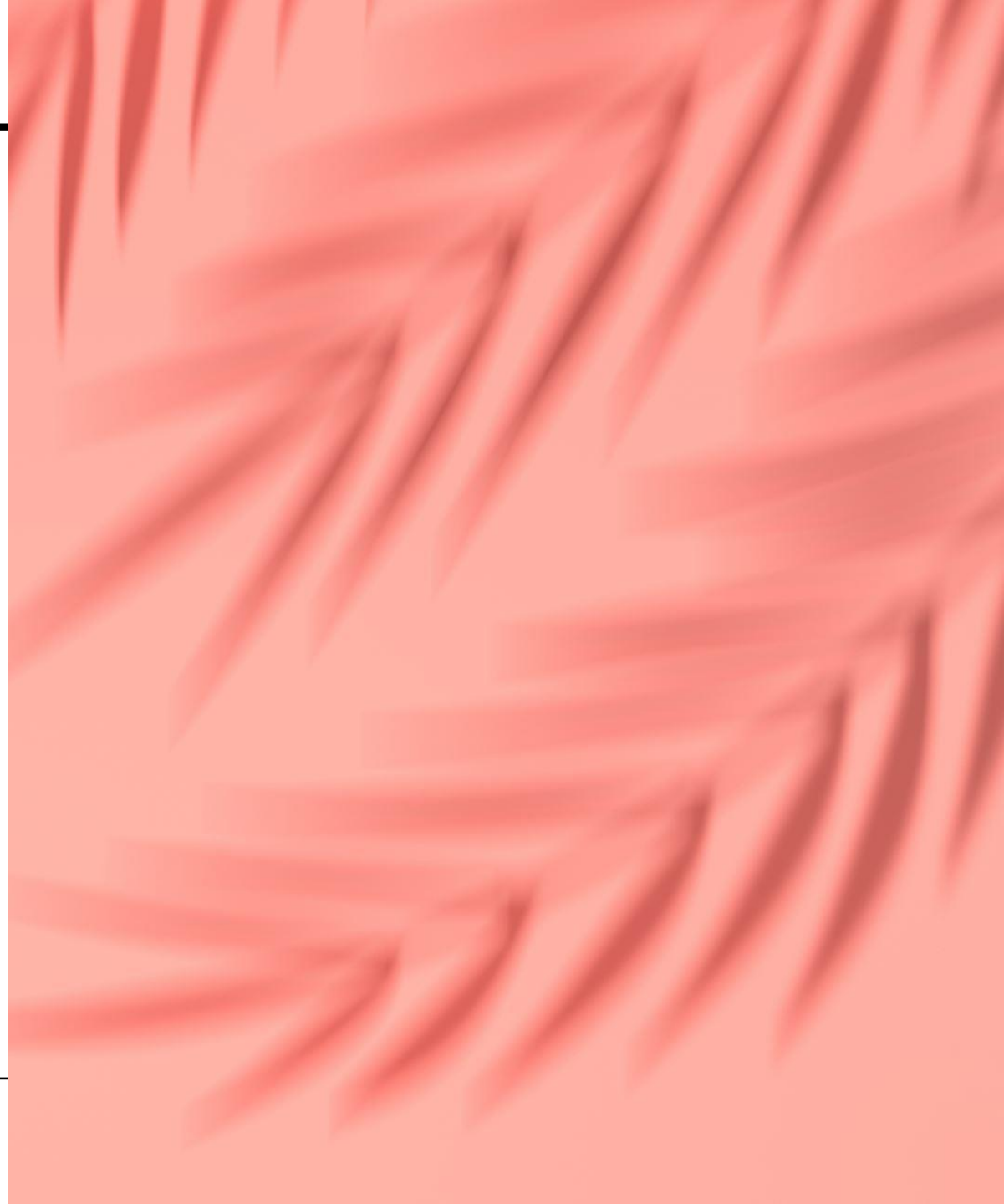
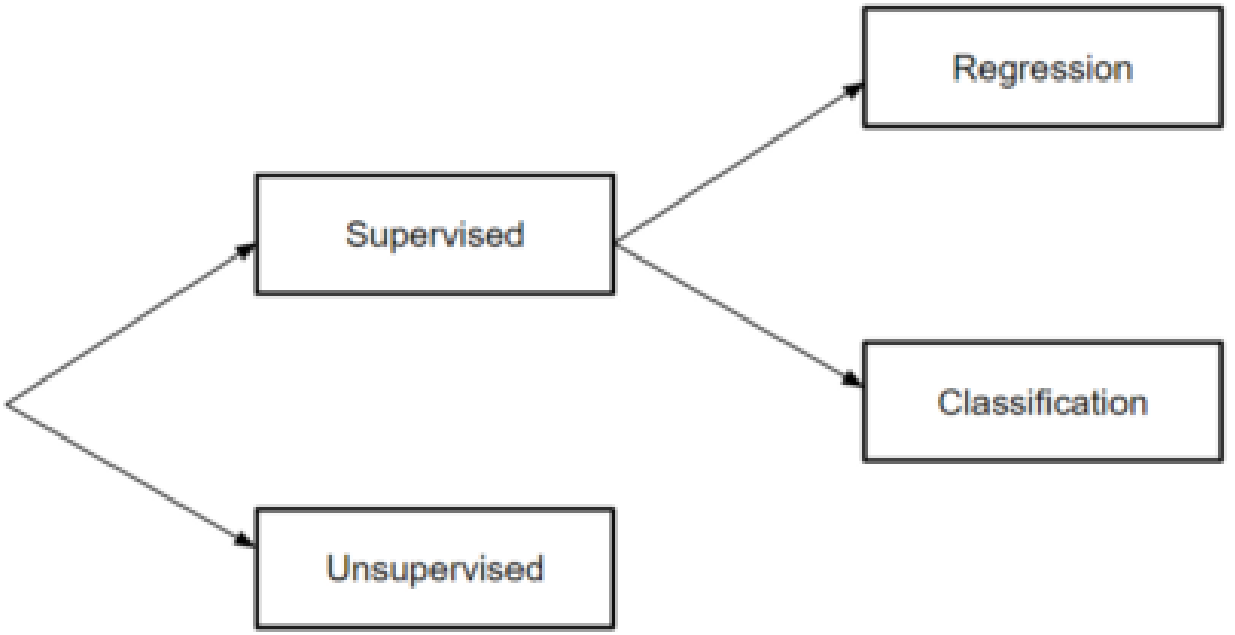
ELECCIÓN DE MODELOS ML

Artefacto tercera entrega.

BREVE INTRODUCCIÓN MODELOS

Artefacto de tercera entrega.

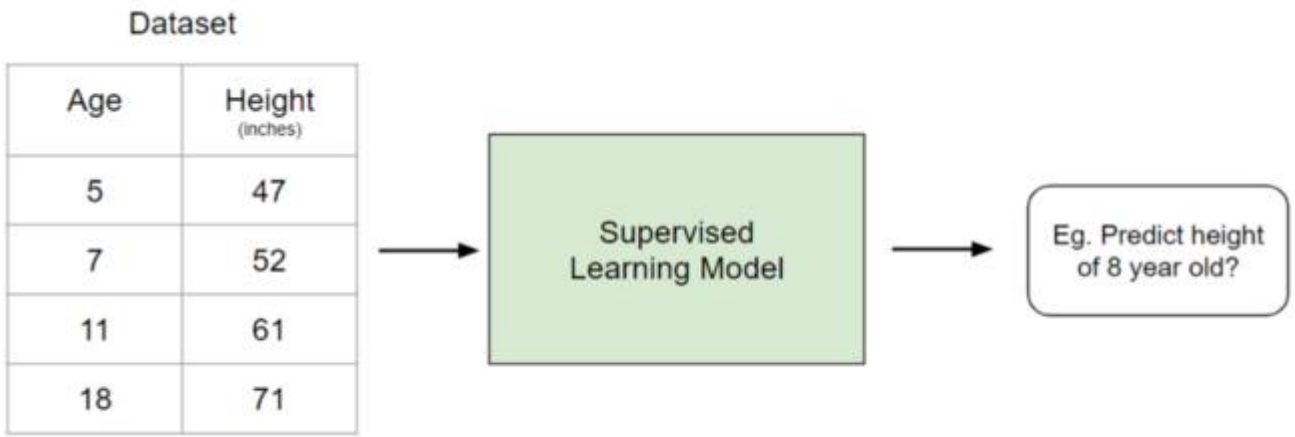




Todos los modelos de aprendizaje automático se clasifican como **supervisados** o **no supervisados** . Si el modelo es un modelo supervisado, se subcategoriza como modelo de **regresión** o de **clasificación** .

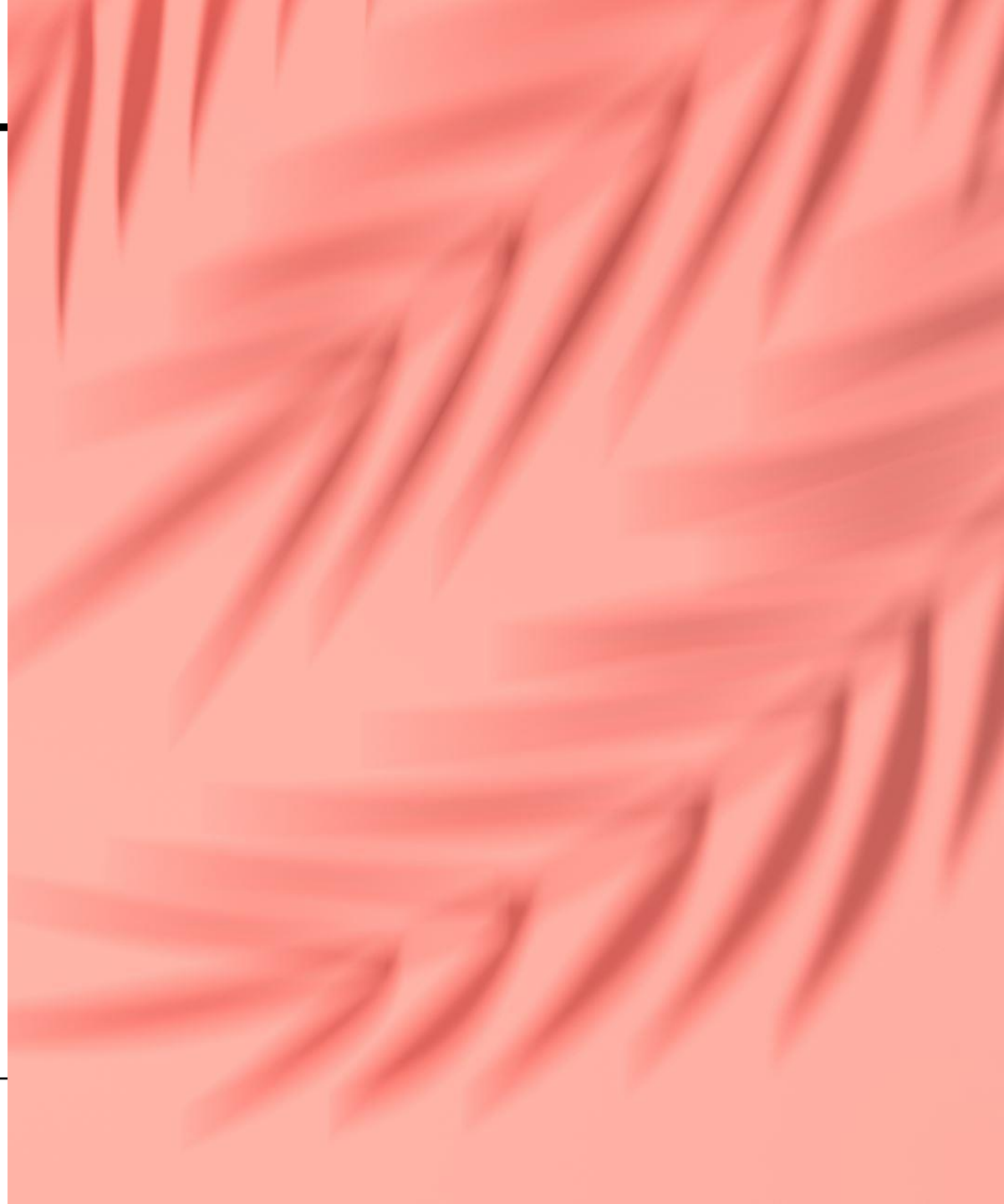
El aprendizaje supervisado implica aprender una función que asigna una entrada a una salida en función de pares de entrada-salida.

Por ejemplo, si tuviera un conjunto de datos con dos variables, edad (entrada) y altura (salida), podría implementar un modelo de aprendizaje supervisado para predecir la altura de una persona en función de su edad.



REGRESIÓN

Artefacto de tercera entrega.

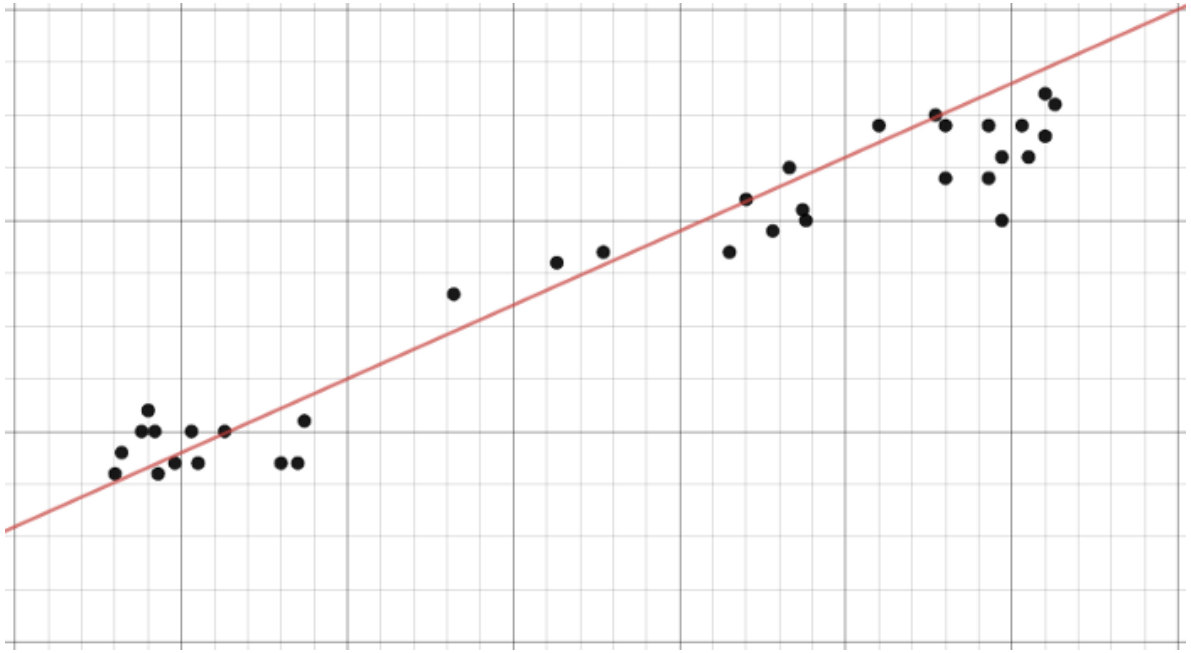


REGRESION

En los modelos de **regresión , la salida es continua.** A continuación se presentan algunos de los tipos más comunes de modelos de regresión.

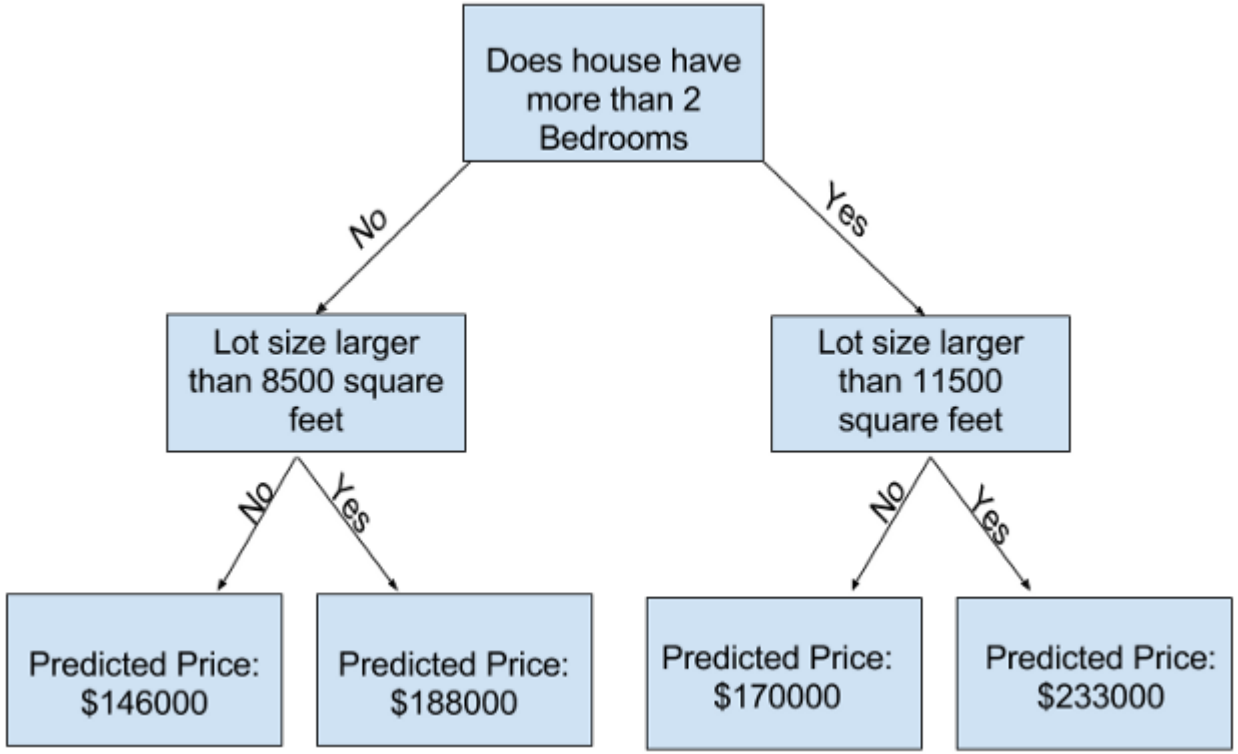
REGRESION LINEAL

La idea de la regresión lineal es simplemente encontrar una línea que mejor se ajuste a los datos. Las extensiones de la regresión lineal incluyen la regresión lineal múltiple y la regresión polinomial.



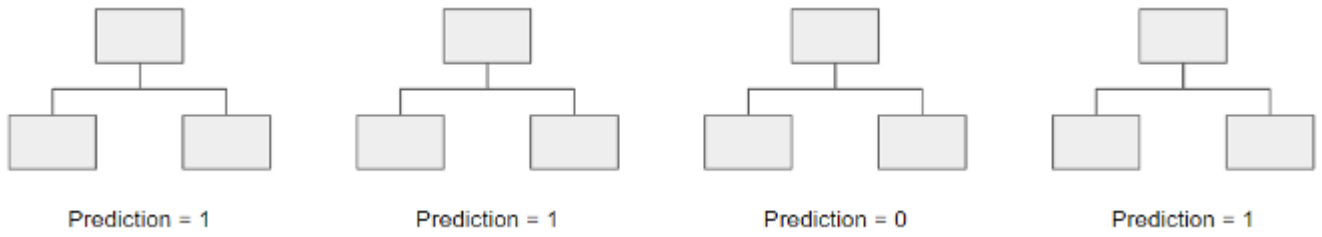
ARBOL DE DECISION

Los árboles de decisión son un modelo popular que se utiliza en la investigación de operaciones, la planificación estratégica y el aprendizaje automático. Cada cuadrado se llama **nodo** , y cuantos más nodos tenga, más preciso será el árbol de decisión (generalmente). Los últimos nodos del árbol de decisión, donde se toma una decisión, se denominan **hojas** del árbol. Los árboles de decisión son intuitivos y fáciles de construir, pero se quedan cortos cuando se trata de precisión.



BOSQUE ALEATORIO

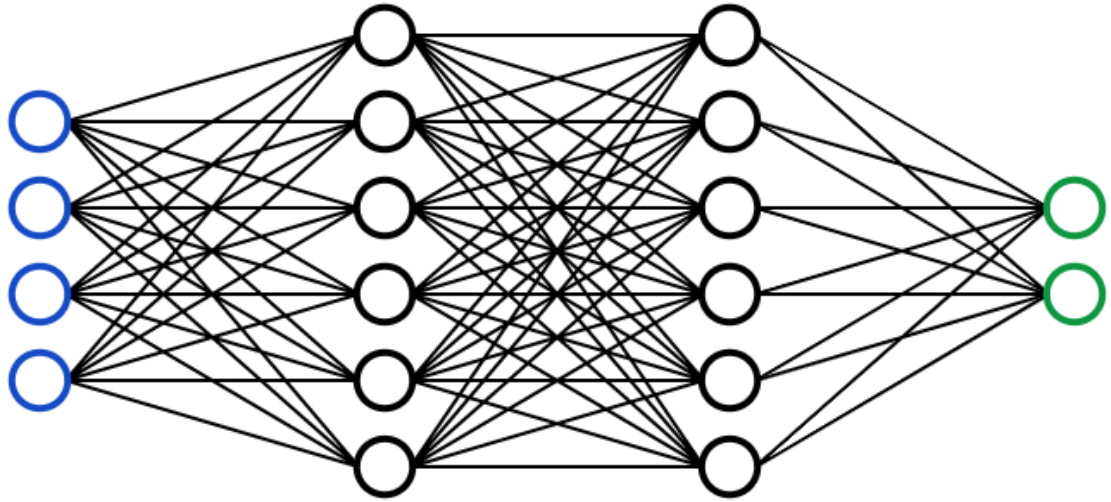
Los bosques aleatorios son una técnica de aprendizaje en conjunto que se basa en árboles de decisión. Los bosques aleatorios implican la creación de múltiples árboles de decisión utilizando conjuntos de datos de arranque de los datos originales y la selección aleatoria de un subconjunto de variables en cada paso del árbol de decisión. Luego, el modelo selecciona el modo de todas las predicciones de cada árbol de decisión. ¿Cuál es el punto de esto? Al basarse en un modelo de "ganancias mayoritarias", reduce el riesgo de error de un árbol individual.



BOSQUE ALEATORIO

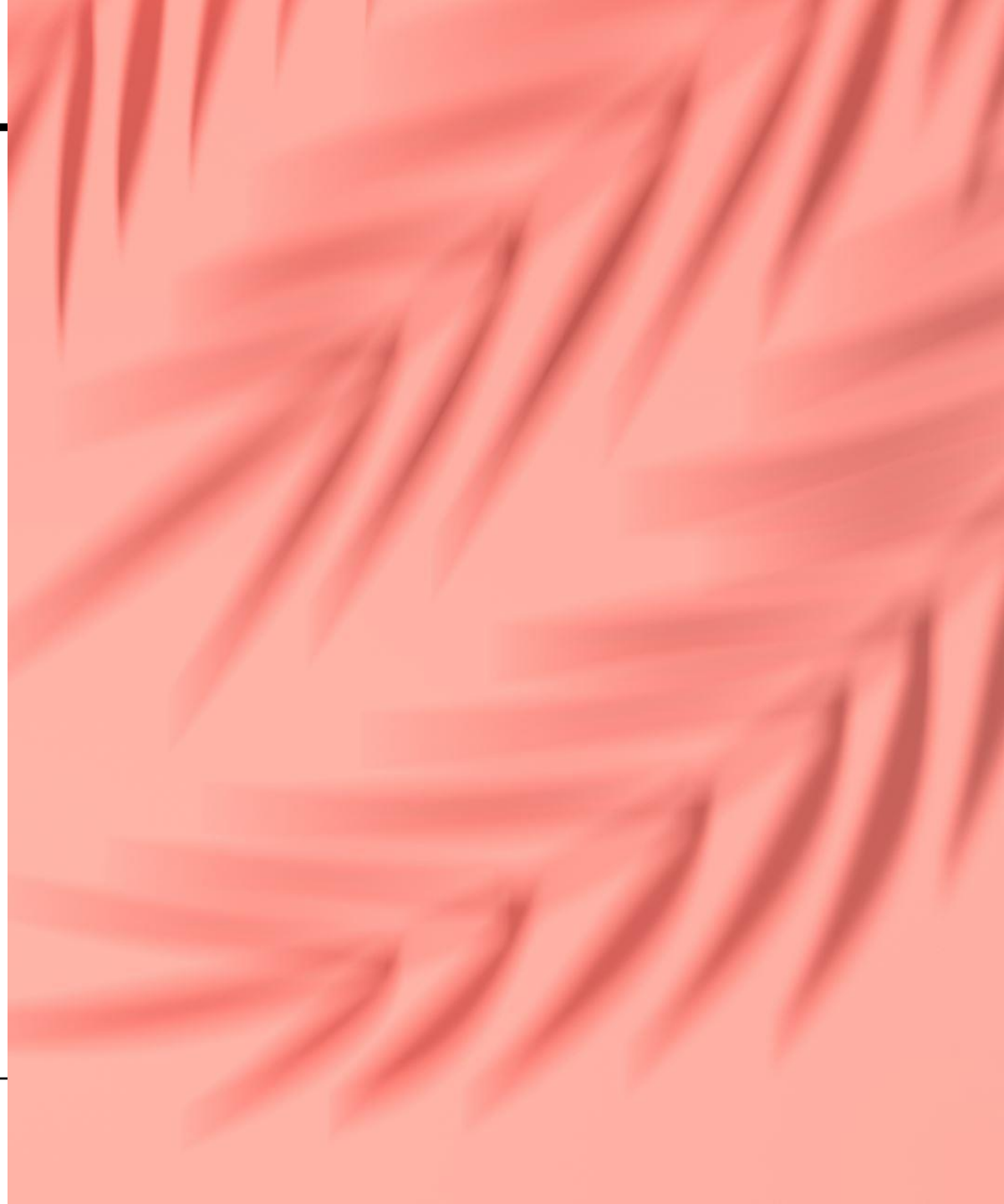
Una red neuronal es esencialmente una red de ecuaciones matemáticas. Toma una o más variables de entrada y, al pasar por una red de ecuaciones, da como resultado una o más variables de salida.

Los círculos azules representan la capa de entrada, los círculos negros representan las capas ocultas y los círculos verdes representan la capa de salida. Cada nodo en las capas ocultas representa tanto una función lineal como una función de activación por la que pasan los nodos en la capa anterior, lo que finalmente conduce a una salida en los círculos verdes.



CLASIFICACIÓN

Artefacto de tercera entrega.



Clasificación

En los modelos de clasificación, la salida es discreta. A continuación se presentan algunos de los tipos más comunes de modelos de clasificación.

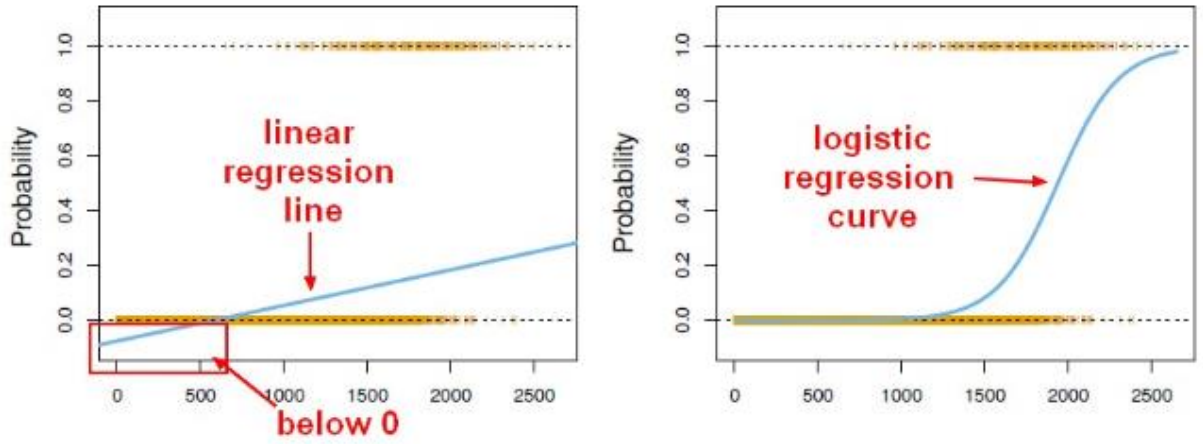
REGRESIÓN LOGÍSTICA

La regresión logística es similar a la regresión lineal, pero se usa para modelar la probabilidad de un número finito de resultados, generalmente dos. En esencia, se crea una ecuación logística de tal manera que los valores de salida solo pueden estar entre 0 y 1.

- **Salida de regresión lineal como probabilidades**

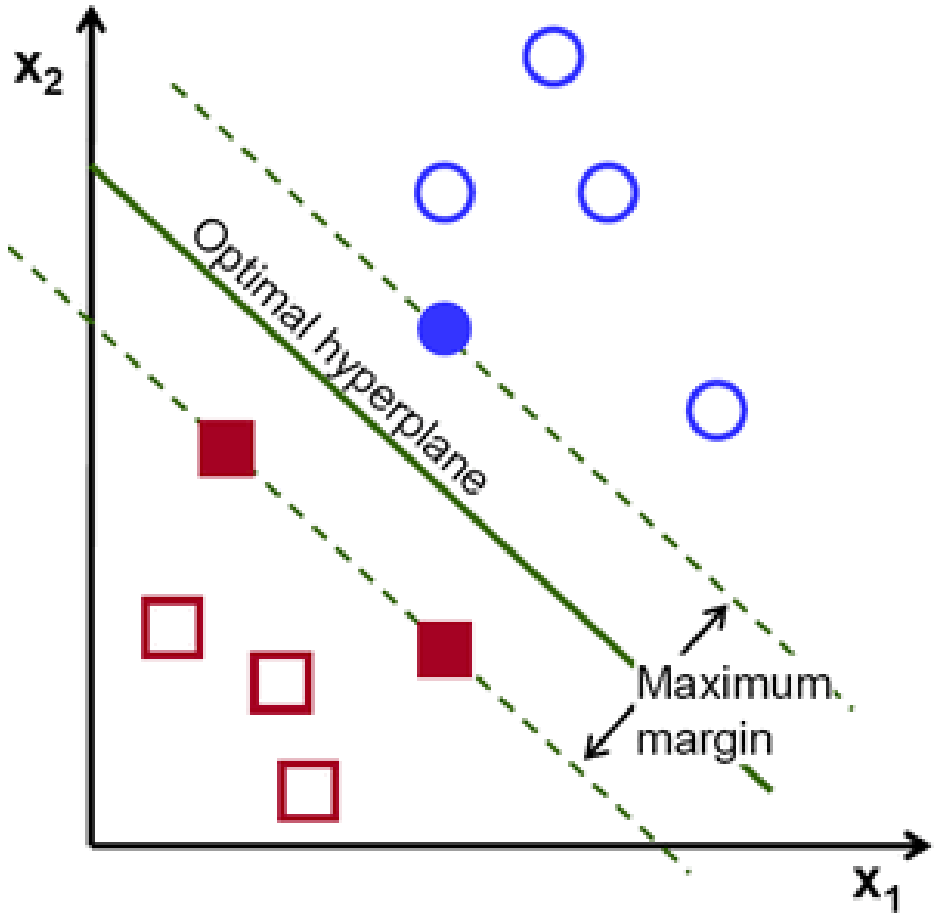
  Es tentador usar la salida de la regresión lineal como probabilidades, pero es un error porque la salida puede ser negativa y mayor que 1, mientras que la probabilidad no. Como la regresión en realidad podría producir probabilidades que podrían ser menores que 0, o incluso mayores que 1, se introdujo la regresión logística.

  Fuente: http://gerardnico.com/wiki/data_mining/simple_logistic_regression



MAQUINAS DE VECTORES DE SOPORTE

Una **máquina de vectores de soporte** es una técnica de clasificación supervisada que en realidad puede volverse bastante complicada pero es bastante intuitiva en el nivel más fundamental.
Supongamos que hay dos clases de datos. Una máquina de vectores de soporte encontrará un **hiperplano** o un límite entre las dos clases de datos que maximiza el margen entre las dos clases . Hay muchos planos que pueden separar las dos clases, pero solo un plano puede maximizar el margen o la distancia entre las clases.



NAIVE BAYES

Naive Bayes es otro clasificador popular utilizado en Data Science. La idea detrás de esto está impulsada por el Teorema de Bayes:

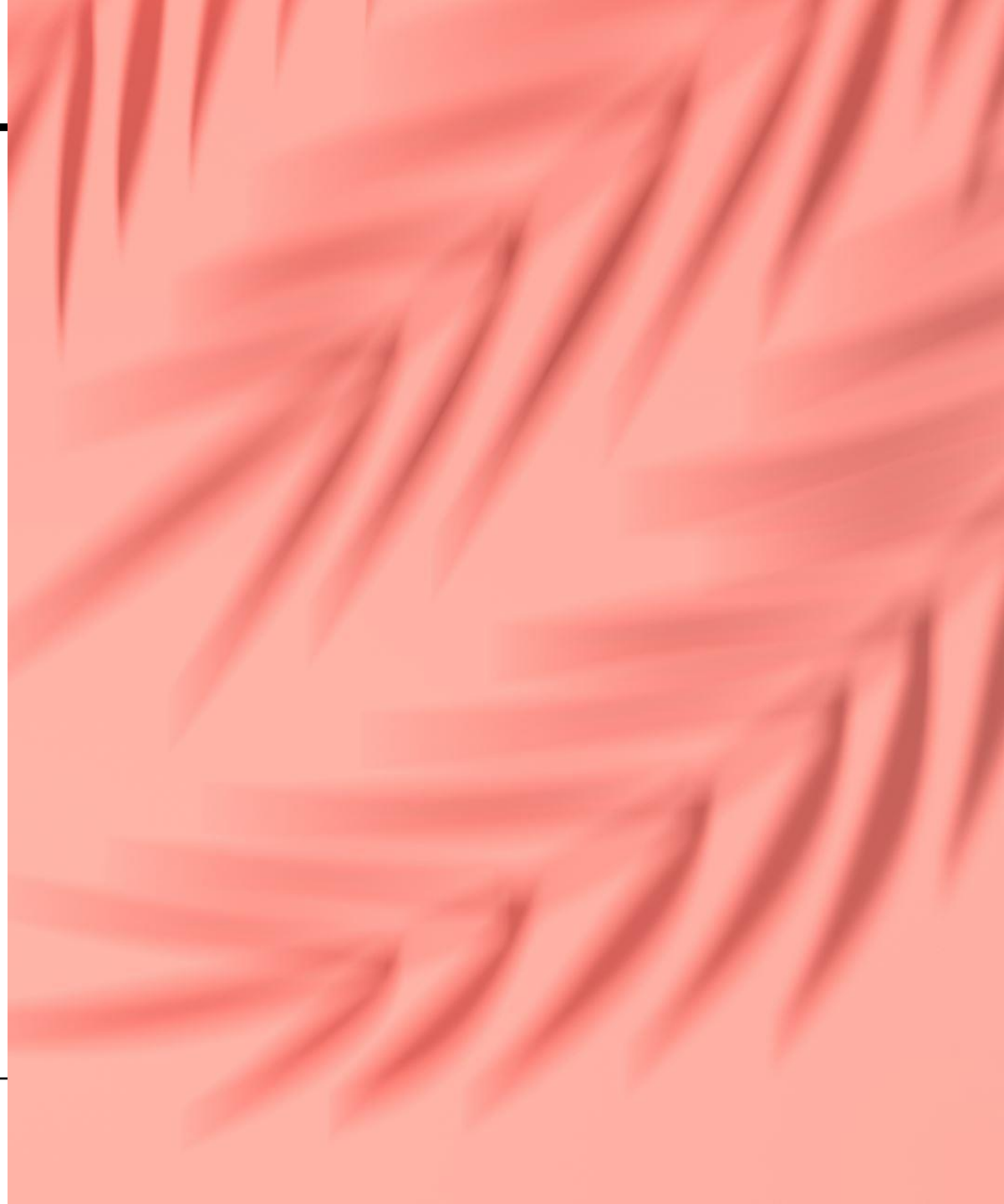
$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

En lenguaje sencillo, esta ecuación se usa para responder la siguiente pregunta. “¿Cuál es la probabilidad de y (mi variable de salida) dada X? Y debido a la suposición ingenua de que las variables son independientes dada la clase, puedes decir que:

$$P(X|y) = P(x2|y) * P(x2|y) * \ldots * P(xn|y)$$

ALGORITMOS Y MODELOS ELEGIDOS

Artefacto de tercera entrega.



REGRESION LINEAL SIMPLE.

- Predecir el rendimiento académico en función de cantidad de respuestas correctas.
- Predecir porcentaje de aciertos promedio en función de pruebas anteriores
- Predecir porcentaje de error promedio en función de pruebas anteriores

Alumno
Grupal
Prueba

De hecho, incluso la mejor información no cuenta una historia completa. La investigación de regresión generalmente se utiliza en la investigación para establecer que existe una relación entre las variables. Sin embargo, la correlación no es equivalente a la causalidad: una conexión entre dos variables no significa que una cause que ocurra la otra. De hecho, incluso una línea en una regresión lineal simple que se ajuste bien a los enfoques de información puede no garantizar una relación de circunstancias y resultados lógicos.

NAIVE BAYES

- Tomar datos de predicciones lineales y calcular predicciones combinadas para distintos alumnos, grupos y pruebas de distintos tópicos.

Alumno
Grupal
Prueba

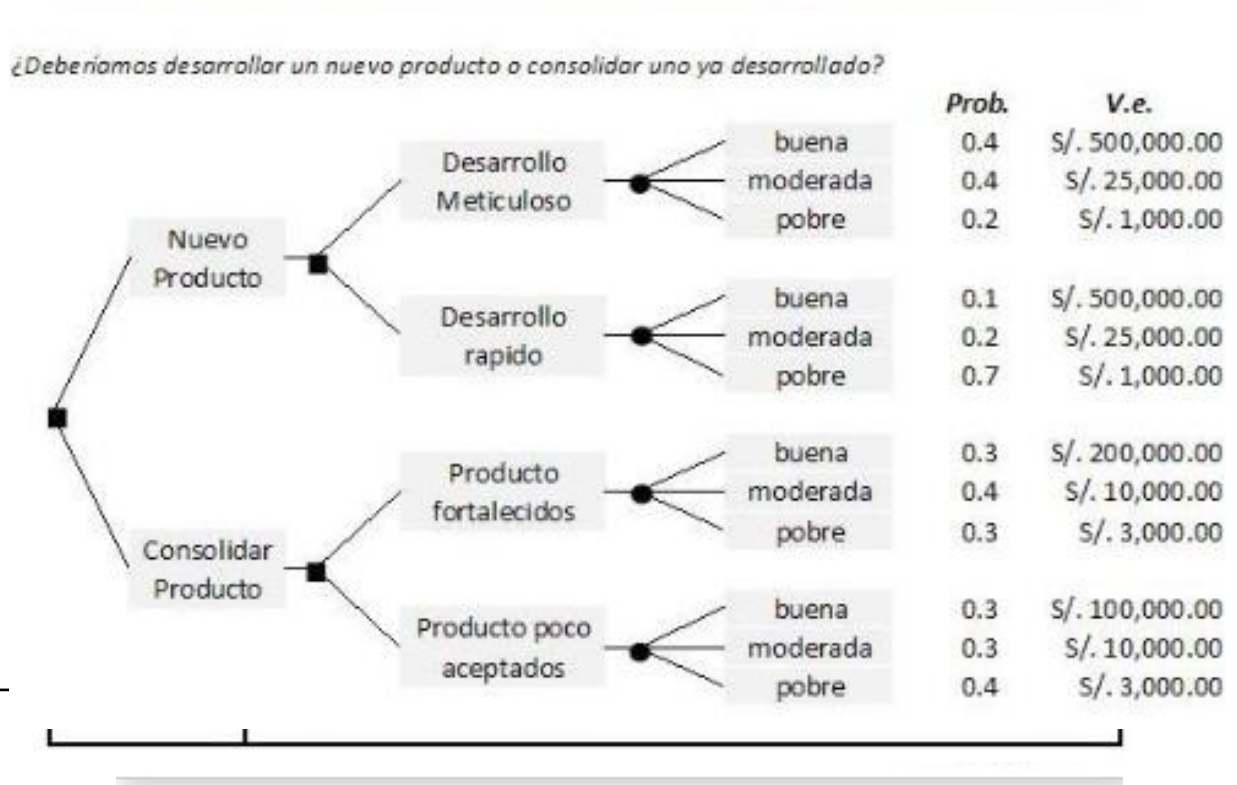
Desventajas de Naive Bayes

- Si su conjunto de datos de prueba tiene una variable categórica de una categoría que no estaba presente en el conjunto de datos de entrenamiento, el modelo Naive Bayes le asignará una probabilidad cero y no podrá hacer ninguna predicción al respecto. Este fenómeno se llama 'Frecuencia cero' y tendrá que usar una técnica de suavizado para resolver este problema.
- Este algoritmo también es conocido como un pésimo estimador. Por lo tanto, no debe tomar demasiado en serio los resultados de probabilidad de 'predict_proba'.
- Asume que todas las características son independientes. Si bien puede sonar genial en teoría, en la vida real, difícilmente encontrará un conjunto de funciones independientes.

Arboles de decisión

- Evaluación de pruebas con opciones predeterminadas
- Determinar VALOR ESPERADO de un comportamiento. (Decir si bajo un parámetro algo fue positivo o negativo)

Alumno
Grupal
Prueba



Matrices de ponderación

- Para el buscador y recomendaciones

Sección de comunidad

Buscador (Recommender System Information)

Norm.
.216
.0833
.1833
.133

X

POO.	Python Básico	Java Básico	C Básico.
1	1	1	0
1	0	0	1
1	0	1	1

Cuestionarios candidatos.

=

POO.	Python Básico	Java Básico	C Básico.
.216	.083	.183	0
.216	0	0	.133
.216	0	.183	.133

Matriz de ponderación.

=

Puntaje.
.482
.349
.532

C3

Matriz de recomendación.