EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

Machine Learning 0 - Intro

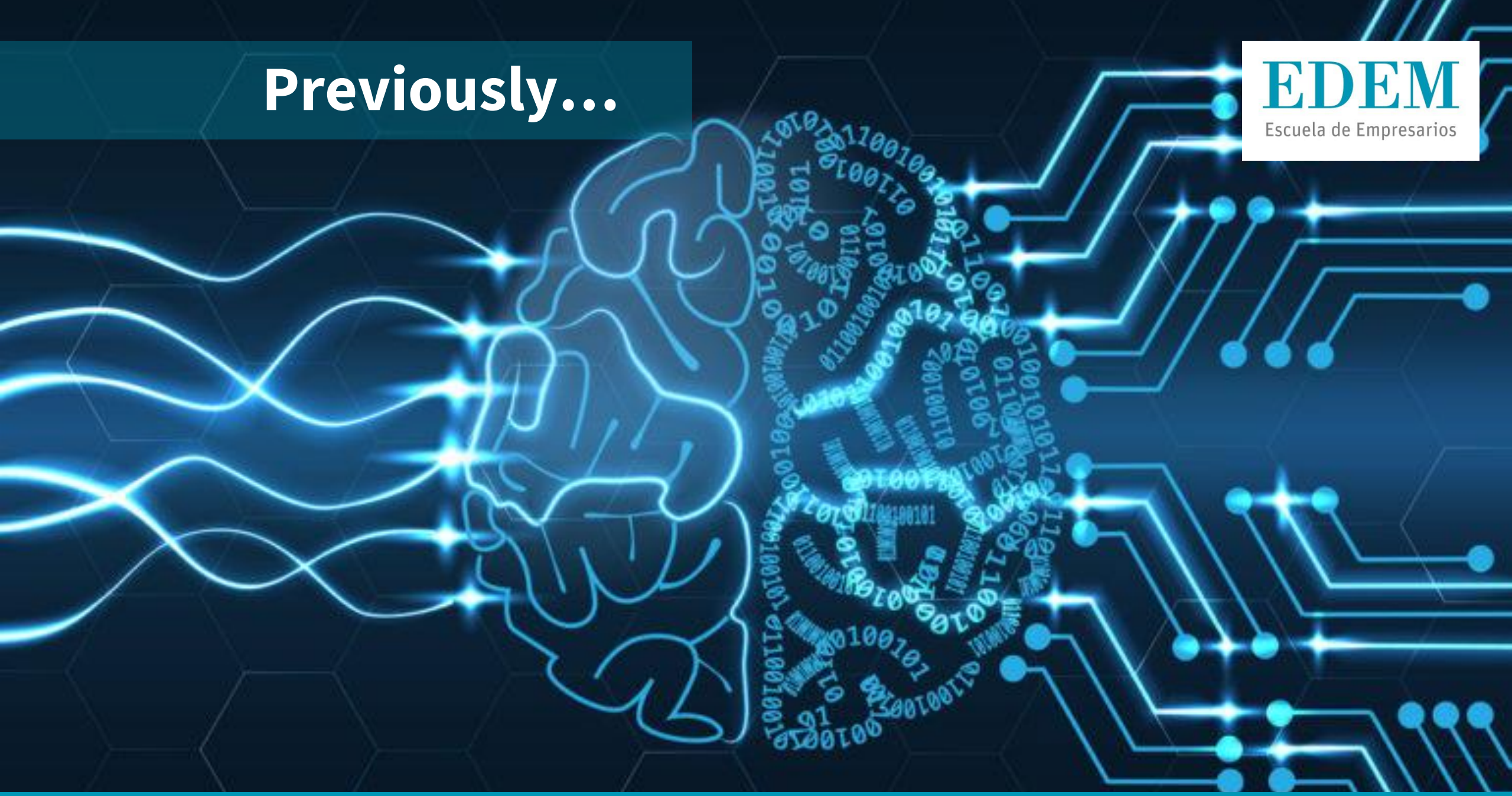
horus·ml

Jesús Prada Alonso - HORUS ML

Curso Máster en Data Analytics - Edición 4

Fecha 30/03/2023

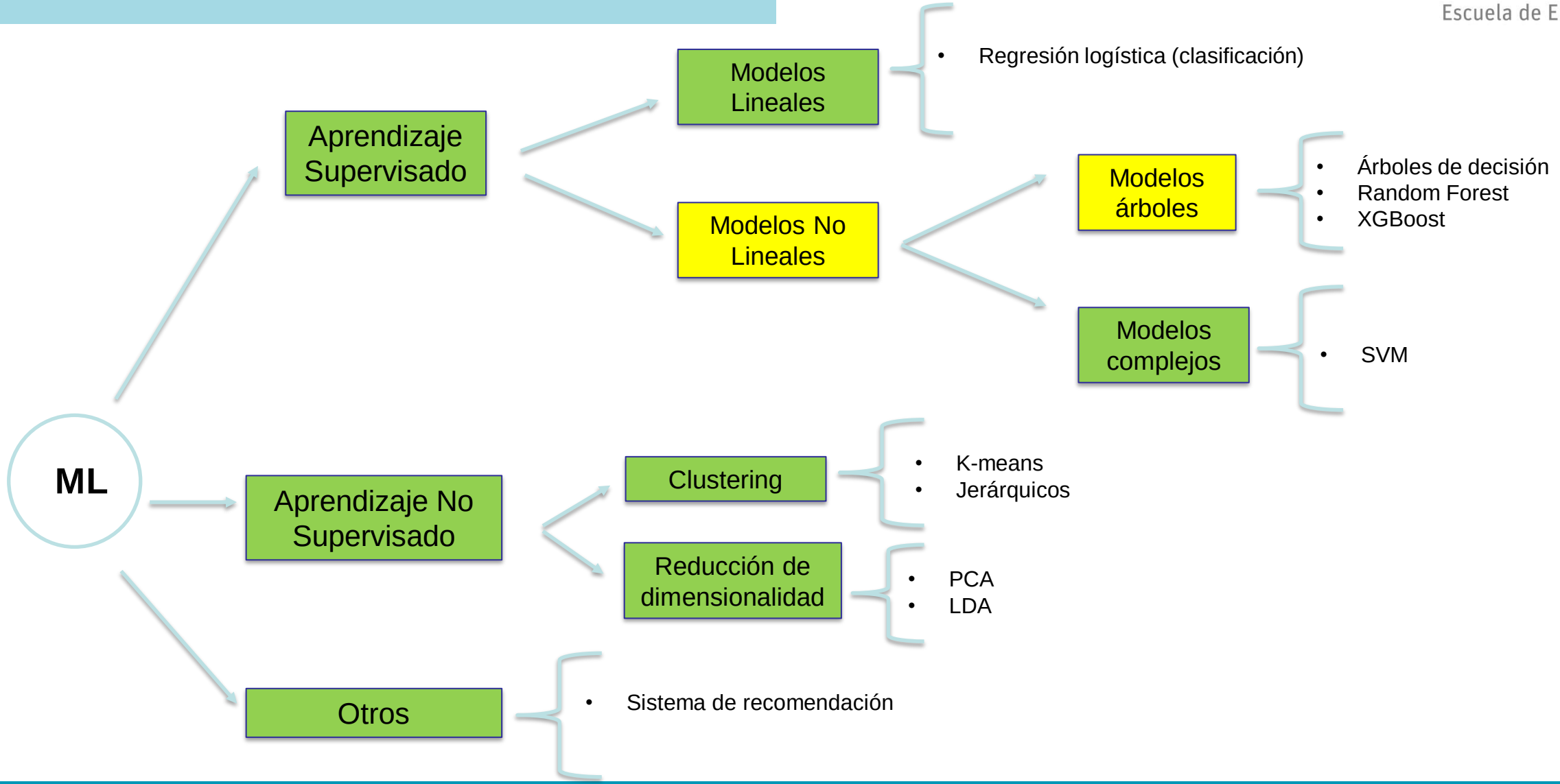
Previously...



EDEM
Escuela de Empresarios

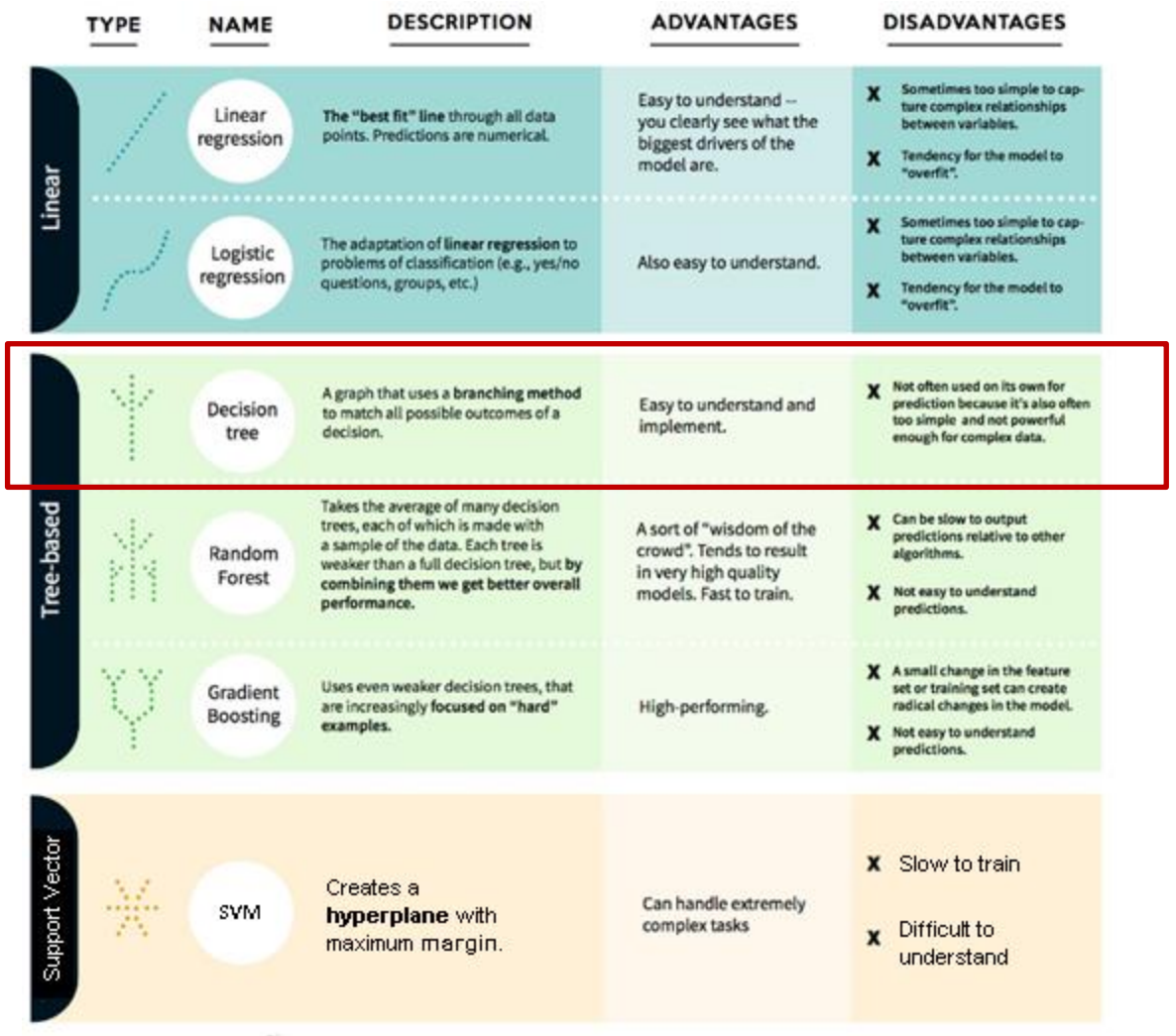
RESUMEN MODELOS

EDEM
Escuela de Empresarios



COMPARACIÓN ALGORITMOS

EDEM
Escuela de Empresarios



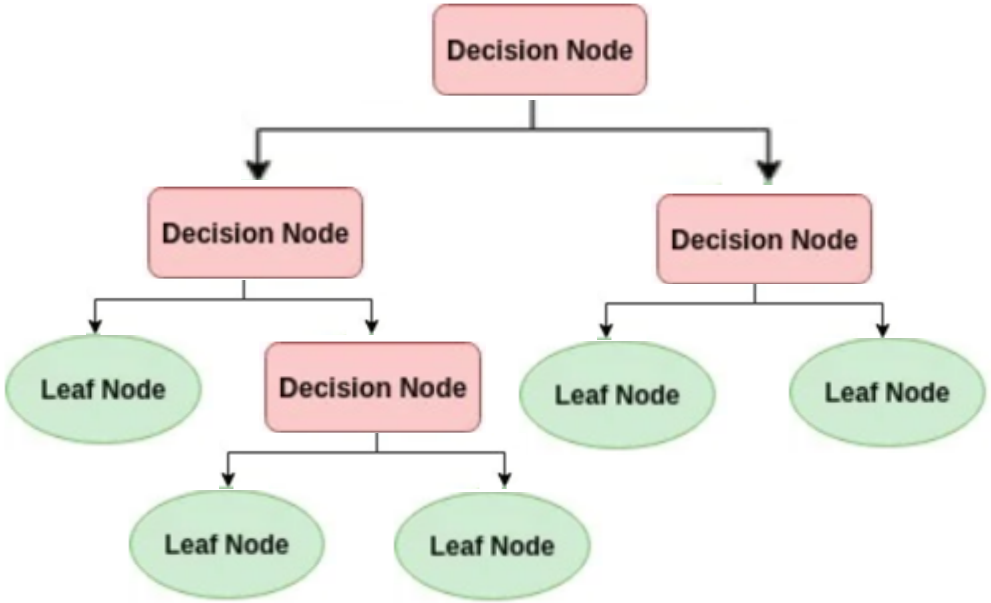
TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Linear	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Tree-based	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
Tree-based	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Support Vector	SVM	Creates a **hyperplane** with maximum margin.	Can handle extremely complex tasks	X Slow to train X Difficult to understand

ÁRBOLES DE DECISIÓN

EDEM
Escuela de Empresarios

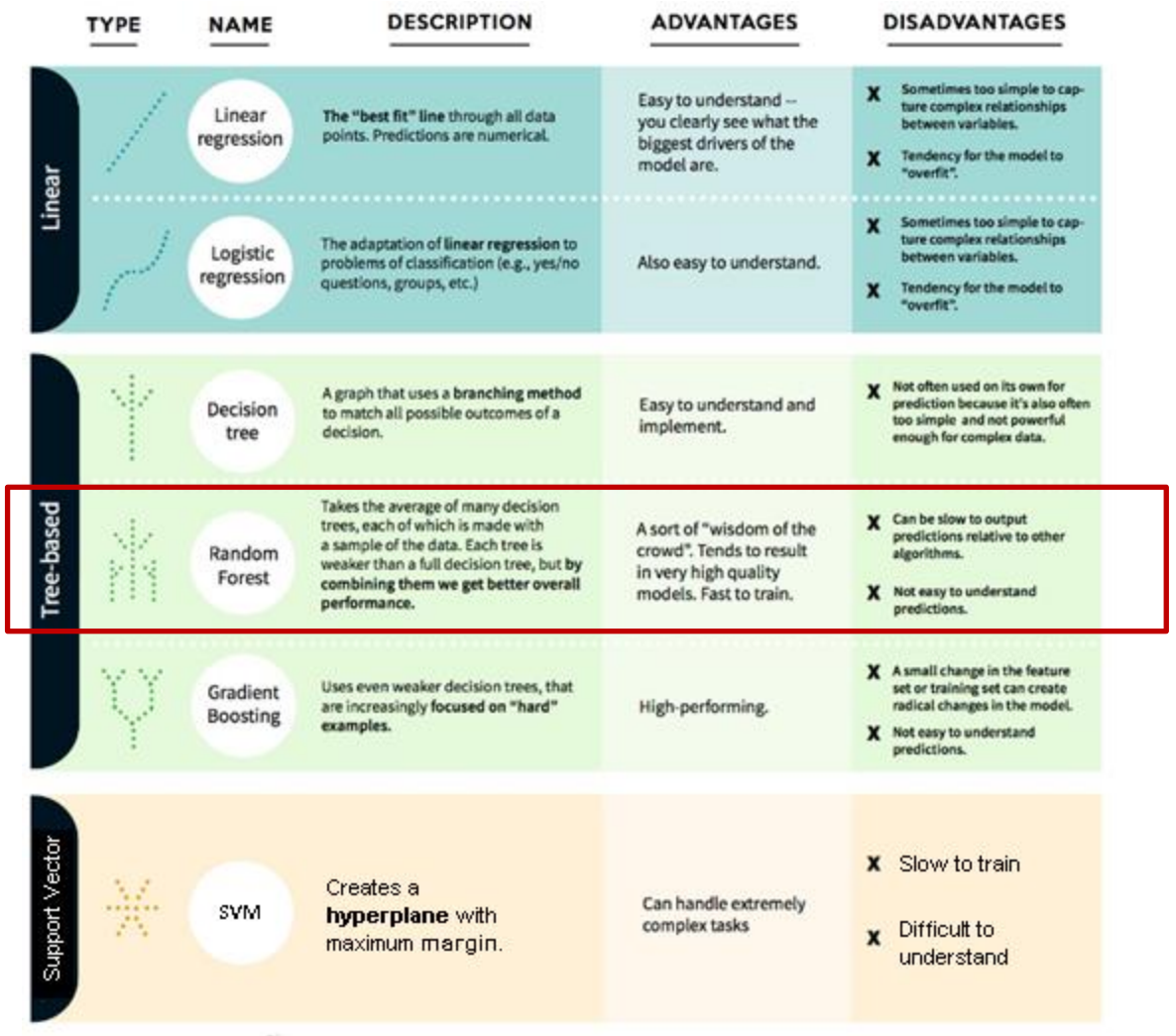
Definición (II)

- Cada nodo representa las variables de los datos, cada rama una decisión y cada nodo hoja un output.
- El nodo superior es el nodo raíz y representa la variable más importante.
- El árbol va aprendiendo de la división de los nodos recursivamente.
- Cuando un nuevo dato debe ser clasificado recorre el árbol desde el nodo raíz al nodo hoja, respondiendo a las preguntas de cada nodo en función de los atributos y el camino correcto en cada respuesta.



COMPARACIÓN ALGORITMOS

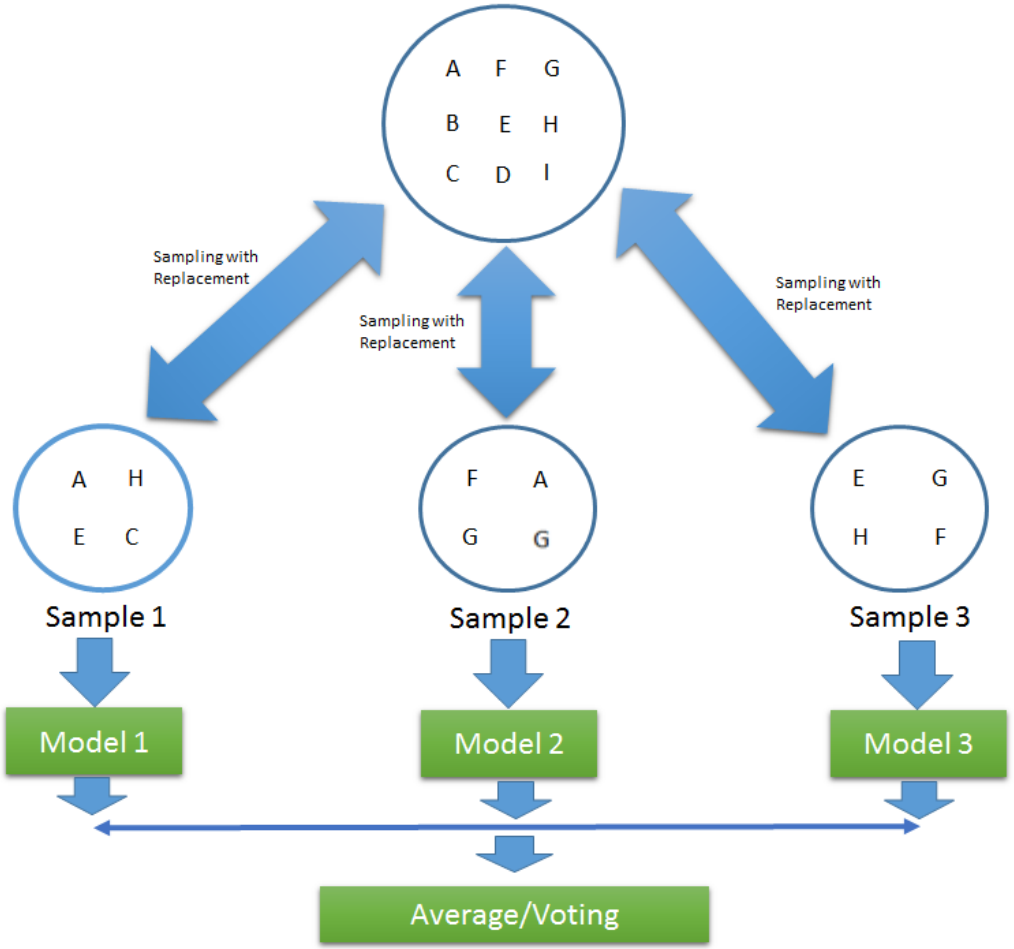
EDEM
Escuela de Empresarios



TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Linear	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Tree-based	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
Tree-based	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Support Vector	SVM	Creates a **hyperplane** with maximum margin.	Can handle extremely complex tasks	X Slow to train X Difficult to understand

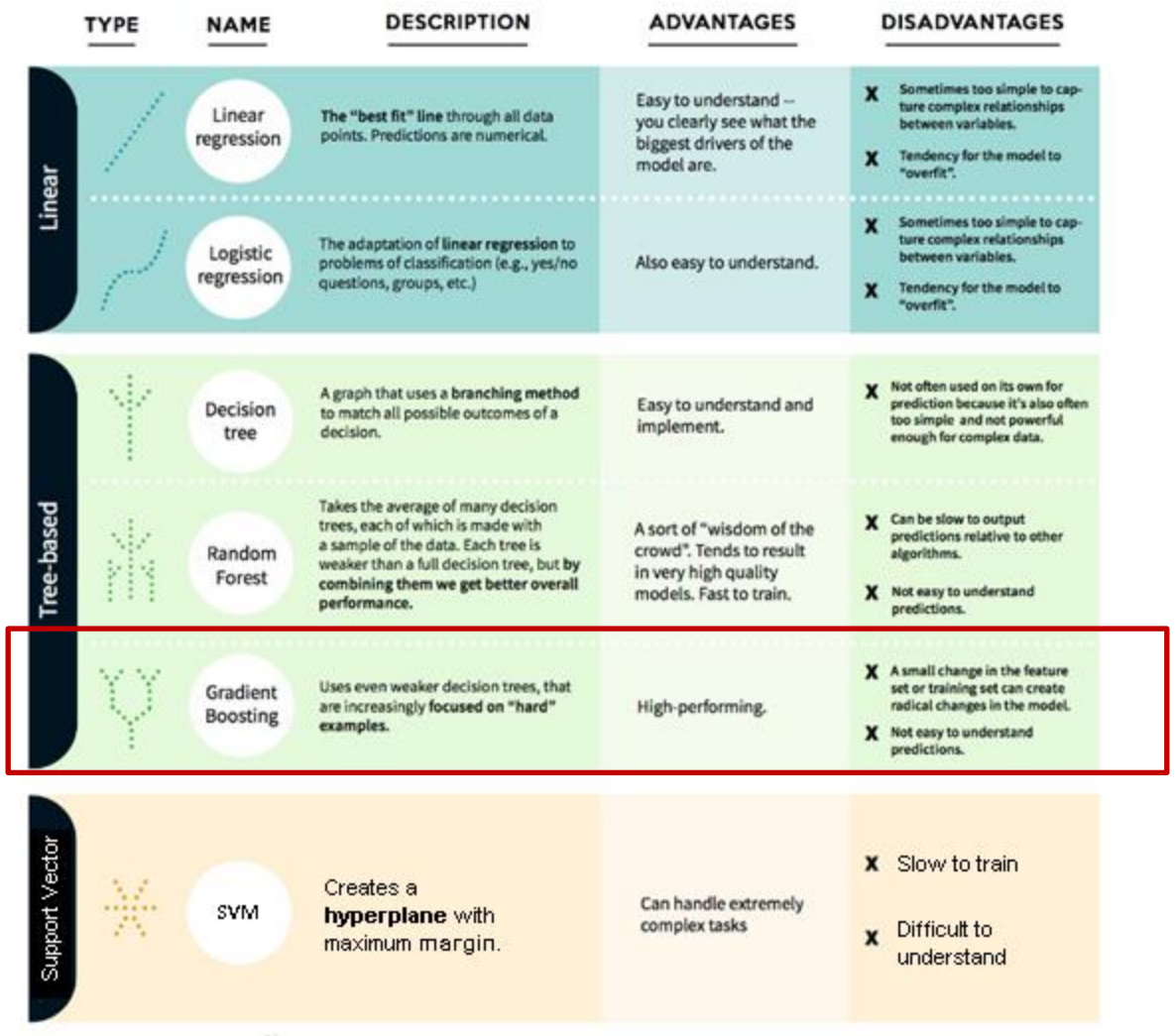
EDEM
Escuela de Empresarios

Bagging. Esquema



COMPARACIÓN ALGORITMOS

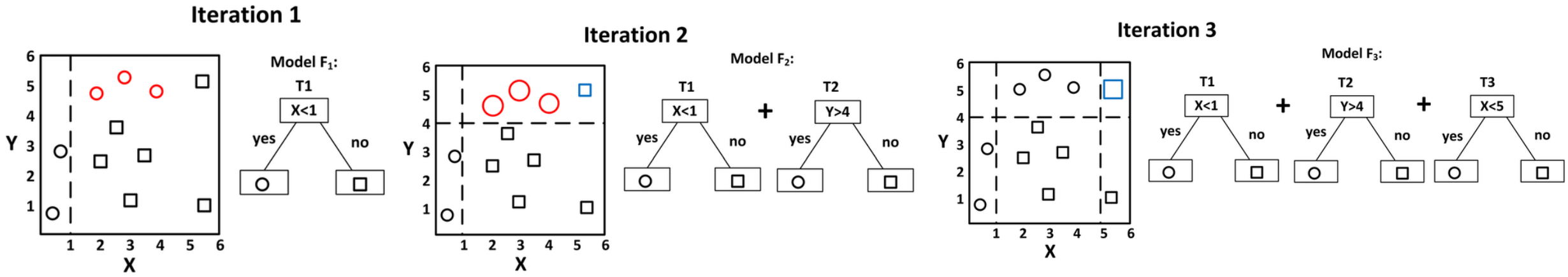
EDEM
Escuela de Empresarios



TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Linear	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Tree-based	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
Tree-based	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Support Vector	SVM	Creates a **hyperplane** with maximum margin.	Can handle extremely complex tasks	X Slow to train X Difficult to understand

EDEM
Escuela de Empresarios

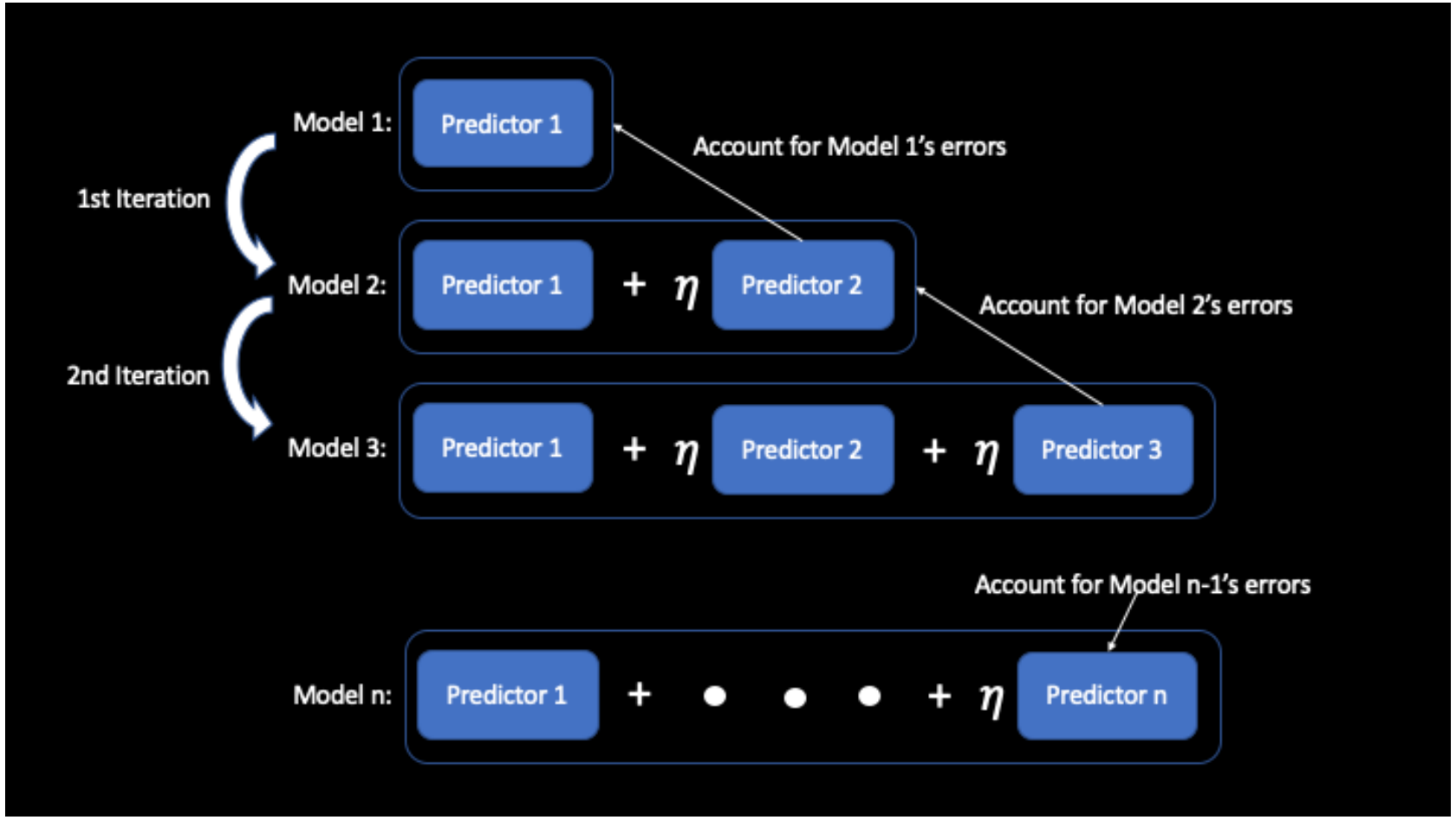
Boosting. Esquema



DEFINICIÓN (V)

EDEM
Escuela de Empresarios

Gradient Boosting. Representación visual



HIPERPARÁMETROS (XI)

EDEM
Escuela de Empresarios

Resumen

Hiperparámetro	Efecto	Rango	Valor por defecto
nrounds	Número de iteraciones de boosting a realizar	[1,∞]	100
eta	Inversamente relacionado con el step size.	[0,1]	0,3
gamma	Mínima mejora en la loss function requerida.	[0,∞]	0
max_depth	Profundidad máxima de cada árbol.	[1,∞]	6
min_child_weight	Suma mínima de pesos necesaria en un nodo hijo.	[0,∞]	1
subsample	Ratio de submuestreo de las instancias de entrenamiento.	(0,1]	1
colsample_bytree	Ratio de submuestreo de las columnas del dataset.	(0,1]	1
num_parallel_tree	Número de árboles calculados en cada iteración.	[1,∞]	1
lambda	Controla la regularización L2.	[0,∞]	1
alpha	Controla la regularización L1.	[0,∞]	0
early_stopping_rounds	Numero de iteraciones sin mejora permitidas.	[1,∞]	None

RESUMEN

EDEM
Escuela de Empresarios

Ventajas

- Una de las familias de modelos de mayor **potencial predictivo**.
- Cada **modelo individual** es muy **poco costoso** computacionalmente.
- **Poco** gasto de **memoria**.
- El elevado número de hiperparámetros permite su **ajuste a problemas muy variados**.
- La implementación oficial está **paralelizada** internamente.
- Permite reducir el preprocesado (scaling, valores no informados, variables categóricas).

Desventajas

- Modelo complejo → **Poco interpretable de forma directa**.
- **Muchos hiperparámetros** a probar, por lo que las búsquedas en rejilla pueden ser extensas y costosas.
- **Sensible** a la elección de **hiperparámetros**.

VARIACIONES DE XGBOOST

EDEM
Escuela de Empresarios

EDEM
Escuela de Empresarios

LightGBM

- LightGBM es otro framework de **gradient boosting** que utiliza modelos basados en **árboles**.
- En **XGBoost**, los árboles crecen en **profundidad**, mientras que en **LightGBM** lo hacen en hojas (**anchura**).
- A nivel de **ajuste predictivo**, está en una escala **similar a XGBoost**.
- Ventajas:
  - **Mayor velocidad** de entrenamiento.
  - **Menor** uso de **memoria**.
- Desventajas:
  - No integrado en framework **scikit-learn**.
  - **Menos directo de** instalar y **usar.**
  - Desarrollo **menos estable.**
  - **Menos documentación.**

EDEM
Escuela de Empresarios

CatBoost

- CatBoost también es un framework de **gradient boosting** que utiliza modelos basados en **árboles**.
- **Optimizado para** datasets con variables categóricas con **muchas categorías**.
- Su nivel de **ajuste predictivo** es **normalmente peor** que en **XGBoost salvo en el caso anterior**.
- Ventajas:
  - **Buenos resultados** empleando los **hiperparámetros por defecto**.
  - **Mejores predicciones** si tenemos **muchas categorías**.
- Desventajas:
  - Desarrollo **menos estable.**
  - **Menos documentación.**
  - **Normalmente peor**es resultados **que** en **XGBoost en muchas situaciones.**

EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

Jesús Prada Alonso
jesus.prada@horusml.com