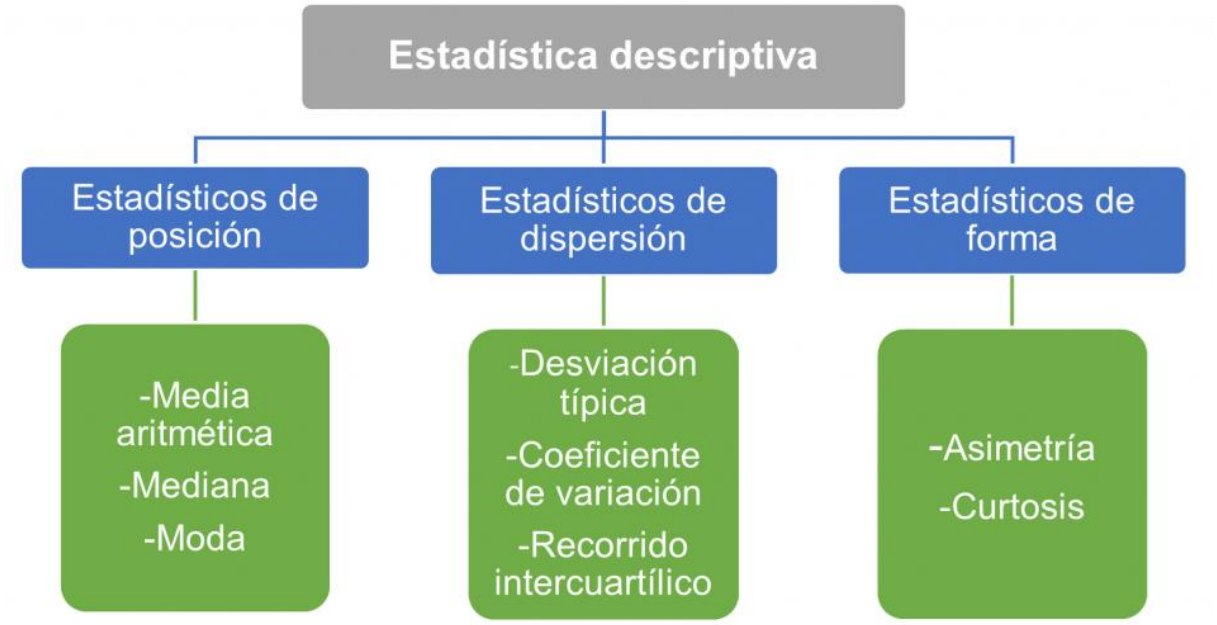
OBJETIVO



El objetivo de esta semana es realizar una revisión de la estadística descriptiva y aplicarla al Análisis Exploratorio de Datos (EDA)

En esta segunda sesión la práctica de laboratorio consistirá en llevar a cabo un análisis exploratorio de datos de la tragedia del Titanic para ello utilizaremos las principales librerías de Python, como son: numpy, pandas, scikit learn, matplotlib.

INTRODUCCION



La estadística descriptiva es el término que se le da al análisis de datos que ayuda a describir, mostrar o resumir datos de una manera significativa de modo que puedan surgir patrones a partir de dichos datos, eso se realiza con la ayuda de gráficos o valores de resumen. Sin embargo, la estadística descriptiva no nos permitirá sacar conclusiones más allá de los datos que hemos analizado ni llegar a conclusiones con respecto a las hipótesis que podríamos haber hecho. Es simplemente una manera de describir nuestros datos.

La estadística descriptiva es muy importante porque si presentamos nuestros datos sin procesar, sería difícil visualizar lo que muestran los datos, especialmente si son muchos. Por lo tanto, la estadística descriptiva nos permite presentar los datos de una manera más significativa, lo que permitirá una interpretación más sencilla de los datos. Por ejemplo, si tenemos los resultados de 100 exámenes de un curso, podríamos estar interesados en el desempeño general de los estudiantes. También estaríamos interesados en la distribución o difusión de las notas. La estadística descriptiva nos permite hacer esto.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son formas de describir la posición central de una distribución de frecuencia para un grupo de datos. En nuestro ejemplo anterior, la distribución de frecuencias es simplemente la distribución y el patrón de calificaciones obtenidas por los 100 estudiantes desde la más baja hasta la más alta. Podemos describir esta posición central utilizando una serie de estadísticas, incluida la moda, la mediana y la media.

Una medida de tendencia central es un valor único que intenta describir un conjunto de datos mediante la identificación de la posición central dentro de ese conjunto de datos. Como tales, las medidas de tendencia central a veces se denominan medidas de ubicación central. También se clasifican como estadísticas de resumen. La media (a menudo llamada promedio) es probablemente la medida de tendencia central que más conocemos, pero hay otras, como la mediana y la moda.

La media, la mediana y la moda son todas medidas válidas de tendencia central, pero en diferentes condiciones, algunas medidas de tendencia central se vuelven más apropiadas que otras. En las siguientes secciones, veremos la media, la moda y la mediana, y aprenderemos cómo calcularlos y en qué condiciones son más apropiados para su uso.

Media

La media (o promedio) es la medida de tendencia central más popular y conocida. Se puede usar tanto con datos discretos como continuos, aunque su uso es más frecuente con datos continuos. La media es igual a la suma de todos los valores del conjunto de datos dividida por el número de valores del conjunto de datos.

Media de la muestra:

$$\overline{x} = \frac{\sum x}{n}$$

Media de la población:

$$\mu = \frac{\sum x}{n}$$

Una propiedad importante de la media es que incluye todos los valores de su conjunto de datos como parte del cálculo. Además, la media es la única medida de tendencia central donde la suma de las desviaciones de cada valor de la media es siempre cero.

La media tiene una desventaja principal: es particularmente susceptible a la influencia de valores atípicos. Estos son valores que son inusuales en comparación con el resto del conjunto de datos por ser especialmente pequeños o grandes en valor numérico.

Mediana

La mediana es la puntuación media de un conjunto de datos que se ha organizado en orden de magnitud. La mediana se ve menos afectada por los valores atípicos y los datos sesgados.

Para calcular la mediana, supongamos que tenemos los siguientes datos (número impar de datos):

65	55	89	56	35	14	56	55	87	45	92

Primero ordenamos los datos en orden ascendente:

14	35	45	55	55	**56**	56	65	87	89	92

En el siguiente ejemplo tenemos un numero par de datos:

65	55	89	56	35	14	56	55	87	45

Primero ordenamos los datos en orden ascendente:

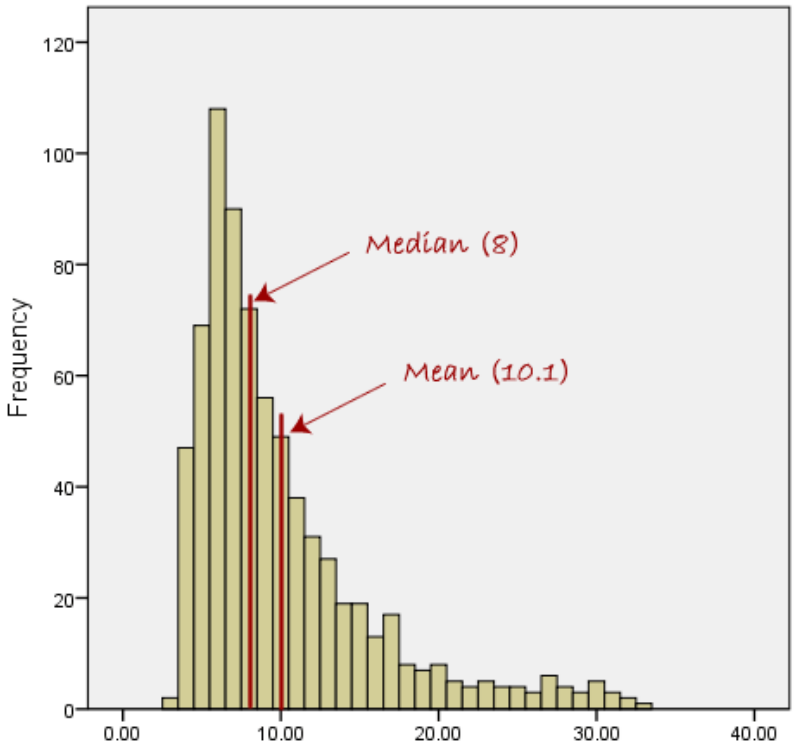
14	35	45	55	**55**	**56**	56	65	87	89

La mediana será el resultado del promedio de las notas 5 y 6 es decir: 55.5

Preferiremos usar la mediana sobre la media (o la moda) cuando nuestros datos están sesgados (es decir, la distribución de frecuencia de nuestros datos está sesgada). Si consideramos la distribución normal, ya que es la más evaluada en estadística, cuando los datos son perfectamente normales, la media, la mediana y la moda son idénticas. Además, todos representan el valor más típico en el conjunto de datos. Sin embargo, a medida que los datos se vuelven sesgados, la media pierde su capacidad de proporcionar la mejor ubicación central para los datos porque los datos sesgados los alejan del valor típico. Sin embargo, la

mediana conserva mejor esta posición y no está tan fuertemente influenciada por los valores sesgados.

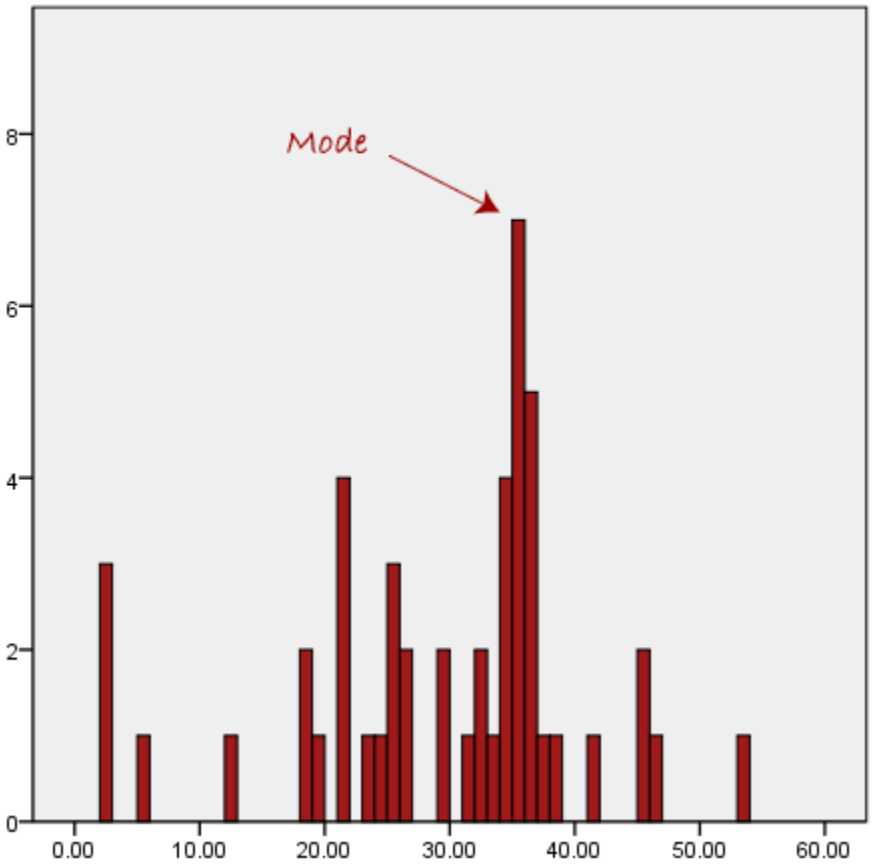
En el caso de datos sesgados, en la siguiente imagen encontramos que la media está siendo arrastrada en el sentido directo del sesgo. En estas situaciones, generalmente se considera que la mediana es el mejor representante de la ubicación central de los datos. Cuanto más sesgada sea la distribución, mayor será la diferencia entre la mediana y la media, y se debe poner mayor énfasis en usar la mediana en lugar de la media. Un ejemplo clásico de la distribución sesgada hacia la derecha salario, donde los que ganan más brindan una representación falsa del ingreso típico si se expresan como una media y no como una mediana.



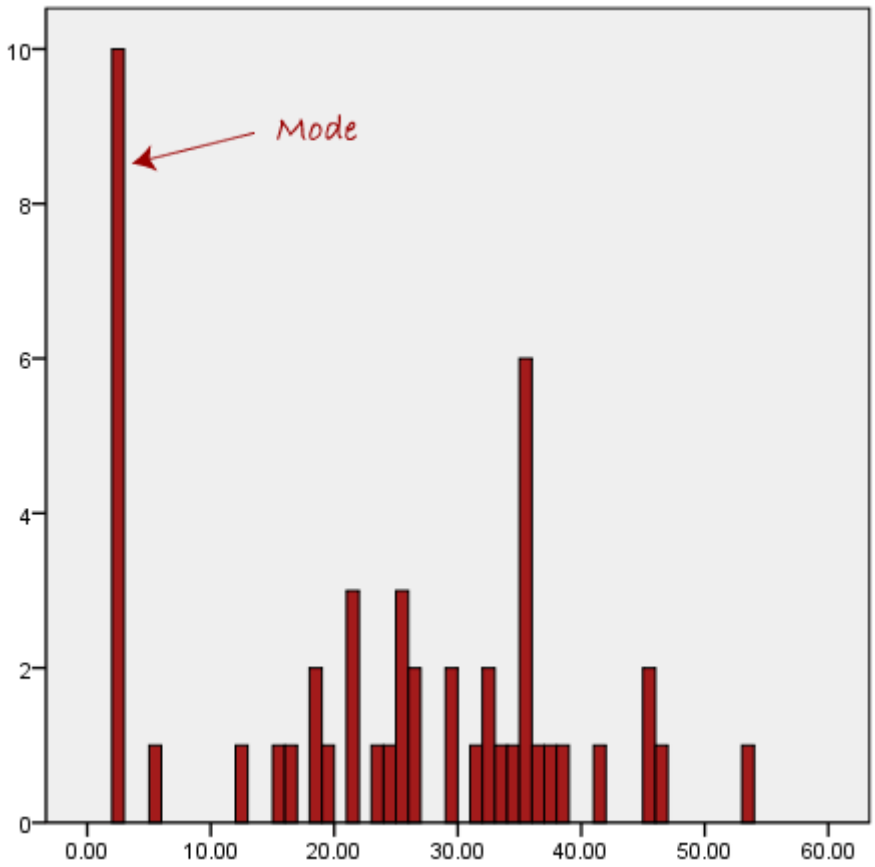
Moda

La moda es la puntuación más frecuente en nuestro conjunto de datos. En un histograma representa la barra más alta en un gráfico de barras o histograma. Por lo tanto, a veces puede considerar la moda como la opción más popular.

A continuación, se presenta un ejemplo de una moda:



Un problema con la moda es que no nos proporcionará una muy buena medida de tendencia central cuando la marca más común está lejos del resto de los datos en el conjunto de datos, como se muestra en el siguiente diagrama:



MEDIDAS DE DISPERSIÓN

Introducción

Las medidas de dispersión, se utiliza para describir la variabilidad en una muestra o población. Por lo general, se usa junto con una medida de tendencia central, como la media o la mediana, para proporcionar una descripción general de un conjunto de datos.

Por ejemplo, la puntuación media de los 100 alumnos puede ser de 65 sobre 100. Sin embargo, no todos los alumnos habrán obtenido 65 puntos. Más bien, sus puntajes se distribuirán. Unos serán más bajos y otros más altos. Las medidas de dispersión nos ayudan a resumir cuán dispersas están estas puntuaciones. Para describir este diferencial, tenemos a nuestra disposición una serie de estadísticas, algunas de ellas son los cuartiles, la varianza, la desviación standard y la correlación.

Cuartiles

Los cuartiles nos informan sobre la dispersión de un conjunto de datos dividiéndolo en cuartos, al igual que la mediana lo divide por la mitad. Por ejemplo, considere las calificaciones de los 100 estudiantes, a continuación, que se han ordenado de la calificación más baja a la más alta. En este caso datos no agrupados.

$$Q_1 = x_i + d \cdot (x_{i+1} - x_i)$$

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

N = 100;
Primer Cuartil => (N + 1)/4 = 25.25;
Segundo Cuartil => 2 (N + 1)/4 = 50.50;
Tercer Cuartil => 3(N + 1)/4 = 75.75
Q1 = 55 + 0.25(45 – 45) = 25
Q2 = 58 + 0.50(59 – 58) = 58.5
Q3 = 71 + 0.75(71 – 71) = 71

Varianza

Los cuartiles son útiles, pero también son algo limitados porque no tienen en cuenta todas las notas de nuestro grupo de datos. Para tener una idea más representativa de la dispersión, debemos tener en cuenta los valores reales de cada puntaje en un conjunto de datos. La varianza y la desviación estándar son tales medidas.

La varianza alcanza valores positivos elevando al cuadrado cada una de las desviaciones. La suma de estas desviaciones al cuadrado nos da la suma de los cuadrados, que luego podemos dividir por el número total de notas en nuestro grupo de datos (en otras palabras, 100 porque hay 100 estudiantes) para encontrar la varianza. Por lo tanto, para nuestros 100 estudiantes, la varianza es 211,89, como se muestra a continuación:

$$variance = \frac{\sum(X-\mu)^2}{N}$$

$$= \frac{21188.75}{100}$$

$$= 211.89$$

Where μ = mean, X = score, ∑= the sum of, N = number of scores, ∑X = "add up all the scores"

Como medida de variabilidad, la varianza es útil. Si las notas en nuestro grupo de datos están muy dispersas, la varianza será un número grande. Por el contrario, si las notas se distribuyen muy cerca de la media, la varianza será un número menor. Sin embargo, hay dos problemas potenciales con la varianza. En primer lugar, debido a que las desviaciones de las notas con respecto a la media se elevan al cuadrado, esto da más peso a las puntuaciones extremas. Si nuestros datos contienen valores atípicos (en otras palabras, uno o un pequeño número de puntajes que están particularmente lejos de la media y quizás no representan bien nuestros datos en su conjunto), esto puede deshacer el peso de estas notas. En segundo lugar, la varianza no está en las mismas unidades que las puntuaciones en nuestro conjunto de datos: la varianza se mide en unidades al cuadrado. Esto significa que no podemos ubicarlo en nuestra distribución de frecuencia y no podemos relacionar directamente su valor con los valores de nuestro conjunto de datos. Por lo tanto, la cifra de 211,89, nuestra varianza, parece algo arbitraria. Calcular la desviación estándar en lugar de la varianza corrige este problema. No obstante, el análisis de la varianza es extremadamente importante en algunos análisis estadísticos.

Desviación Standard

La desviación estándar es una medida de la dispersión de puntajes dentro de un conjunto de datos. Por lo general, estamos interesados en la desviación estándar de una población. Sin embargo, como a menudo se nos presentan datos de una muestra solamente, podemos estimar la desviación estándar de la población a partir de una desviación estándar de la muestra. Estas dos desviaciones estándar (desviaciones estándar de la muestra y de la población) se calculan de manera diferente. En estadística, generalmente se nos presenta el tener que calcular las desviaciones estándar de la muestra, aunque también se mostrará la fórmula para una desviación estándar de la población.

La desviación estándar se usa junto con la media para resumir datos continuos, no datos categóricos. Además, la desviación estándar, como la media, normalmente solo es adecuada cuando los datos continuos no están significativamente sesgados o tienen valores atípicos.

Fórmula de la desviación standard para una muestra:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Fórmula de la desviación estándar para una población:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

Covarianza

La desviación típica es un indicador de dispersión de una variable. ¿Qué pasa cuando tienes más de una variable? ¿Existe alguna forma de saber cómo se relaciona una con la otra?

La Covarianza es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas.

$$\sigma_{XY} = \sum_{i=1}^{N} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

La covarianza positiva: Cuando uno variable crece la otra variable también. Tienen una relación directa.

La covarianza negativa: Cuando una variable crece la otra variable decrece. Tienen una relación Inversa.

Coeficiente de Correlación

La correlación es un indicador para saber si hay relación (LINEAL) entre dos variables numéricas para esto se utiliza el coeficiente de correlación o correlación de Pearson.

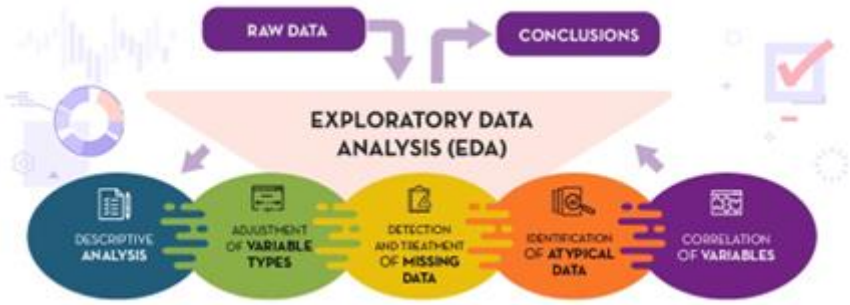
La correlación contesta preguntas como las siguientes:

- ¿La práctica de algún deporte está relacionada con una vida más longeva?
- ¿Existe una relación entre la cantidad de carne ingerida diariamente y el cáncer?
- ¿Mayor estudio implica mejores notas en un examen?

La correlación es un ratio entre la dispersión entre las dos variables conjuntamente (covarianza) y la desviación standard de cada variable.

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

ANÁLISIS EXPLORATORIO DE DATOS - EDA



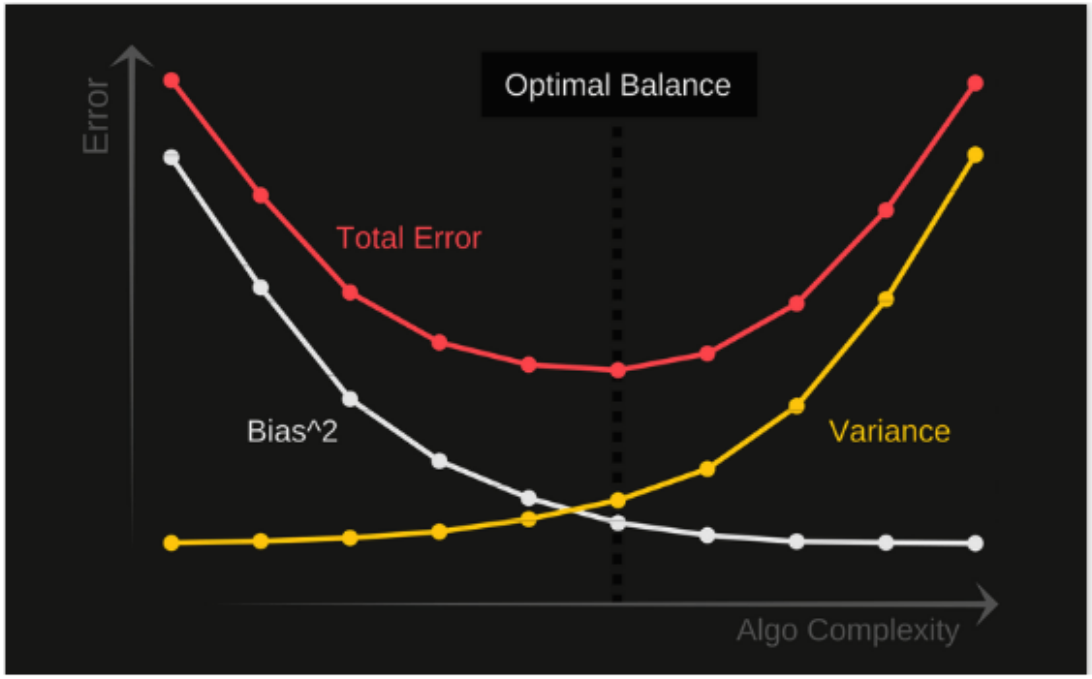
Antes de realizar análisis de datos, con fines estadísticos o predictivos, usando por ejemplo técnicas de aprendizaje automático, es necesario entender la materia prima (raw data) con la que vamos a trabajar. Es necesario comprender y evaluar la calidad de los datos para, entre otros aspectos, detectar y tratar los datos atípicos (outliers) o incorrectos, evitando posibles errores que puedan repercutir en los resultados del análisis.

EDA consiste en aplicar un conjunto de técnicas estadísticas destinadas a explorar, describir y resumir la naturaleza de los datos, de forma que podamos entender claramente como están relacionadas nuestras variables de interés.

Todo esto nos permite identificar posibles errores, revelar la presencia de outliers, comprobar la relación entre variables (correlaciones) y su posible redundancia, y realizar un análisis descriptivo de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos.

MACHINE LEARNING

Bias-variance tradeoff

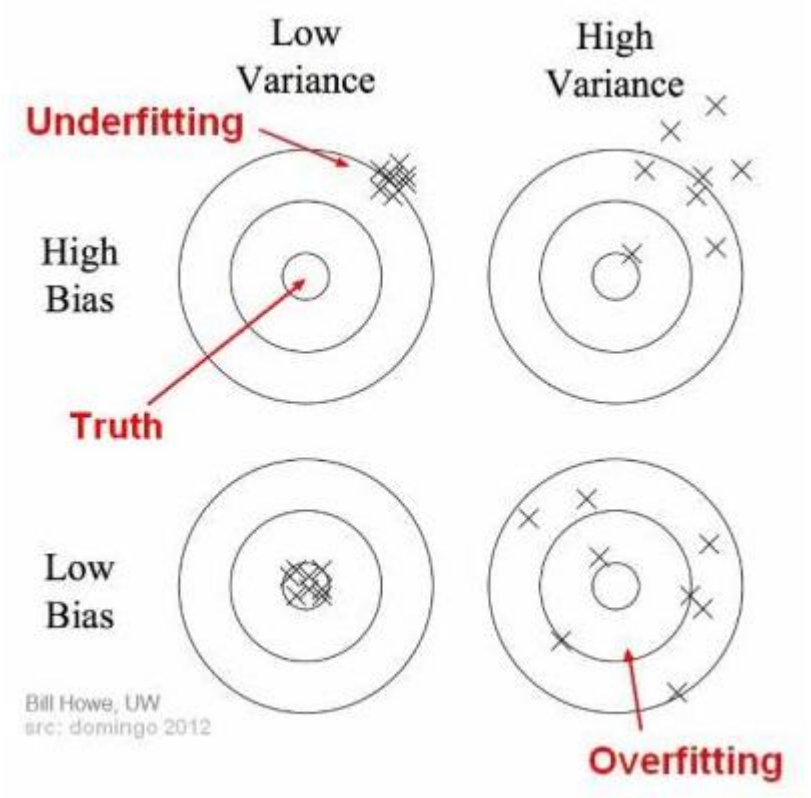


Cada vez que discutimos la predicción del modelo, es importante comprender los errores de predicción (sesgo y varianza). Existe una compensación entre la capacidad de un modelo para minimizar el sesgo y la varianza. Obtener una comprensión adecuada de estos errores nos ayudara no solo a construir modelos precisos, sino también a evitar el error de sobreajuste y ajuste insuficiente.

Bias: El sesgo es la diferencia entre la predicción promedio de nuestro modelo y el valor correcto que estamos tratando de predecir. El modelo con alto sesgo presta muy poca atención a los datos de entrenamiento y simplifica demasiado el modelo. Siempre conduce a un alto error en los datos de entrenamiento y prueba.

Variance: La varianza es la variabilidad de la predicción del modelo para un punto de datos dado o un valor que nos indica la dispersión de nuestros datos. El modelo con alta varianza presta mucha atención a los datos de entrenamiento y no generaliza sobre los datos que no ha visto antes. Como resultado, estos modelos funcionan muy bien con los datos de entrenamiento, pero tienen altas tasas de error con los datos de prueba.

Evaluación del modelo



En el diagrama anterior, el centro del objetivo es un modelo que predice perfectamente los valores correctos. A medida que nos alejamos de la diana, nuestras predicciones empeoran cada vez más. Podemos repetir nuestro proceso de creación de modelos para obtener resultados separados en el objetivo.

En el aprendizaje supervisado, el ***underfitting*** ocurre cuando un modelo no puede capturar el patrón subyacente de los datos. Estos modelos suelen tener un alto sesgo y una baja varianza. Ocurre cuando tenemos muy poca cantidad de datos para construir un modelo preciso o cuando intentamos construir un modelo lineal con datos no lineales.

En el aprendizaje supervisado, el ***overfitting*** ocurre cuando nuestro modelo captura el ruido junto con el patrón subyacente en los datos. Ocurre cuando entrenamos mucho nuestro modelo sobre un conjunto de datos ruidoso. Estos modelos tienen un sesgo bajo y una varianza alta.