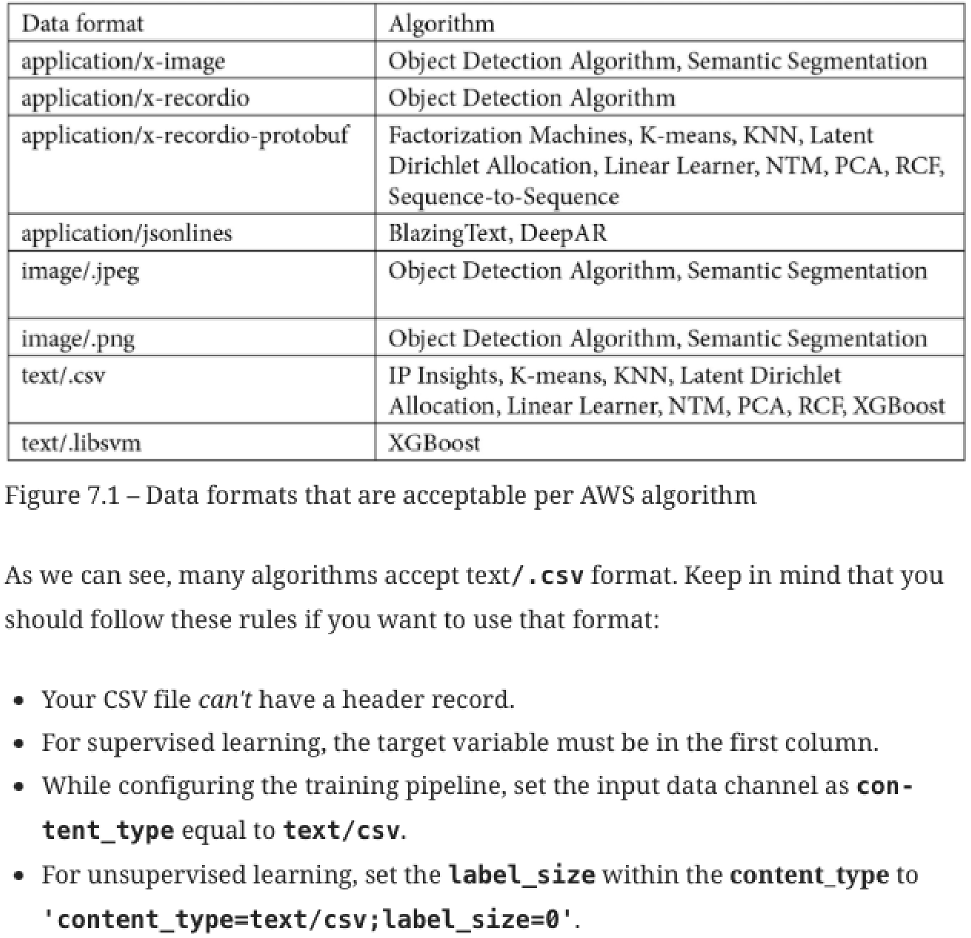
Tech Experts

Semana 6

Storing the training data



Data format	Algorithm
application/x-image	Object Detection Algorithm, Semantic Segmentation
application/x-recordio	Object Detection Algorithm
application/x-recordio-protobuf	Factorization Machines, K-means, KNN, Latent Dirichlet Allocation, Linear Learner, NTM, PCA, RCF, Sequence-to-Sequence
application/jsonlines	BlazingText, DeepAR
image/.jpeg	Object Detection Algorithm, Semantic Segmentation
image/.png	Object Detection Algorithm, Semantic Segmentation
text/.csv	IP Insights, K-means, KNN, Latent Dirichlet Allocation, Linear Learner, NTM, PCA, RCF, XGBoost
text/.libsvm	XGBoost

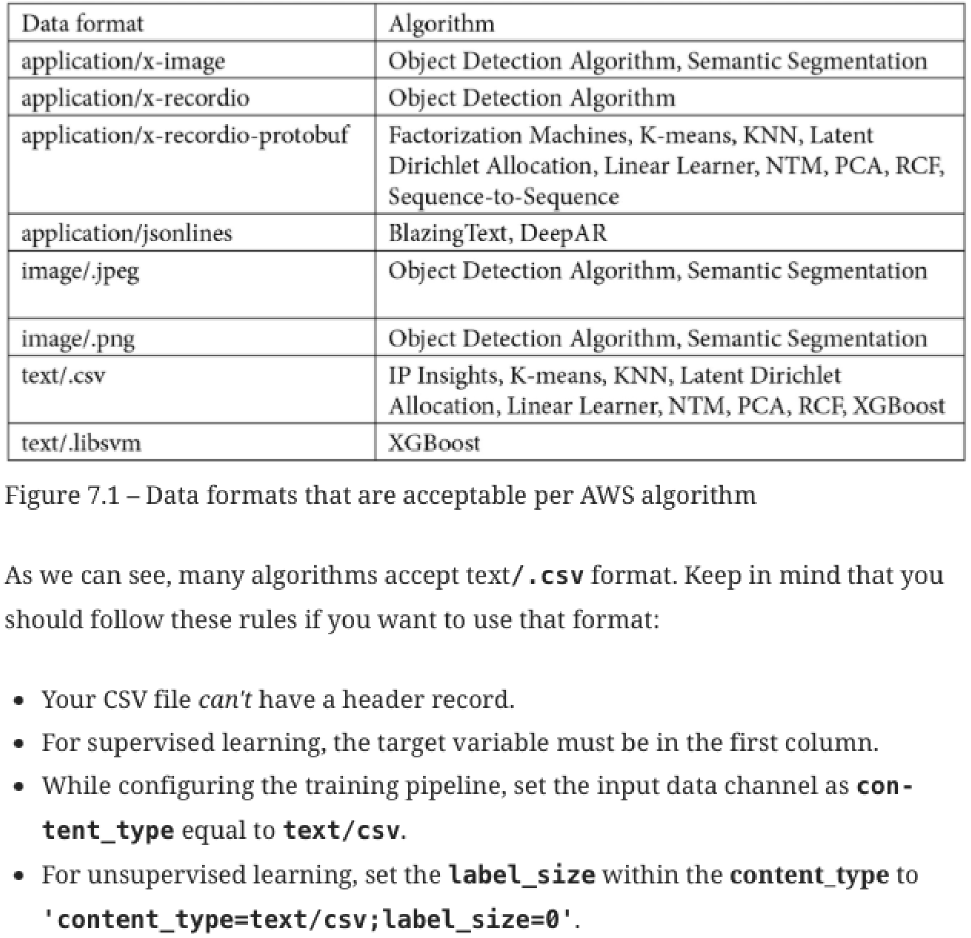
Figure 7.1 – Data formats that are acceptable per AWS algorithm

As we can see, many algorithms accept text/`.csv` format. Keep in mind that you should follow these rules if you want to use that format:

- Your CSV file *can't* have a header record.
- For supervised learning, the target variable must be in the first column.
- While configuring the training pipeline, set the input data channel as **`content_type`** equal to **`text/csv`**.
- For unsupervised learning, set the **`label_size`** within the **content_type** to **`'content_type=text/csv;label_size=0'`**.

- Mais importantes: csv e recordIO-protobuf
- recordIO-protobuf é otimizado e funciona melhor com algoritmos built-in aws
- RecordIO-protobud aceita dois tipos de input: pipe mode e file mode
- Pipe mode: os dados são enviados em stramming para a instancia de treino diretamente do S3, otimizando o armazenamento.
- File mode: Os dados são copiados para o volume da instancia de treinamento.

Tipos de prediçao



Data format	Algorithm
application/x-image	Object Detection Algorithm, Semantic Segmentation
application/x-recordio	Object Detection Algorithm
application/x-recordio-protobuf	Factorization Machines, K-means, KNN, Latent Dirichlet Allocation, Linear Learner, NTM, PCA, RCF, Sequence-to-Sequence
application/jsonlines	BlazingText, DeepAR
image/.jpeg	Object Detection Algorithm, Semantic Segmentation
image/.png	Object Detection Algorithm, Semantic Segmentation
text/.csv	IP Insights, K-means, KNN, Latent Dirichlet Allocation, Linear Learner, NTM, PCA, RCF, XGBoost
text/.libsvm	XGBoost

Figure 7.1 – Data formats that are acceptable per AWS algorithm

As we can see, many algorithms accept text/`.csv` format. Keep in mind that you should follow these rules if you want to use that format:

- Your CSV file *can't* have a header record.
- For supervised learning, the target variable must be in the first column.
- While configuring the training pipeline, set the input data channel as **`content_type`** equal to **`text/csv`**.
- For unsupervised learning, set the **`label_size`** within the **content_type** to **`'content_type=text/csv;label_size=0'`**.

- Mais importantes: csv e recordIO-protobuf
- recordIO-protobuf é otimizado e funciona melhor com algoritmos built-in aws
- RecordIO-protobud aceita dois tipos de input: pipe mode e file mode
- Pipe mode: os dados são enviados em stramming para a instancia de treino diretamente do S3, otimizando o armazenamento.
- File mode: Os dados são copiados para o volume da instancia de treinamento.

Tipos de algoritmos

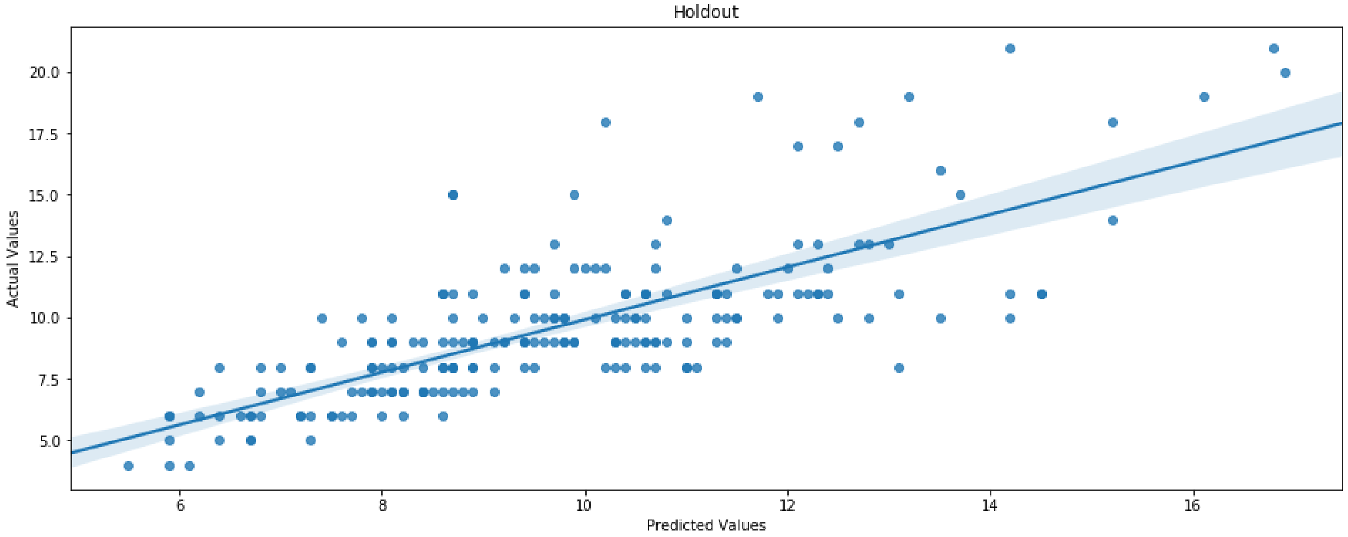
- Supervised learning
- Unsupervised learning
- Textual analysis
- Image processing

- Classificacao
- Regressão
- Previsao (forecasting)
- Object2Vec
- Clustering
- Reducao de dimensionalidade
- IP Insights
- Natual Language Processing

Supervised learning

- Linear learner algorithm
- XGBoost algorithm
- K-Nearest Neighbor algorithm
- Object2Vec algorithm
- DeepAR Forecasting algorithm

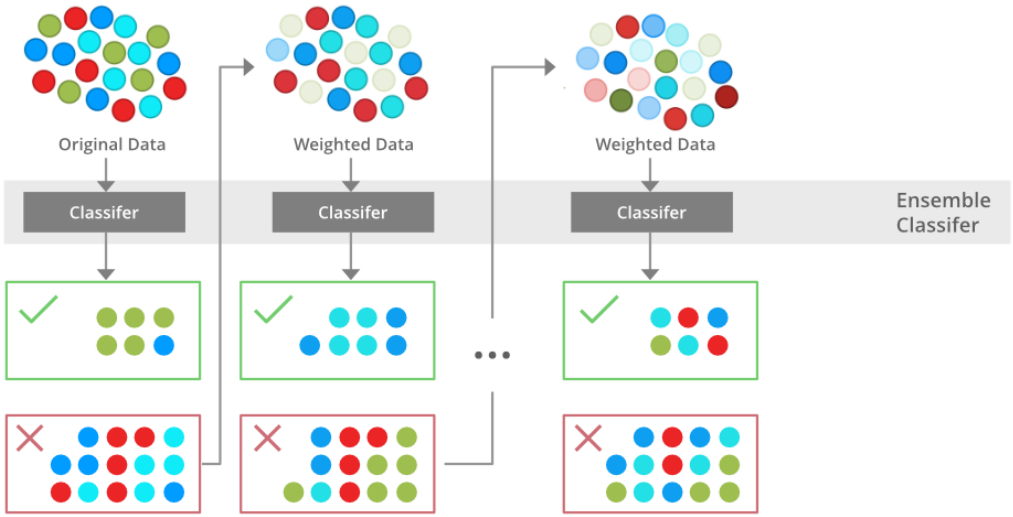
•Linear learner algorithm



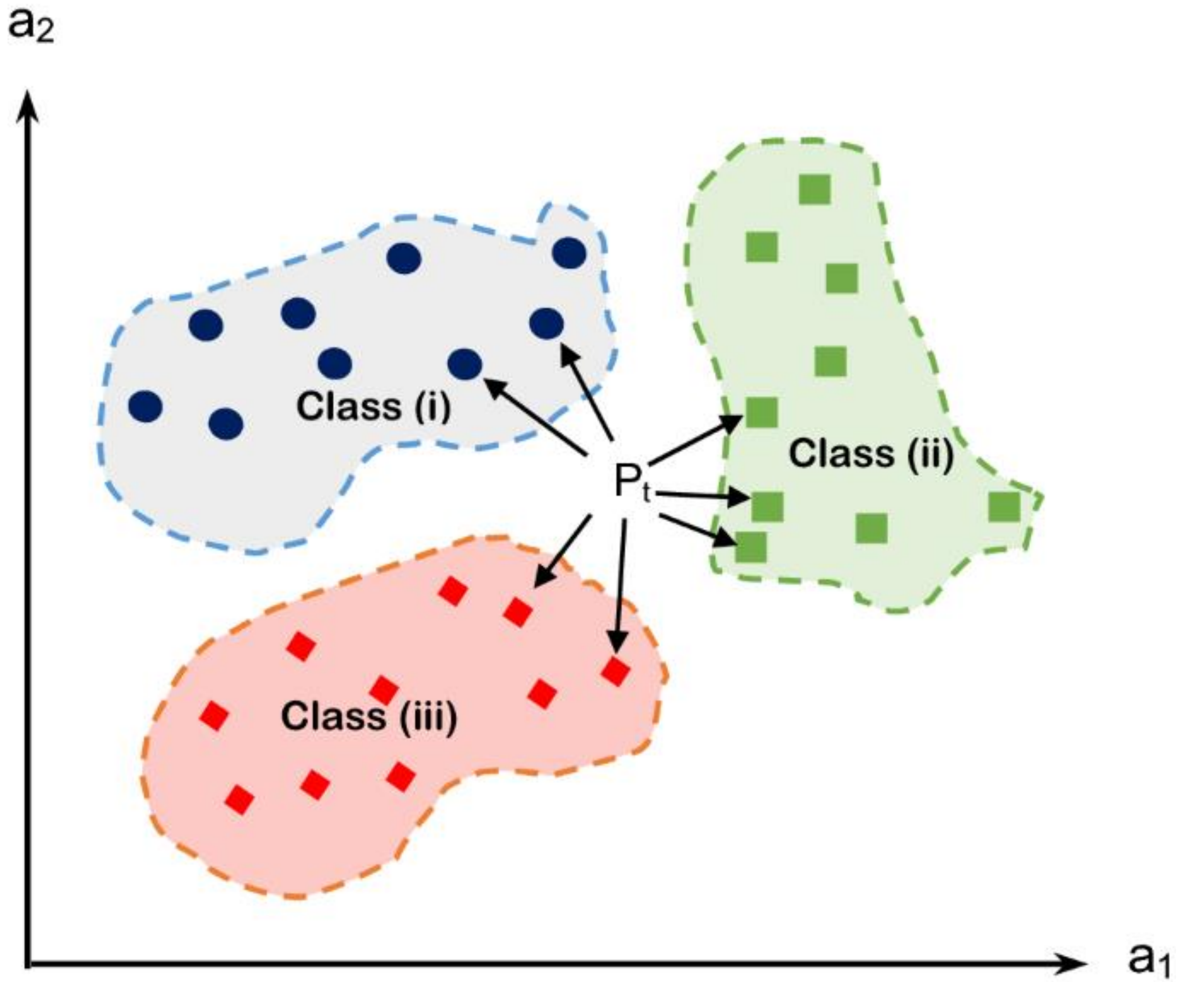
- Regressao
- Classificao, com nome **logistc regression**

•Xgboost

- Regressao e classificao

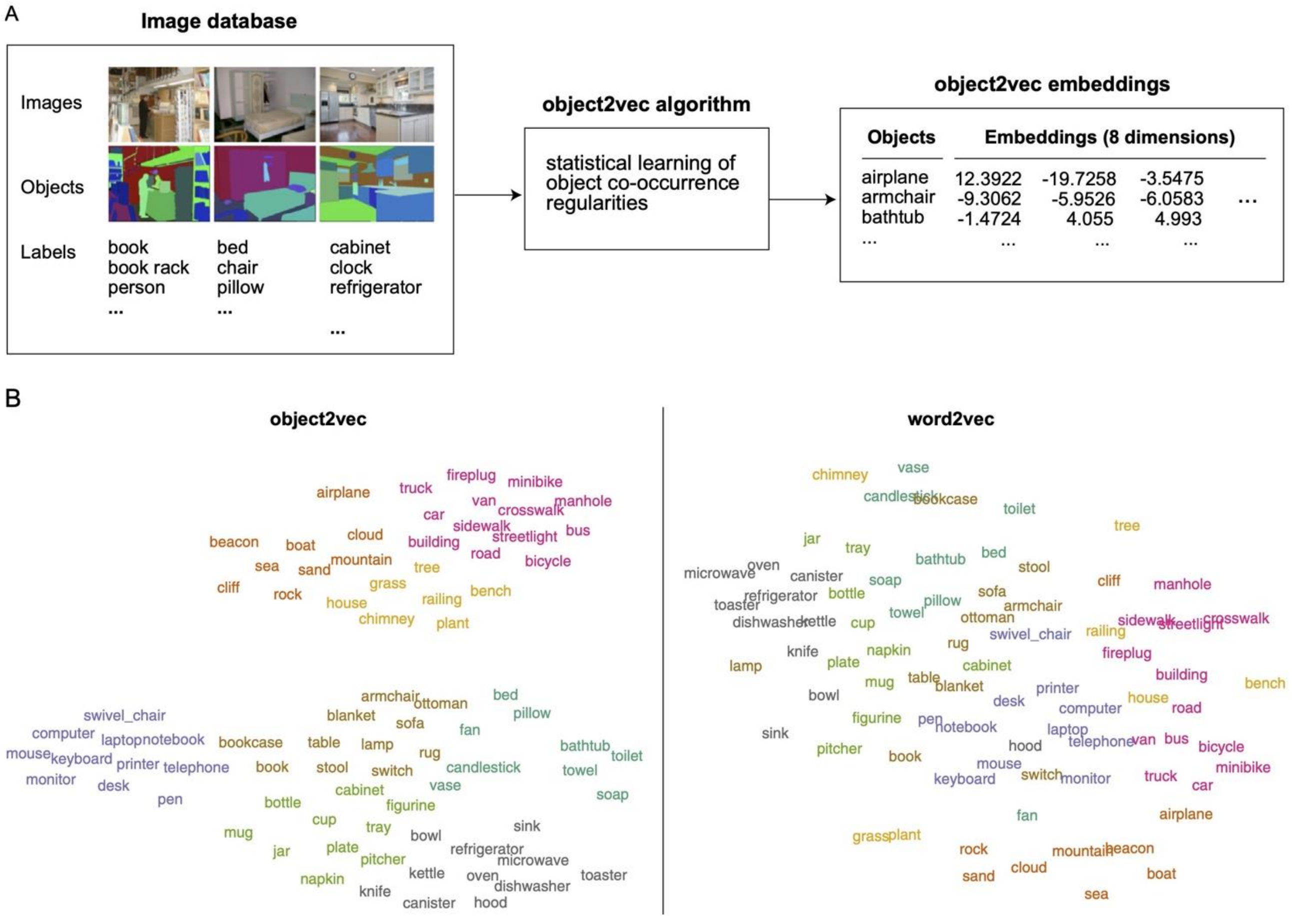


- KNN



- Regressao e classificao

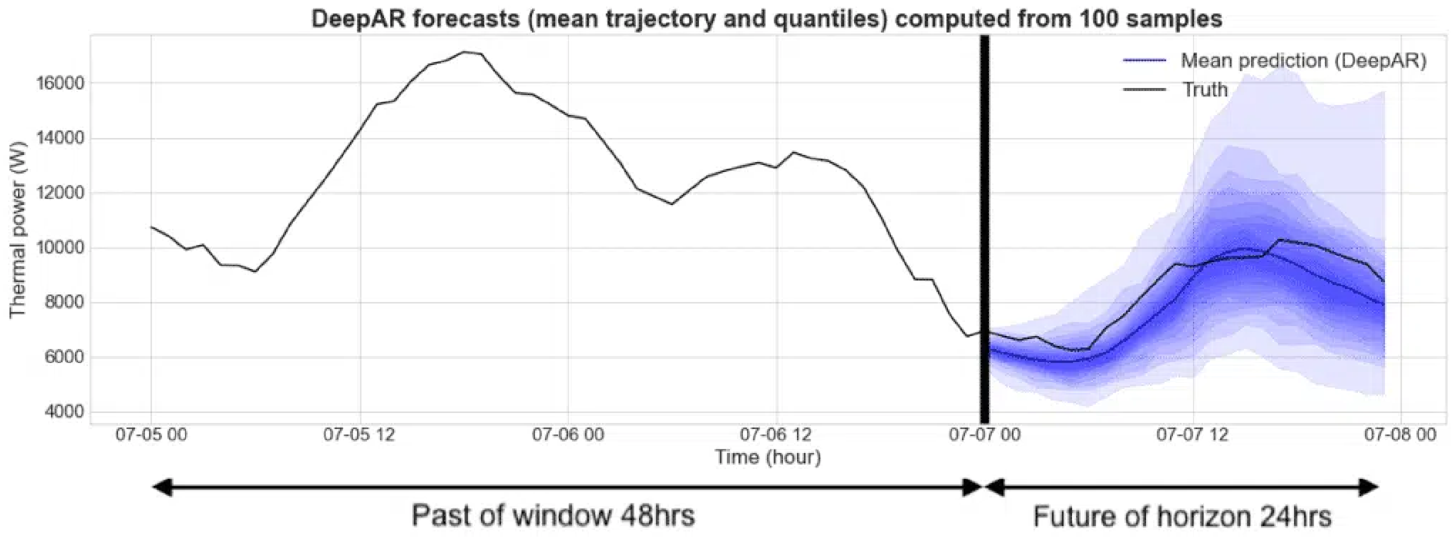
•Object2vec



Objects	Embeddings (8 dimensions)			
airplane	12.3922	-19.7258	-3.5475	...
armchair	-9.3062	-5.9526	-6.0583	
bathtub	-1.4724	4.055	4.993	
...	...	...	...	

•DeepAR

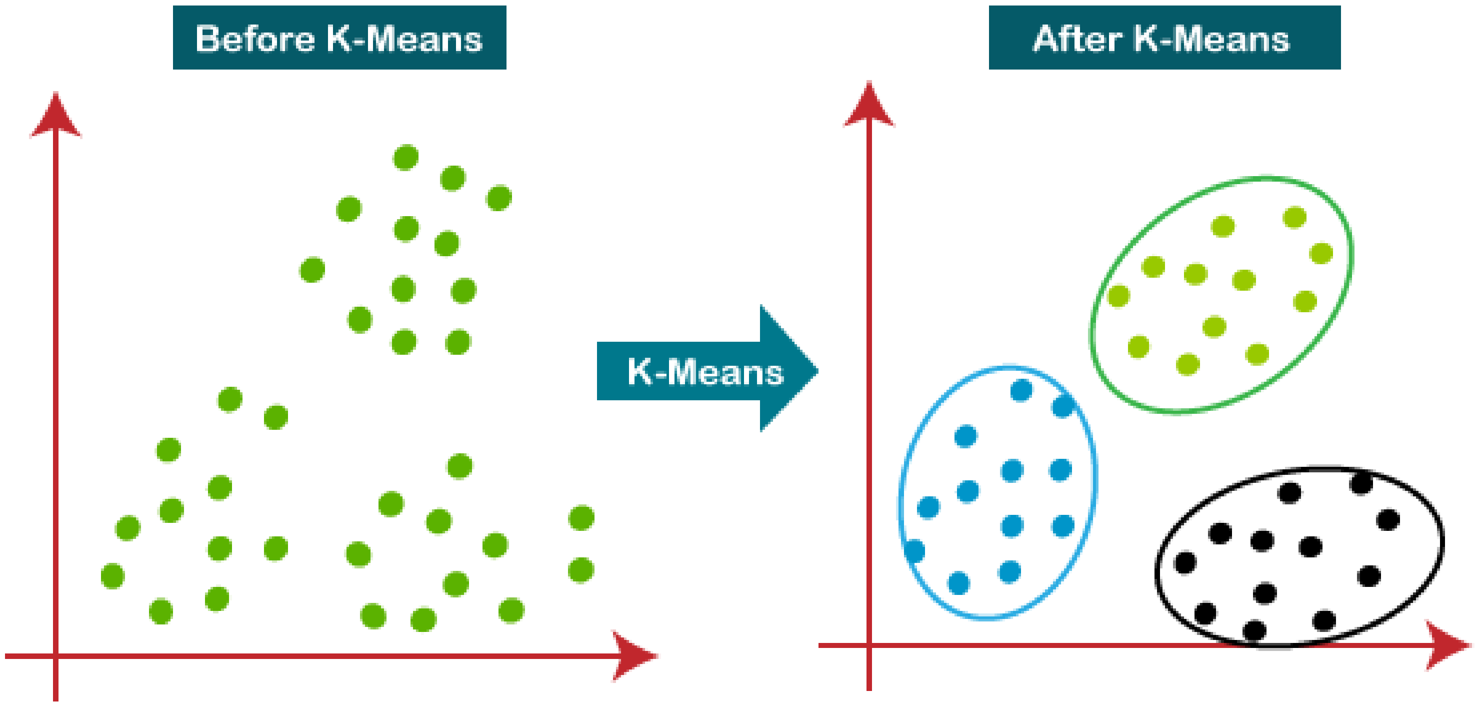
- Forecasting



Supervised learning

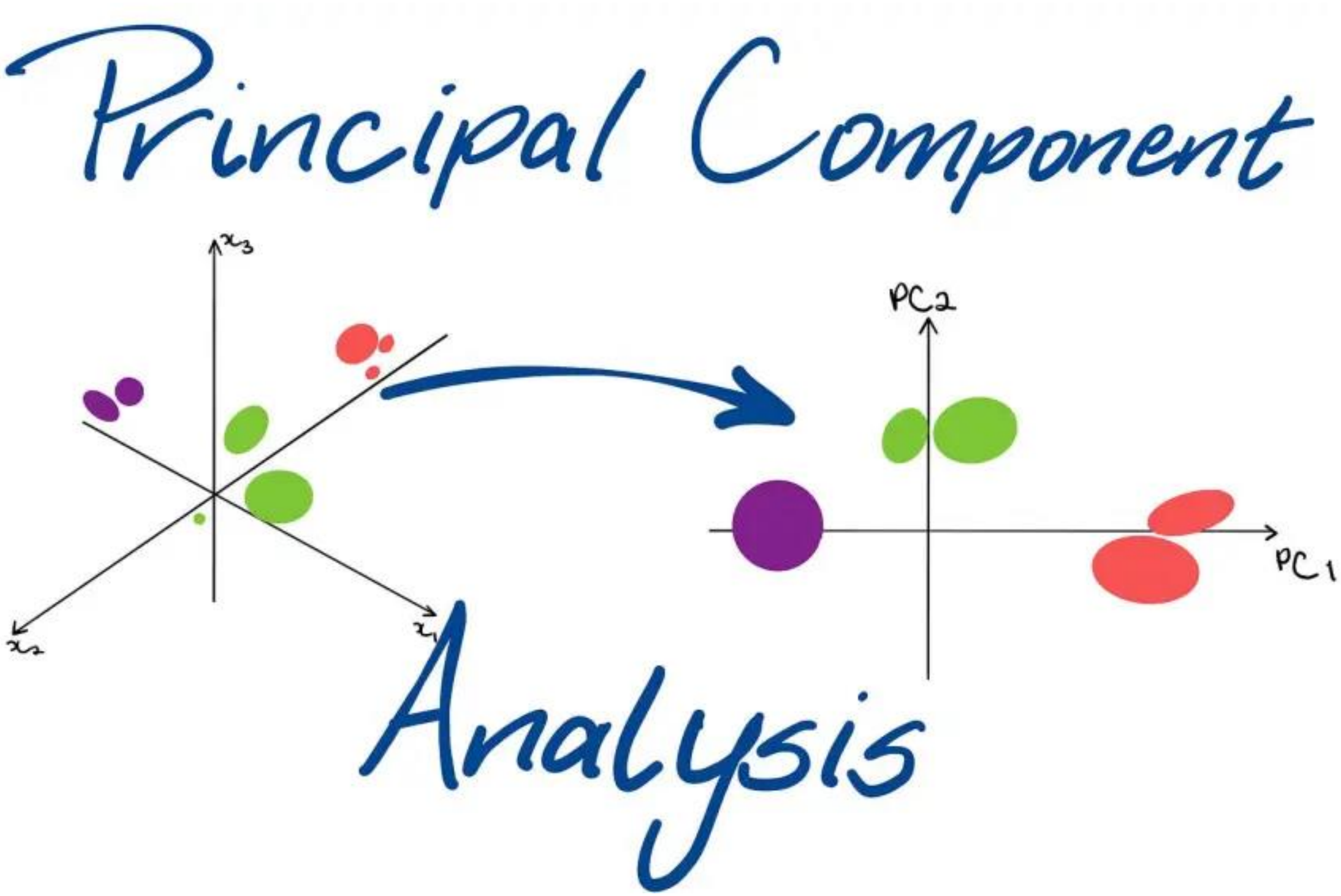
- K-means algorithm
- **Principal Component Analysis** (PCA)
- IP Insights
- **Random Cut Forest Algorithm** (RCF)

K-means



- Clustering

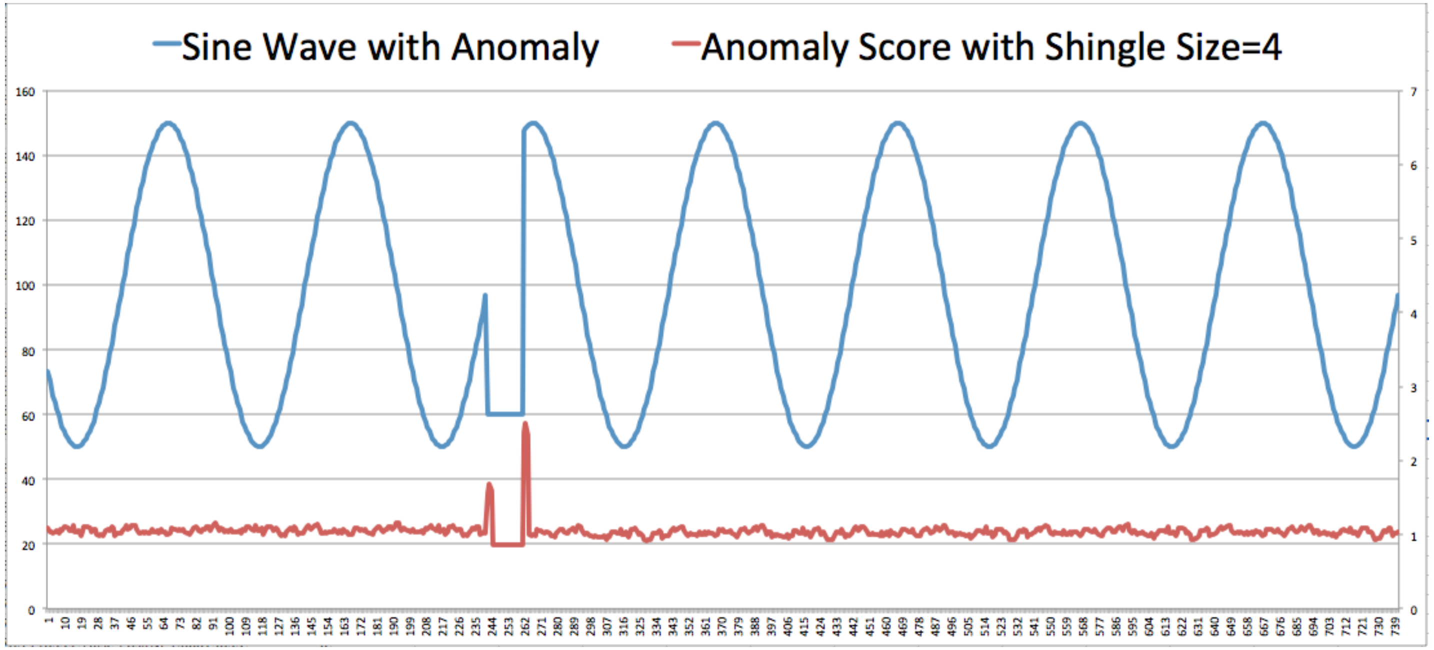
PCA



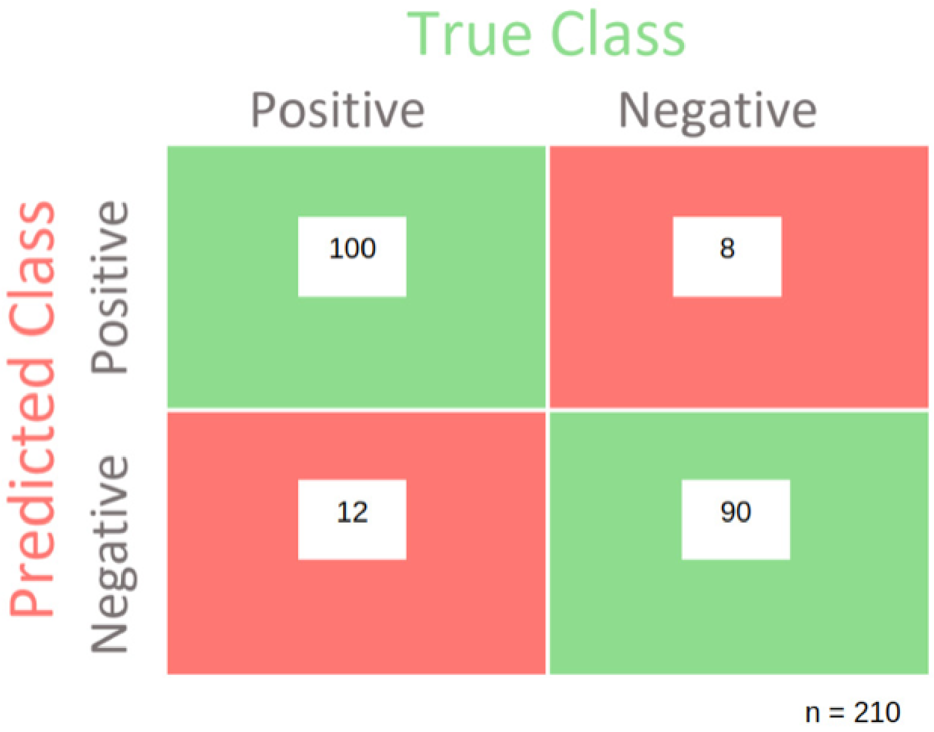
- Reduçao de dimensionalidade

Random cut forest

- Deteccao de anomalias



Avaliando modelos de classificacao

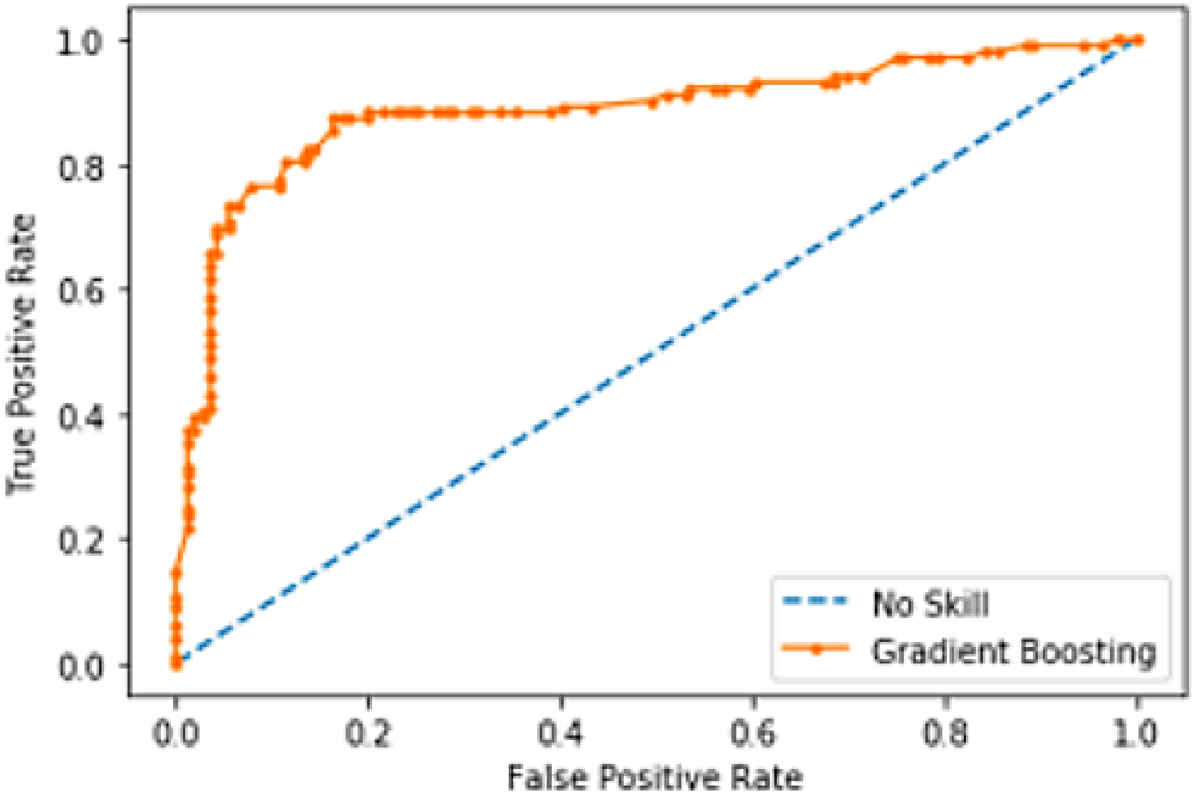


		True Class	
		Positive	Negative
Predicted Class	Positive	100	8
	Negative	12	90

n = 210

- Acuracia = TP+TN/n
- Recall = TP/TP+TN
- Precisao = TP/TP+FP
- Acuracia é útil em bases bem balanceadas
- Otimizar recall quando falsos negativos são aceitáveis
- Otimizar precisão quando falsos positivos são aceitáveis

Curva ROC



- Curva que sumariza os trade off entre true positive e false positive
- O ideal da curva ROC é formar uma esquina perfeita entre o ponto 0.0/1.0
- AUC: Area under the curve: Quanto maior melhor(entre 1 e 0)
- F1 score medica armonica entre precisao e recall

Avaliando modelos de regressao

- Mean Absolute Error(MAE): penaliza grandes erros

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} (\underbrace{y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i}_{\text{actual value}})^2$$

- Root Mean Squared Error: mais utilizada pois é mais fácil de interepretar

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Otimizando modelos

- GridSearch: testa diferentes combinações de algoritmos e hyperparamentos
  - Deve ser definido qual métrica deve ser otimizada, quantas combinações devem ser testadas
  - Voce pode definir manualmente os limites das combinações
  - Pode utilizar Random Search para que o grid Search utilize valores randômicos para teste
  - Pode utilizar otimizacao Bayesiana, utiliza abordagem probabilisca para encontrar os melhores parametros

Informacoes adicionais

- Estudar a seção 4 do curso da udemy, principalmente algorimtmos não tratados aqui como:
- LDA
- Word2Vec
- Neural Topic Model
- seq2seq

Q1

- A real estate company wants to create a machine learning (ML) model to predict housing prices based on a historical dataset. The dataset contains 32 features.
  Which algorithm will meet these requirements?
- A - Logistc regression
- B – Linear Regression
- C – K-means
- D – Principal Component Analysis (PCA)

Q1

- A real estate company wants to create a machine learning (ML) model to predict housing prices based on a historical dataset. The dataset contains 32 features.
  Which algorithm will meet these requirements?
- A - Logistc regression -> Classificao
- B – Linear Regression
- C – K-means -> Agrupamento de objetos similares, não supervisionado
- D – Principal Component Analysis (PCA) -> redução de dimensão, quando queremos reduzir o numero de features.

Q2

- A company is building a website that offers a variety of comedy content for adults and children. The company intends to automate the process of ingesting the content and tagging the content as safe for viewing by children as the positive class. The company's top priority is to avoid showing inappropriate content to children.
- What is the MOST relevant metric for the company to use to evaluate the machine learning (ML) model for this task?
- A – Recall
- B – Accuracy
- C – AUC/ROC
- D – Precision

Q2

- A company is building a website that offers a variety of comedy content for adults and children. The company intends to automate the process of ingesting the content and tagging the content as safe for viewing by children as the positive class. The company's top priority is to avoid showing inappropriate content to children.
- What is the MOST relevant metric for the company to use to evaluate the machine learning (ML) model for this task?
- A – Recall -> Mais relevante para minimizar falsos negativos, queremos minimizar falsos positivos.
- B – Accuracy -> Predicoes corretas de forma global
- C – AUC/ROC -> Reflete os trade off entre false positivo e falso negativo, queremos otimizar falsos positivos.
- D – Precision

Q3

- A machine learning (ML) specialist is training a model by using a supervised learning algorithm. The ML specialist split the dataset to use 80% of the data for training and 20% of the data for testing. While evaluating the model, the ML specialist discovers that the model is 97% accurate for the training dataset and 75% accurate for the test dataset.
- Which action should the ML specialist take?
- A – Ignore the difference and deploy the model
- B – Change the hyperparameters to reduce overfitting of the model, retrain the model
- C – Balance the model by adding data to the test set. Take data from the end of training set, and move the data to the test set. Balance the model with 70% of the data in the training set and 30% of the data in the test set
- D – Change the hyperparameters to make the model more specific. Retrain the model.

Q3

- A machine learning (ML) specialist is training a model by using a supervised learning algorithm. The ML specialist split the dataset to use 80% of the data for training and 20% of the data for testing. While evaluating the model, the ML specialist discovers that the model is 97% accurate for the training dataset and 75% accurate for the test dataset.
- Which action should the ML specialist take?
- A – Ignore the difference and deploy the model -> o modelo está com overfitting
- B – Change the hyperparameters to reduce overfitting of the model, retrain the model
- C – Balance the model by adding data to the test set. Take data from the end of training set, and move the data to the test set. Balance the model with 70% of the data in the training set and 30% of the data in the test set. -> o ideal seria misturar os dados e redividir novamente, não remover um conjunto especifico dos dados de treino e mover.
- D – Change the hyperparameters to make the model more specific. Retrain the model. -> o modelo esta overfittando, queremos que ele se tornem mais generalista, não mais especifico.

Q3

- A machine learning (ML) specialist is training a model by using a supervised learning algorithm. The ML specialist split the dataset to use 80% of the data for training and 20% of the data for testing. While evaluating the model, the ML specialist discovers that the model is 97% accurate for the training dataset and 75% accurate for the test dataset.
- Which action should the ML specialist take?
- A – Ignore the difference and deploy the model -> o modelo está com overfitting
- B – Change the hyperparameters to reduce overfitting of the model, retrain the model
- C – Balance the model by adding data to the test set. Take data from the end of training set, and move the data to the test set. Balance the model with 70% of the data in the training set and 30% of the data in the test set. -> o ideal seria misturar os dados e redividir novamente, não remover um conjunto especifico dos dados de treino e mover.
- D – Change the hyperparameters to make the model more specific. Retrain the model. -> o modelo esta overfittando, queremos que ele se tornem mais generalista, não mais especifico.

Q4

- A company has 1,000 sentences with sentiments categorized as positive, neutral, or negative.
- Which algorithm should a machine learning (ML) specialist select for training a baseline sentiment model?
- A – K-nn
- LDA
- K-means
- Random cut forests

Q4

- A company has 1,000 sentences with sentiments categorized as positive, neutral, or negative.
- Which algorithm should a machine learning (ML) specialist select for training a baseline sentiment model?
- A – K-nn
- LDA -> não supervisionado
- K-means -> não supervisionado
- Random cut forests -> Identificacao de anomalias