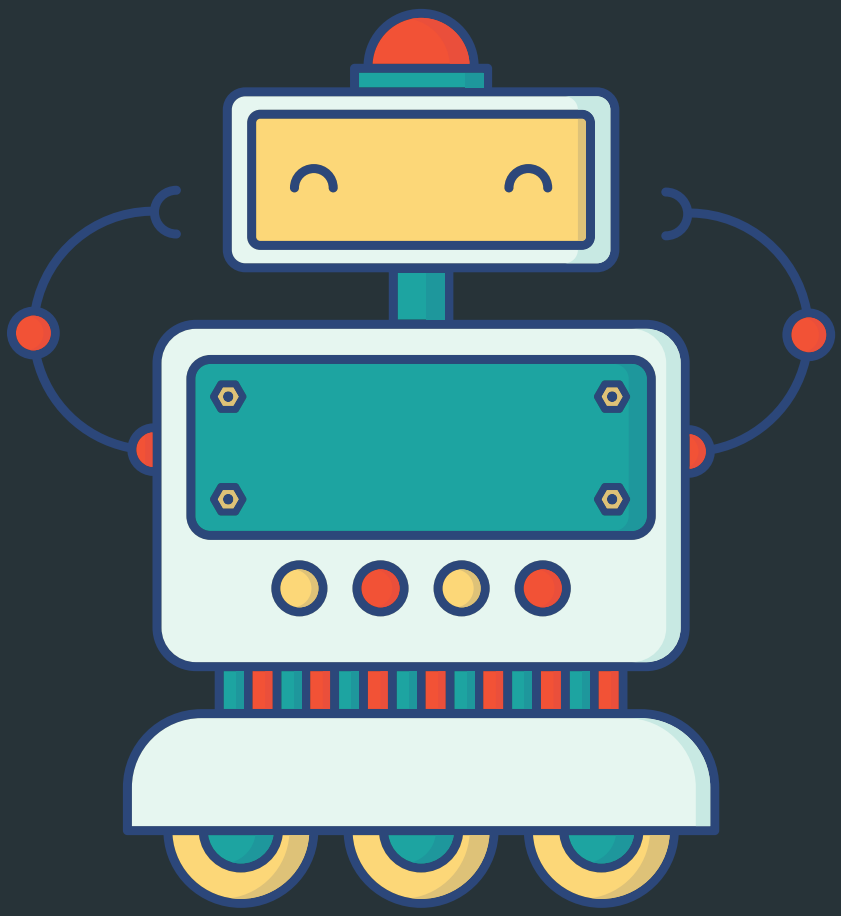
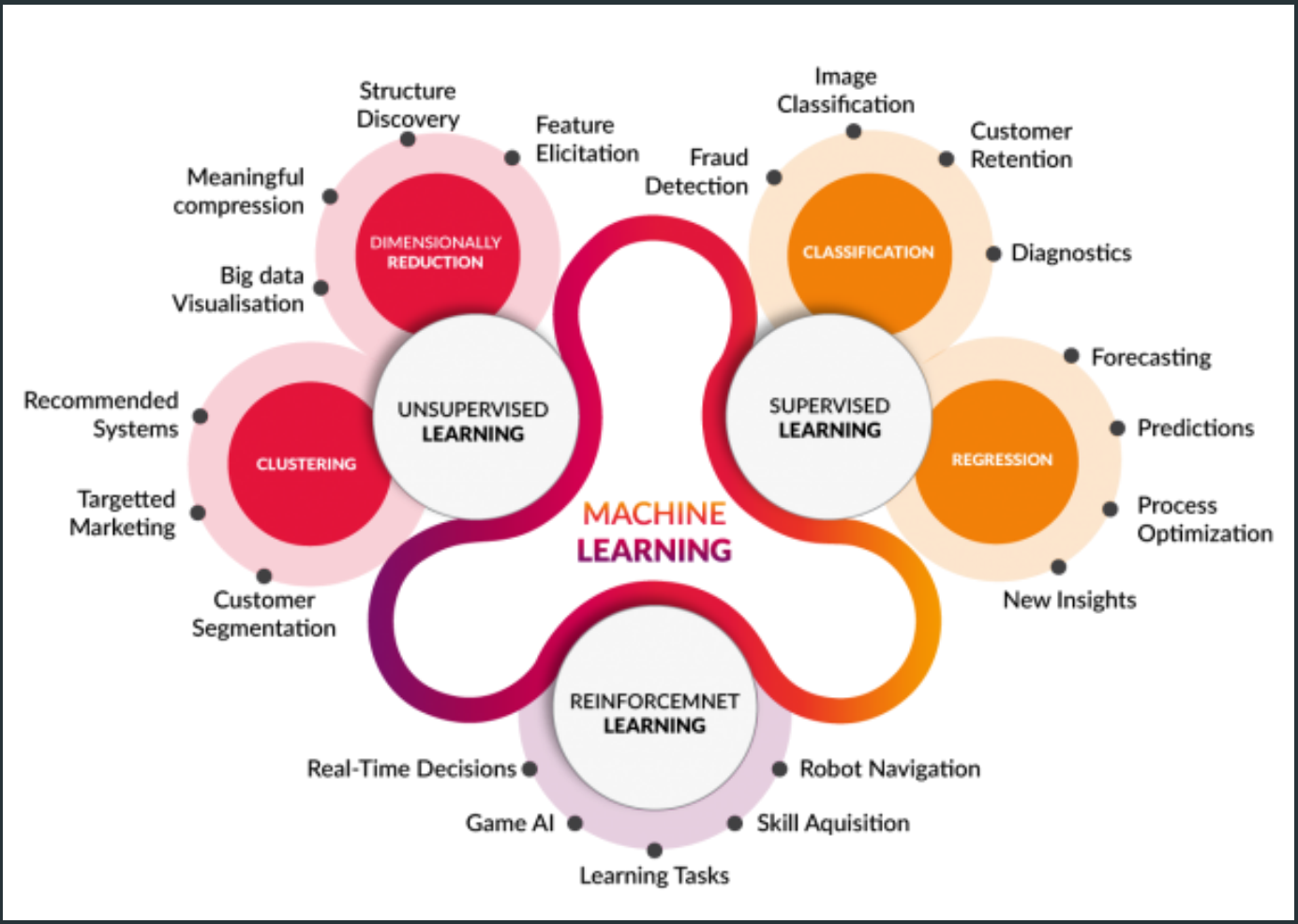


MACHINE LEARNING

PYTHON

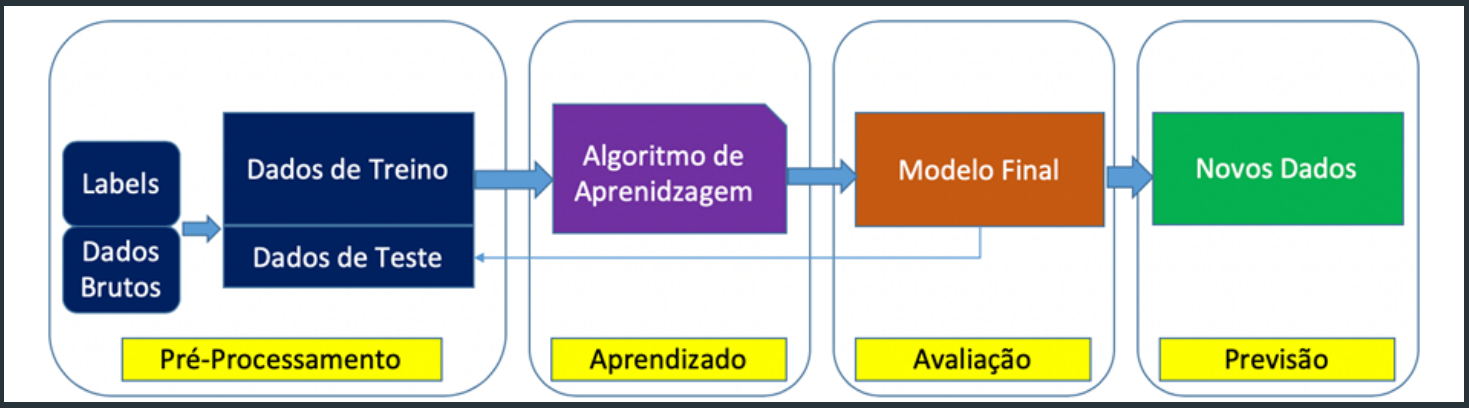
1. Processo de Machine Learning



1.1 - Processo de construção de modelos de Machine Learning

- Transformação de Variáveis
- Feature Selection
- Redução de Dimensionalidade
- Amostragem

- Validação do Modelo
- Otimização



- Seleção do Modelo
- Cross-Validation
- Métricas de Performance
- Otimização

1.2 - Processo de Análise de Dados

1.Definição do problema de negócio

2.Extrair e carregar os dados

3.Análise exploratória dos dados (etapa para compreender os dados)

- Análise descritiva (ex, pandas)
- Análise com visualização matplotlib (ou outros)
- Análise com visualização Seaborn (melhor para análise estatística dos dados)

4.Pré-Processamento: Preparando os dados para Machine Learning

(cada algoritmo de ML requer uma abordagem diferente, conferir documentação)

-> **Transformação de Variáveis**: (técnicas aplicadas à variáveis quantitativas)

- Normalização - altera a escala dos dados - (métodos 1(MinMaxScaler) e 2(Normalizer) do Scikit-learn)
- Padronização - coloca os dados em uma distribuição normal (gaussiana) (não altera a escala dos dados) (fç StandardScaler)
- Binarização - transformar em valores binários (fç Binarizer) (usado em deep learning)

-> **Feature Selection:** (seleção das melhores candidatas a var preditoras, ajuda a reduzir o overfitting)

- Seleção Univariada (SelectKBest, ex chi2) (site com lista de todos os testes estatisticos oferecidos)
- Eliminação Recursiva de Atributos (RFE, LogisticRegression) (nessa técnica é usado ML) (Score)
- Método Ensemble para Seleção de Variáveis (ExtraTreesClassifier) (ML) (aumentar a precisão) (Score)

-> **Redução de Dimensionalidade:** (O PCA requer dados normalizados ->MinMaxScaler)

- Principal Component Analysis (PCA algoritmo de ML não supervisionada) (MinMaxScaler, PCA) (atributos com menos relevância são levados para outro componente) (componente não é uma var, e sim um grupo de vars com variância similar)

-> **Amostragem (Resamling)**

-> **Dados de Treino e de Teste:**

(Treinamos o algoritmo nos dados de treino e fazemos as previsões nos dados de teste e avaliamos o resultado)

- Seleção randomica 70/30 e 65/35
- Divisão estática (ex: ~78.7%)

-> **Cross-Validation** (usado no Pré-Processamento e no Aprendizado) ("Divisao dinamica" ex:~77.5%)

(executa várias vezes o processo Treino/Teste) (Kfold, cross_val_score (treina e testa....), LogisticRegression)

5. Aprendizado (etapas abaixo são realizadas simultaneamente)

-> Seleção do modelo

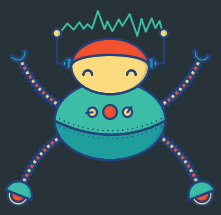
-> Métricas de performance: (Na documentação do scikit-learn) (Acurácia, Curva ROC, Confusion Matrix, Relatório de calssificação... ex de métricas de classificação!)

-> Otimização - Ajuste de Hyperparâmetros (Grid Search Parameter Tuning e Random Search Parameter Tuning) - Otimizando Performance com Métodos Ensemble (Bagging, Boosting, Voting)

6. Avaliação 7. Previsão

2. Algoritmos de Machine Learning

Aprendizagem Supervisionada



Regressão

(prever valor numérico)

- Regressão Linear

  (prever o valor de uma variável contínua)
- SMV- Support Vector Machine

  (requer normalização dos dados)

  (Para dados não-linearmente espaçados)
- KNN - K-Nearest Neighbors

  (requer normalização dos dados)
- Gradient Boosting
- Adaboost
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- CART - Classification and Regression Trees

Classificação

(prever categoria)

- Regressão Logistica (Linear)

  (requer normalização dos dados)
- Árvore de Decisão
- Random Forest

  (também usar para seleção de variáveis)
- SMV - Support Vector Machine

  (requer normalização dos dados)

  (Para dados não-linearmente espaçados)
- Naive Bayes (Não linear)

  (Agoritmo Probabilístico)
- KNN - K-Nearest Neighbors

  (requer normalização dos dados)
- Gradient Boosting
- Adaboost
- CART - Classification and Regression Trees (Não Linear)
- LDA - Linear Discriminant Analysis (Linear)

  (requer normalização dos dados)

Aprendizagem Não-supervisionada

- K-Means
- Redução de Dimensionalidade

Melhores Algoritmos para problemas de Classificação Binária:

- Regressão Logística*
- LDA
- Naive Bayes

Melhores Algoritmos para problemas de Classificação Multi-Class:

- SVM
- Rede Neural

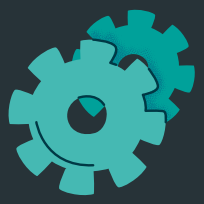
Algumas métricas para Algoritmos de Regressão:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R Squared (R^2)
- Adjusted R Squared (R^2)
- Mean Square Percentage Error (MSPE)
- Mean Absolute Percentage Error (MAPE)
- Root Mean Squared Logarithmic Error (RMSLE)

Algumas métricas para Algoritmos de Classificação:

- Acurácia
- Curva ROC
- Confusion Matrix
- Relatório de calssificação

3. Frameworks de Machine Learning



- Scikit-learn (Python)
- Caret (R)
- TensorFlow (Python, R, Java, C++)
- Apache Mahout (Python, Java)
- Spark Mllib (Scala, Java, Python, R)
- H2O (Java, Python)
- Weka (Java, Python)
- PyTorch, CNTK, MXNet (Python, C++, Java)