How to build Ensemble Models with Machine Learning

PATRICIA RODRÍGUEZ VAQUERO
@patri_vaquero_
patricia.vaquero@stratebi.com

WHO WE ARE

- Stratebi Business Solutions is a Spanish company based in Madrid, with offices in Barcelona, Alicante and Seville, created by a group of professionals with extensive experience in information systems, technological solutions and processes related to Open Source and Business intelligence solutions .
- This experience, acquired during participation in strategic projects in internationally recognized companies, has been made available to our clients.

"WE DEDICATE ALL OUR EXPERIENCE TO THE CREATION OF INTELLIGENT SYSTEMS FOR THE MOST SUITABLE DECISION-MAKING"

WHO WE ARE

- We are mainly dedicated to the world of **Business Intelligence**, **Big Data** and **Machine Learning**. Stratebi has created the most important WebLog in Spanish on Business Intelligence, Data Warehouse, CRM, Dashboards, Scorecard and Open Source technology (**TODOBI**).

"THE TIME HAS COME WHEN THE INFORMATION AVAILABLE IN YOUR COMPANY BECOMES A REAL ASSET CAPABLE OF GENERATING BUSINESS."

- STRATEBI TEAM -

WHO WE ARE

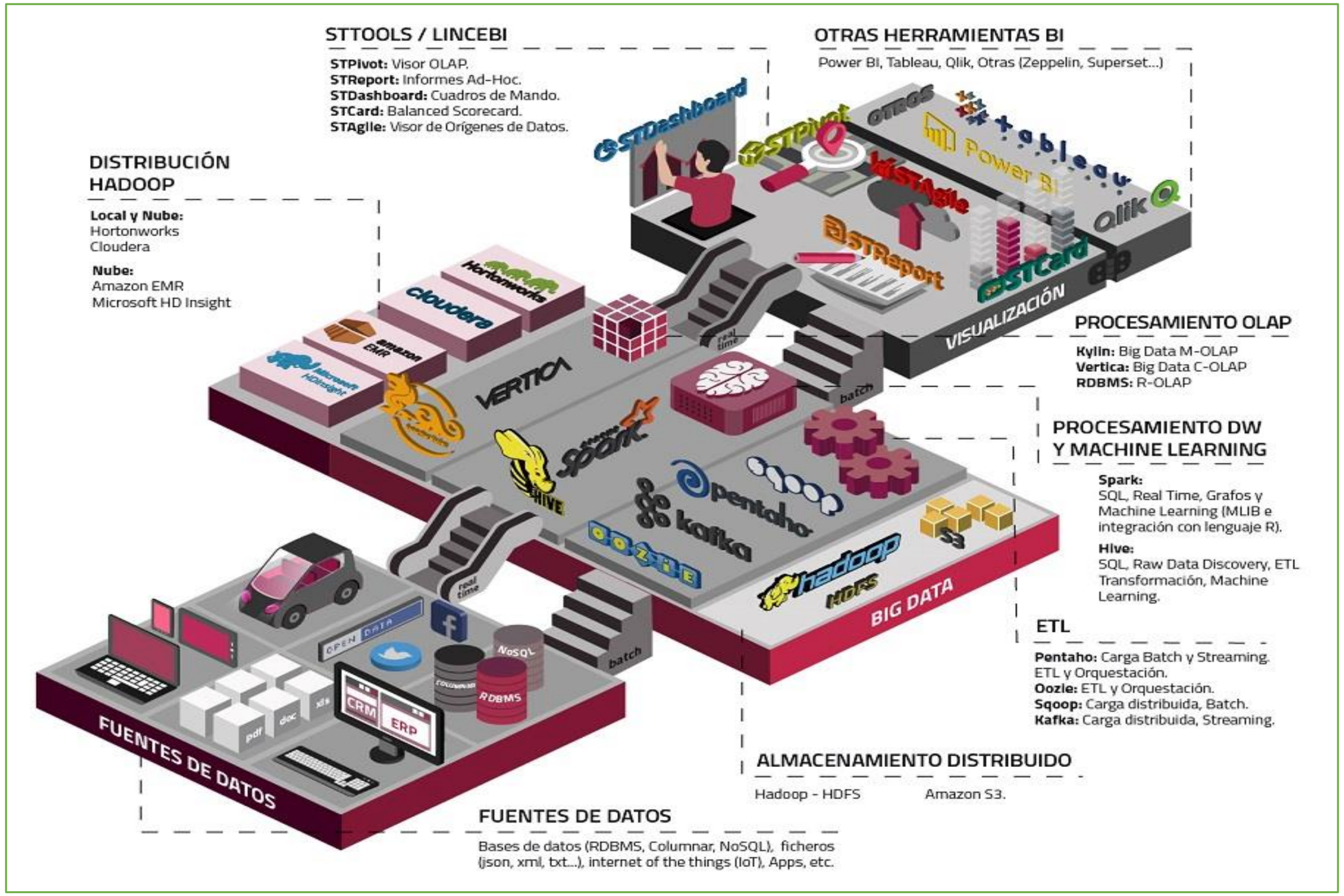
- Stratebi is the only Spanish company that has been present at all the Pentaho Developers held in Europe, having organized the one in Spain.
- At Stratebi we set ourselves the objective of providing companies and institutions with scalable tools adapted to their needs, which form a Business Intelligence strategy capable of making the available information profitable. For this, we are based on the development of Business Intelligence solutions, using Open Source technology.

OUR PARTNERS

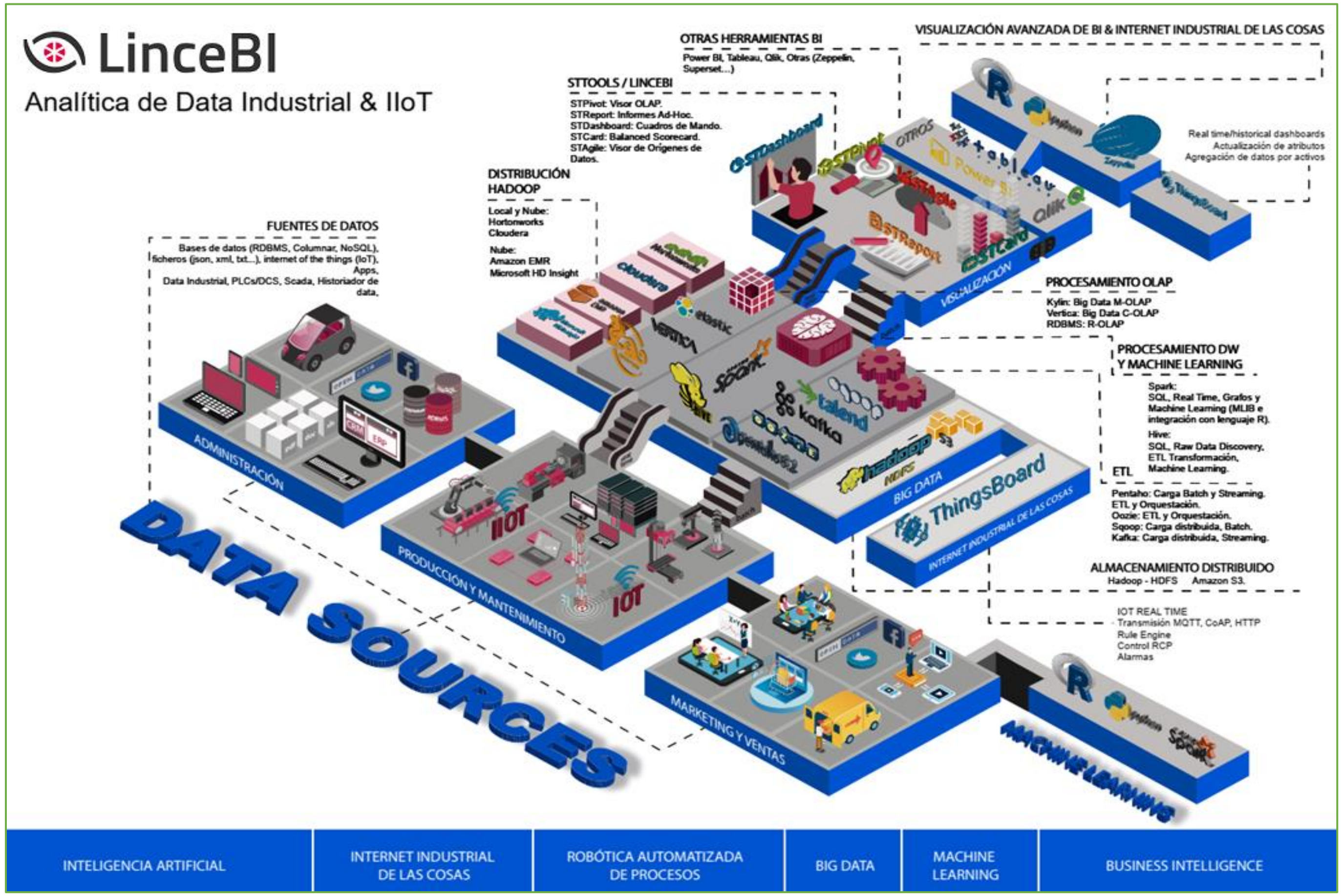
- At Stratebi we always bet on the best technologies.
- We are **Microsoft PowerBI** Certified Partners with extensive experience.
- We continually seek the best partners, both in technological and commercial areas, who can complement our portfolio of proprietary solutions. We have recently been named Certified Partners of **Vertica**, **Talend**, **Microsoft**, **Snowflake**, **Kylligence**, **Pentaho**, etc.

OUR PARTNERS

OUR TECHNOLOGIES



LINCEBI



LinceBI
Analítica de Data Industrial & IIoT
OTRAS HERRAMIENTAS BI
Power BI, Tableau, Qlik, Otras (Zeppelin, Superset...)
VISUALIZACIÓN AVANZADA DE BI & INTERNET INDUSTRIAL DE LAS COSAS
Real time/historical dashboards
Actualización de atributos
Agregación de datos por activos
STTOOLS / LINCEBI
STPivot: Visor OLAP.
STReport: Informes Ad-Hoc.
STDashboard: Cuadros de Mando.
STCard: Balanced Scorecard.
STAgile: Visor de Orígenes de Datos.
DISTRIBUCIÓN HADOOP
Local y Nube:
Hortonworks
Cloudera
Nube:
Amazon EMR
Microsoft HD Insight
FUENTES DE DATOS
Bases de datos (RDBMS, Columnar, NoSQL), ficheros (json, xml, txt...), internet of the things (IoT), Apps,
Data Industrial, PLCs/DCS, Scada, Historiador de data,
PROCESAMIENTO OLAP
Kylin: Big Data M-OLAP
Vertica: Big Data C-OLAP
RDBMS: R-OLAP
PROCESAMIENTO DW Y MACHINE LEARNING
Spark:
SQL, Real Time, Grafos y Machine Learning (MLIB e integración con lenguaje R).
Hive:
SQL, Raw Data Discovery, ETL Transformación, Machine Learning.
ETL
Pentaho: Carga Batch y Streaming. ETL y Orquestación.
Oozie: ETL y Orquestación.
Sqoop: Carga distribuida, Batch.
Kafka: Carga distribuida, Streaming.
ALMACENAMIENTO DISTRIBUIDO
Hadoop - HDFS Amazon S3.
IOT REAL TIME
Transmisión MQTT, CoAP, HTTP
Rule Engine
Control RCP
Alarmas
VISUALIZACIÓN
BIG DATA
ThingsBoard
INTERNET INDUSTRIAL DE LAS COSAS
ADMINISTRACIÓN
PRODUCCIÓN Y MANTENIMIENTO
MARKETING Y VENTAS
DATA SOURCES
MACHINE LEARNING
INTELIGENCIA ARTIFICIAL
INTERNET INDUSTRIAL DE LAS COSAS
ROBÓTICA AUTOMATIZADA DE PROCESOS
BIG DATA
MACHINE LEARNING
BUSINESS INTELLIGENCE

OUR CLIENTS

Private Sector

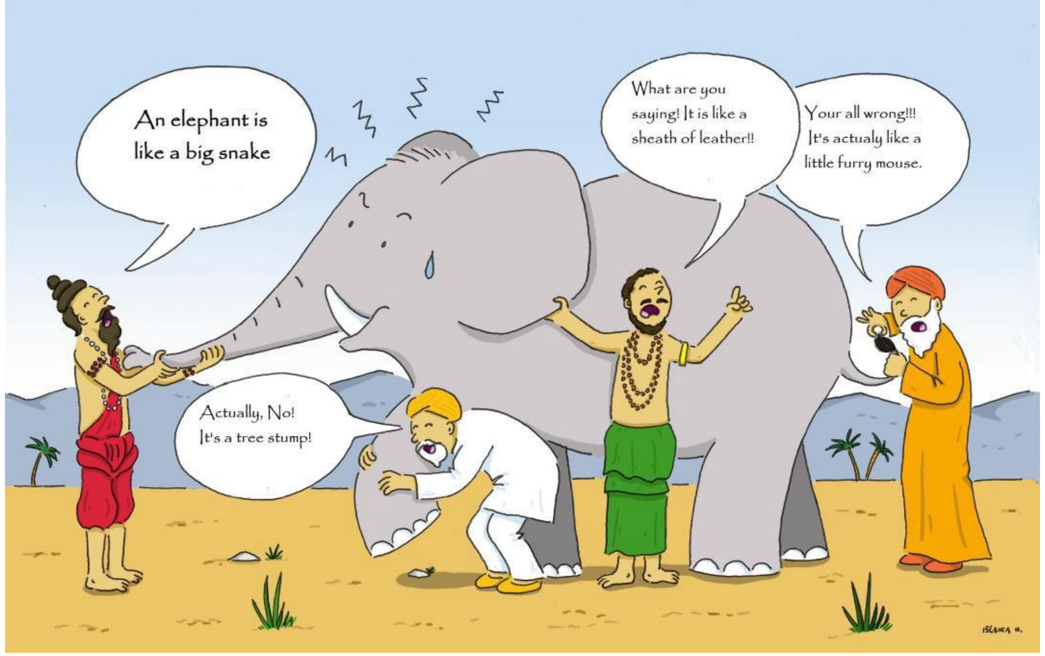
Public Sector

EXAMPLES OF ANALYTICS DEVELOPMENTS

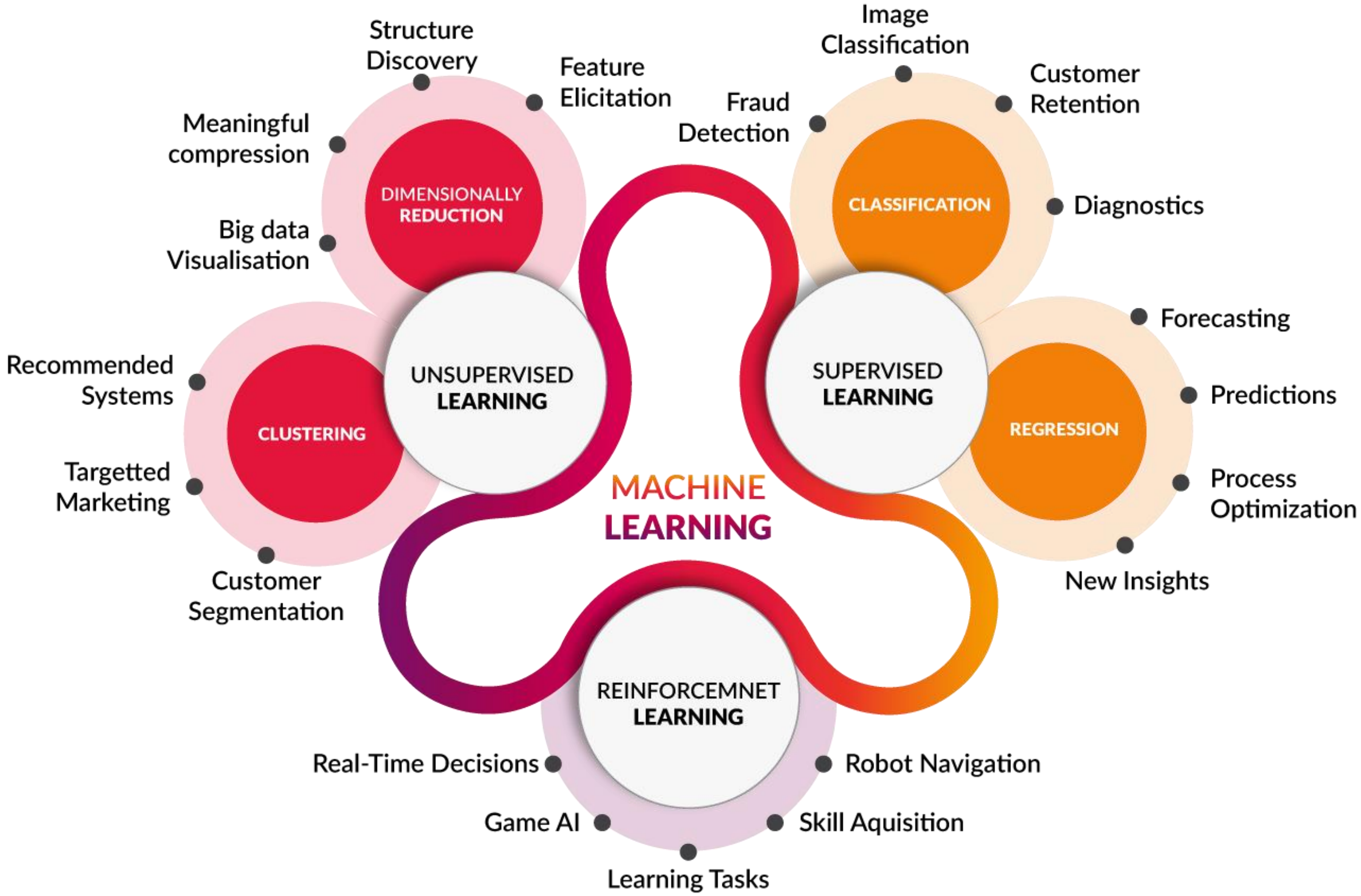
INDEX

1. INTRODUCTION

- Ensemble methods are the data science version of the old saying about two heads being better than one: if one model works well, multiple models working together can do even better.



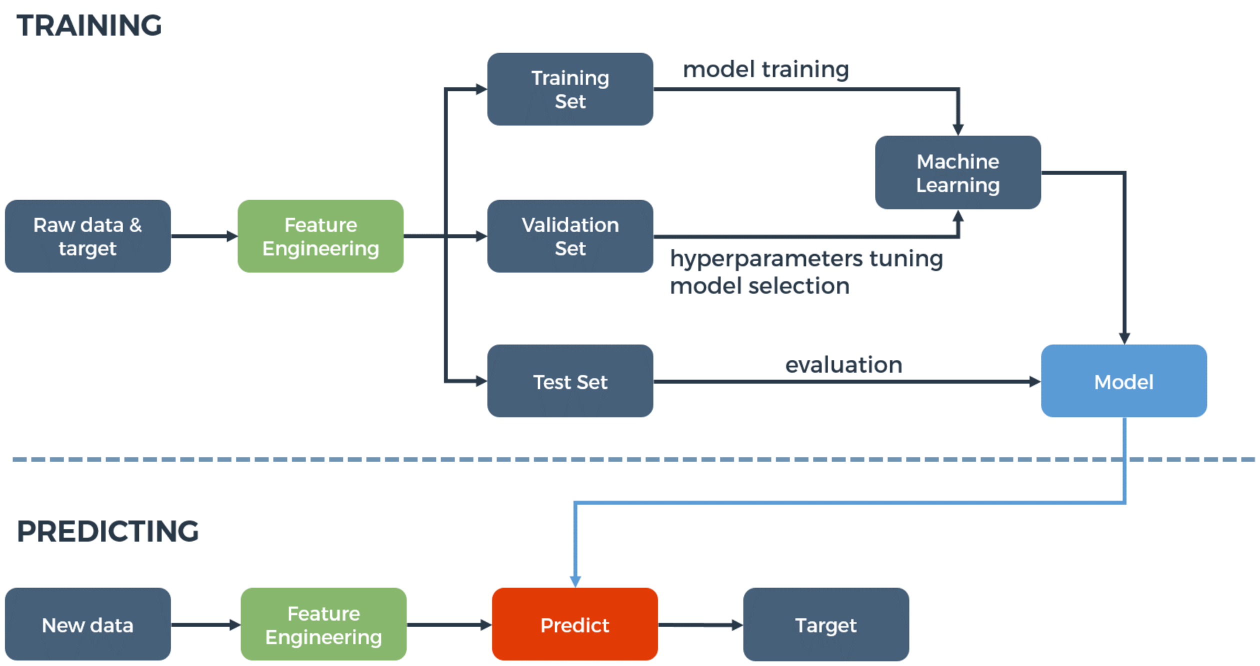
2. MACHINE LEARNING



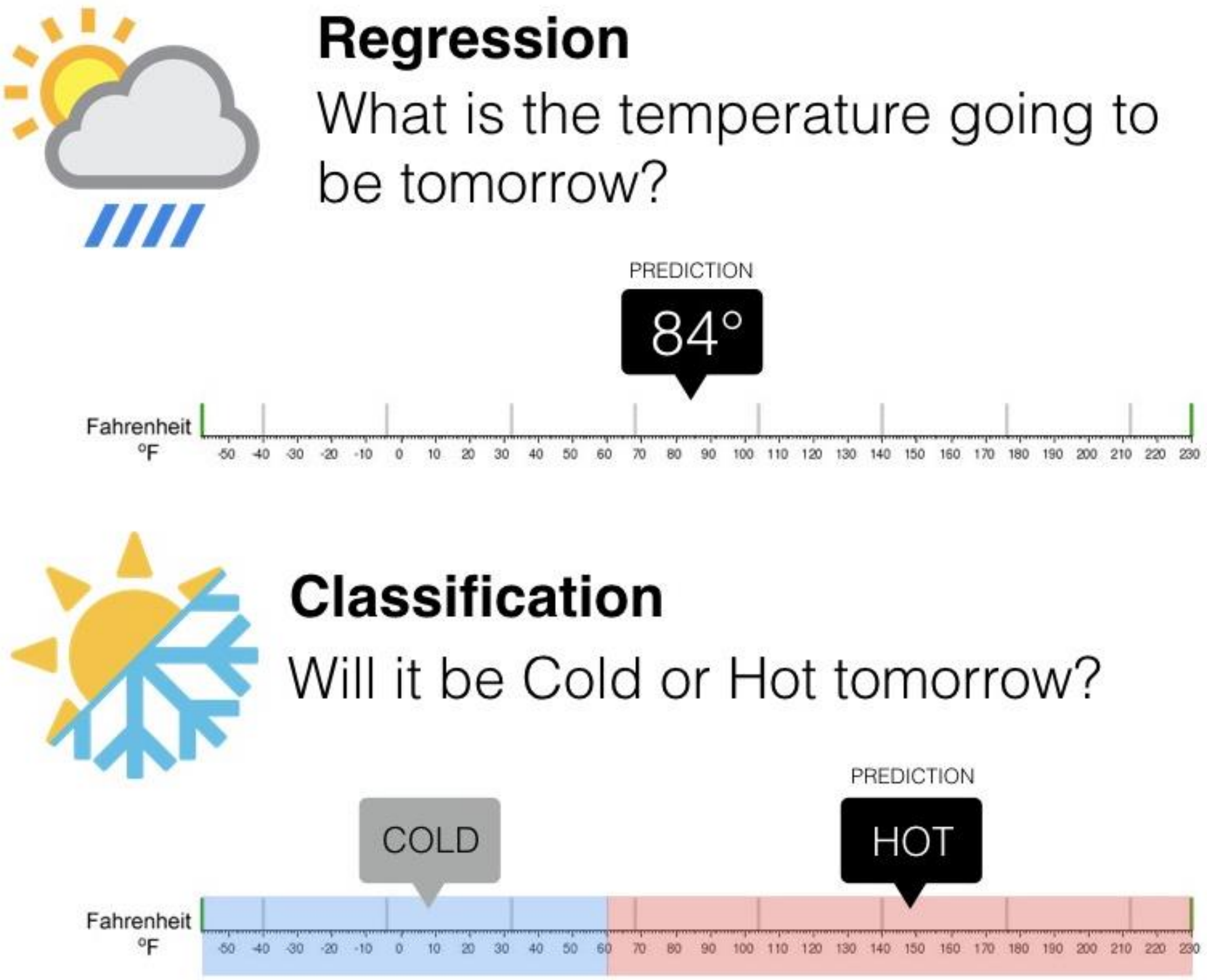
2. MACHINE LEARNING

- There is not a “one and only one" way to solve a problem in the Machine Learning world.
- There are always more than one algorithm that fits, you have to choose which one fits better.

3. SUPERVISED MACHINE LEARNING



3. SUPERVISED MACHINE LEARNING



Classification and *Regression* share the common concept of using past data to make predictions, or take decisions, that's where their similarity ends.

3. SUPERVISED MACHINE LEARNING

Classification problems:

1. *Binary classification:* only two classes to predict, usually 1 or 0.
2. *Multi-Class Classification:* more than two class labels to predict.

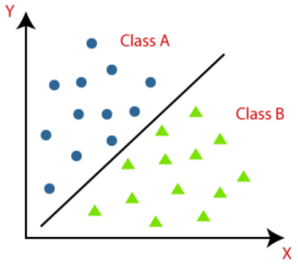
The objective is to predict a discrete number of values for a given input data. The labels (Y) is categorical and represents a finite number of classes.

Regression problems:

- The objective is to predict a continuous value for an input based on past data; the labels (Y) are numerical.

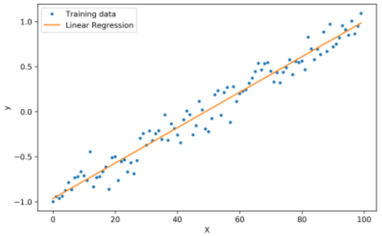
4. POPULAR ALGORITHMS

- *Classification:*



- Decision Trees.
- Logistics Regression.
- Naive Bayes.
- KNN.
- SVM.

- *Regression:*



- Regression Trees.
- Simple Regression.
- Multiple Regression.
- Ridge Regression.
- Lasso Regression.

5. ENSEMBLE METHODS

- They have impressive results in a number of machine learning competitions, where the winning solutions used ensemble methods, and also they have been successfully applied to diverse real-world tasks.
- Ensemble Methods are among the most powerful and easiest to use of predictive analytic algorithms and R/Python have an outstanding collection that includes the best performers.

5. ENSEMBLE METHODS

- Ensemble methods refer to a group of models working together to solve a common problem.
- Rather than depending on a single model for the best solution, ensemble learning utilizes the advantages of several different methods to counteract each model's individual weaknesses.
- The resulting quality & predictive performance will be higher than even the best individual algorithms.

REASONS TO USE ENSEMBLE METHODS

Ensemble helps to reduce Bias and Variance.

- *Reduce variance:* it will be less sensitive to specific training data, increasing the robustness of the model (averaging reduces variance, without increasing bias).
- *Reduce bias:* for simple models, average of models has much greater capacity than single model.

GENERAL IDEA

- Each model in the ensemble makes a prediction.
- A *final prediction* is determined by:
- *Averaging:* take the average of predictions, in case of a regression problem or while predicting probabilities for a classification problem.
- *Majority vote:* take the prediction with maximum votes, in case of predicting the outcomes of a classification problem.
- *Weighted average:* a different weight is assigned to each prediction, then take the average of them, which means giving high or low importance to specific model output.

METHODS FOR CONSTRUCTING ENSEMBLES

- *By manipulating the training set:*

Create multiple training sets by re-sampling the original data according to some sampling distribution or technique.

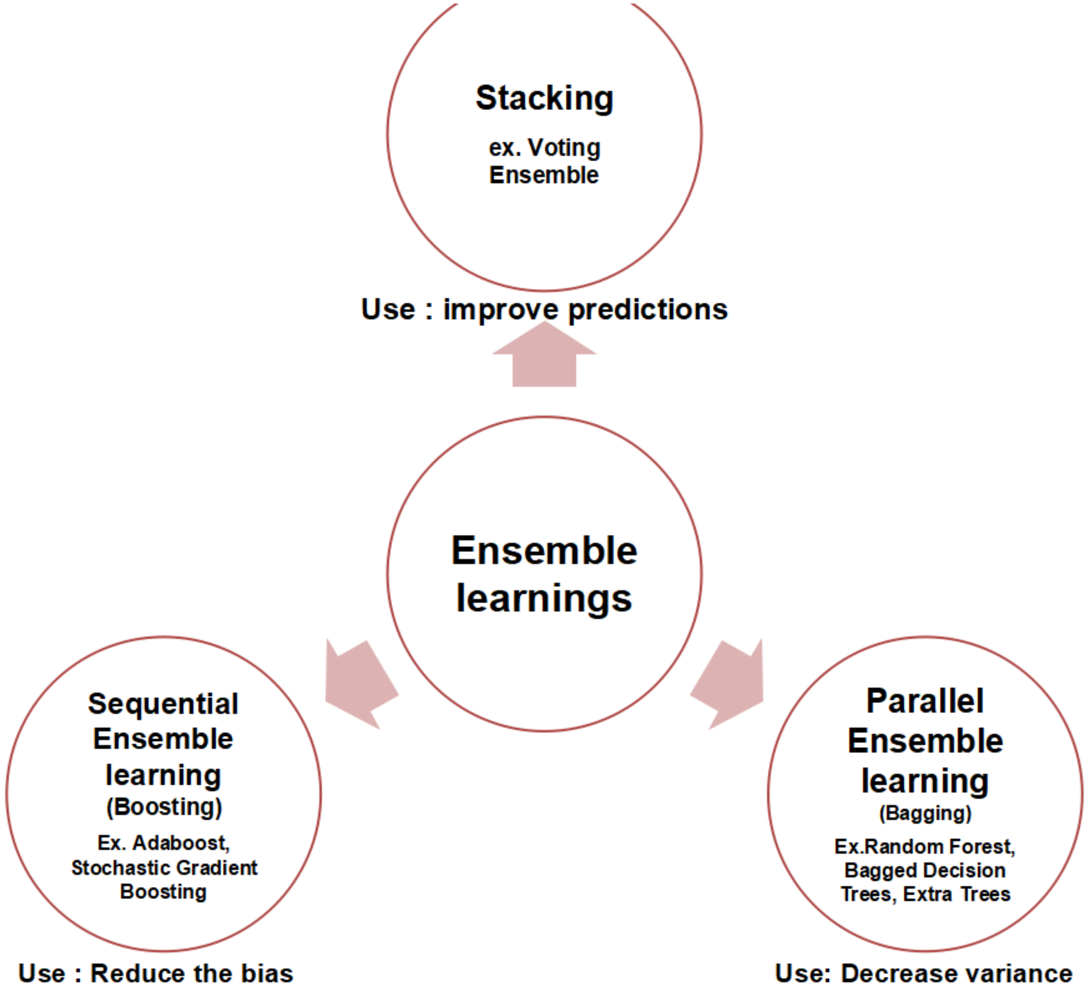
- *By manipulating the input features:*

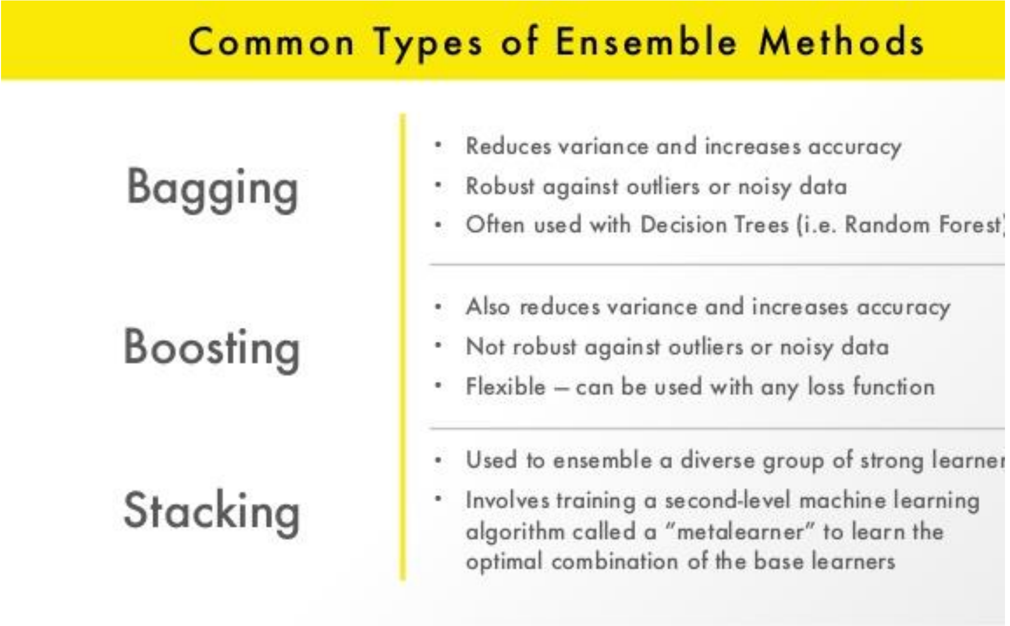
Choose a subset of input features to form each training set.

- *By manipulating the learning algorithm:*

Manipulate the learning algorithm to generate different models.

ENSEMBLE LEARNING TYPES





Common Types of Ensemble Methods

Bagging	• Reduces variance and increases accuracy • Robust against outliers or noisy data • Often used with Decision Trees (i.e. Random Forest)
Boosting	• Also reduces variance and increases accuracy • Not robust against outliers or noisy data • Flexible – can be used with any loss function
Stacking	• Used to ensemble a diverse group of strong learner • Involves training a second-level machine learning algorithm called a "metalearner" to learn the optimal combination of the base learners

ADVANTAGES AND DISADVANTAGES

- *Advantages:*

- Stable and more robust model.
- They often improve predictive performance.
- They're unlikely to overfit.

- *Disadvantages:*

- They suffer from lack of interpretability.
- Ensemble methods are usually computationally expensive.
- The selection of models for creating an ensemble is an art which is really hard to master.

6. EXAMPLES

7. RESULTS

BASELINE MODELS

	Baseline Model 1: Adaboost	Baseline Model 2: Gradient Boosting	Baseline Model 3: Random Forest
Parameters Used	DecisionTreeClassifier(max_depth=1) n_estimators=200	n_estimators=300 learning_rate=0.75 random_state=0	n_estimators=300 min_samples_leaf = 3 random_state=0
Training Accuracy Score	**0.7657564718398401**	**0.8175150714078468**	**0.8237693026854509**
Test Accuracy Score	**0.7678363232915868**	**0.7843469103764714**	**0.7948775809545301**
F1 Score	0.645982680506968	0.6987496059682673	0.703619171829149
Precision Score	0.7551622418879056	0.7344008834897846	0.7686097589932328

ITERATION I

	Gradient Boosting	Random Forest
Parameters Used	learning_rate=0.005 max_depth=7 max_features='sqrt' n_estimators=1750 random_state=10 subsample=1	max_features=None min_samples_leaf=12 min_samples_split=5 n_estimators=150
Training Accuracy Score	**0.8027338083110351**	**0.8048615364776427**
Test Accuracy Score	**0.7963819624656813**	**0.7978487344390538**
F1 Score	0.7073829856231759	0.7123668860705303
Precision Score	0.7678047635808988	0.7643546164446486
Parameter Tuning Time	181.2min (576 fits, 288 candidates, 2 folds)	22.3min (30 fits, 10 candidates, 3 folds)

ITERATION II

	XGBoost	CatBoost	Light GBM
Parameters Used	learning_rate=0.05 max_depth=10 min_child_weight=3 n_estimators=200	depth=10 iterations= 300 l2_leaf_reg= 9 learning_rate= 0.03	learning_rate=0.01 max_depth=25 n_estimators=200 num_leaves=300 silent=False
Training Accuracy Score	**0.8231567748799123**	**0.8023630677971566**	**<u>0.8061833070053838</u>**
Test Accuracy Score	**0.800368573470232**	**0.7994283350257625**	**<u>0.801572078679153</u>**
F1 Score	0.7163317657118426	0.7113083960374601	0.7165269718461208
Precision Score	0.7674338715218139	0.7734871674122911	0.7723850341712035
Parameter Tuning Time (81 fits, 27 candidates, 3 folds)	122.2min	38.7min	12.6min

Contact us:

Madrid: Avda. del Brasil, 17.
Barcelona: C/ Valencia, 63.
Alicante: C/Italia 23.
Sevilla: : Estadio Olímpico, PT La Cartuja

91 788 34 10
www.stratebi.com
info@stratebi.com

facebook.com/stratebiopenbi
@stratebi
linkedin.com/company/stratebi

stratebi
Analytics and Big Data Ninjas

TodoBI
< Business Intelligence >