ConFLOSS

Ambiente Livre

Big Data & Data Science

Low-Code Data Science with Pentaho Machine Intelligence

Marcio Junior Vieira
CEO & Data Scientist, Ambiente Livre

ConFLOSS

Marcio Junior Vieira

Ambiente Livre
Big Data & Data Science

Mini-cv

- 20 anos de experiência em TI, vivência em desenvolvimento, análise e ciência de dados.
- CEO da Ambiente Livre atuando como Cientista de Dados, Engenheiro de Dados e Arquiteto de Software.
- Professor dos MBAs em Big Data & Data Science, Inteligência Artificial e BI da Universidade Positivo.
- Professor do MBA Artificial Intelligence e Machine Learning da FIAP.
- Pesquisador do Laboratório de tecnologias de tomada de decisão da Universidade de Brasília (UNB/Latitude).
- Trabalhando com Free Software e Open Source desde 2000 com serviços de consultoria e treinamento.
- Graduado em Tecnologia em Informática(2004) e pós-graduado em Software Livre(2005) ambos pela UFPR.
- Palestrante FLOSS em: FISL, TDC, Latinoware, Campus Party, Pentaho Day, Ticnova, PgDay e FTSL.
- Organizador Geral: Pentaho Day 2017, 2015, 2019 e apoio nas ed. 2013 e 2014.
- Data Scientist, instrutor e consultor de Big Data e Data Science com tecnologias abertas.
- Ajudou a capacitar equipes de Big Data na IBM, Accenture, Tivit, Serpro, Natura, MP, Netshoes, Embraer, etc.
- Especialista em implantação e customização de Big Data com Hadoop, Spark, Pentaho, Cassandra e MongoDB.
- Contribuidor de projetos internacionais, tais como Pentaho, LimeSurvey, SuiteCRM e Camunda.
- Especialista em implantação e customização de ECM com Alfresco e BPM com Activiti, Flowable e Camunda.
- Certificado (Certified Pentaho Solutions) pela Hitachi Vantara (Pentaho).
- Membro da The Order Of de Bee (comunidade Alfresco para desenvolver o ecossistema Alfresco independente)

Nosso Ecossistema de Serviços

Ambiente Livre
Big Data & Data Science

Big Data e Data Science

Análise de Dados da IoT
Análise Preditiva
Processamento Distribuído
Banco de Dados Colunares

Big Data & Data Lake
Big Data Analytics
Machine Learning

Consultoria | Treinamento | Projeto

CRM e CMS

Marketing e Vendas
Fidelização
SAC e Pós-vendas
Portais de Conteúdo

Customer Relationship Management
Content Management System
Pesquisa de Mercado & SLA

Consultoria | Treinamento | Projeto

ECM e BPM

Gestão de Documentos
Gerenciamento de Mídias
Processo de Negócio
BPMN e BPMS

Enterprise Content Management
Records Management
Business Process Management

Consultoria | Treinamento | Projeto

Business Intelligence

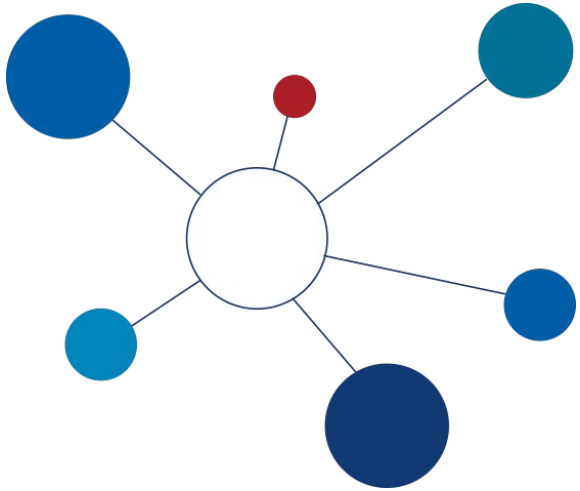
Painéis de Indicadores
Cubos de Análise
Relatórios Gerênciais
Tomada de Decisão

Business Intelligence & Analitycs
Dashboards e OLAP
Data Integration & Data Mining

Consultoria | Treinamento | Projetos

pentaho · Alfresco · hadoop · mongoDB · Cassandra · suiteCRM · LPI · Spark · The Apache Software Foundation · LIFERAY Enterprise. Open Source. For Life. · LimeSurvey · PostgreSQL the world's most advanced open source database · Apache Flink · camunda · Flowable making process personal · moodle · MariaDB · Drupal · ConFLOSS Conferência de Free/Libre e Open Source Software

Agenda



① Low Code.
② Data Science Low Code.
③ Pentaho
④ Pentaho Data Mining.
⑤ Pentaho Data Integration.
⑥ Pentaho Machine Intelligence.

Low Code

Definição

- O termo “Low Code” foi criado em 2014 para denotar plataformas que tinham interfaces de desenvolvimento baseadas em GUI (Graphical User Interface).
- Codificação tradicional sem a necessidade de conhecer explicitamente a linguagem de programação.
- Elimina a necessidade de criar estruturas, vincular diferentes bancos de dados e realizar outras tarefas que normalmente são necessárias para codificar um software ou um aplicativo.
- Desenvolvimento mais simples e fácil.
- Pessoas sem conhecimento em codificação podem desenvolver aplicativos.

Por que muitas empresas adotam Low Code?

Definição

- Grandes gargalos para serem resolvidos antes que um novo software seja criado e implementado.
- Com o desenvolvimento dessa ferramenta, é possível a uma empresa receba um programa ou software rapidamente.
- Plataformas Low Code:
  * Interfaces de arrastar e soltar.
  * Modelagem Visual.
  * Segurança e escalabilidade.

Data Science Low Code

Ambiente Livre
Big Data & Data Science

Definição

- Aplica os conceitos de Low Code a area de Ciência de dados.
- Simplifica a necessidade de Conhecimento avançado em estatística e tratamento de dados.

ConFLOSS
Conferência de Free/Libre e Open Source Software

Ferramentas - Data Science / IA - Low Code

Definição

- Muitas opções em Cloud!
- Algumas opções Open Source!

3 Pilares do Pentaho

- Plataforma abrangente para integração de dados e Business Analytics.

Data Integration

Business Analytics

Big Data

Weka

- Desenvolvido pela Universidade de Waikato (**W**aikato **E**nvironment for **K**nowledge **A**nalysis)
- Licença GPL
- Desenvolvido em Java
- Iniciado o desenvolvimento em 1993.
- O software foi adquirido pela Pentaho Corporation em 2016 (Hoje chamada de Hitachi Vantara).

- Site do projeto: http://www.cs.waikato.ac.nz/ml/weka/

Funcionalidades

- Aprendizado de máquina
- Mineração de Dados
- Pré-processamento
- Classificação
- Regressão
- Agrupamento
- Regras de associação
- Atributo de seleção
- Experiências
- Workflow
- Visualização

Atual mantenedor

- Líder do projeto Weka / PDM.
- 15 anos de experiência como pesquisador acadêmico em ciências da computação.
- Diversas publicações em conferências de aprendizado de máquina, mineração de dados e revistas.
- Se formou no pós-doutorado da Universidade de Waikato, na Nova Zelândia.

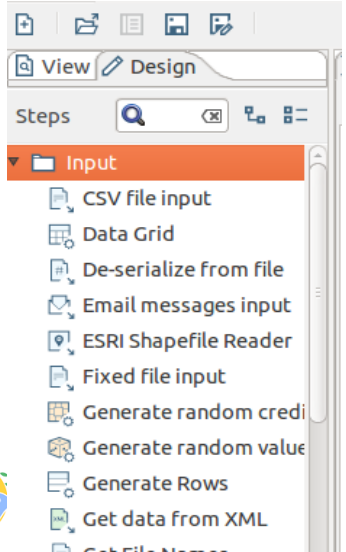
- Blog: markahall.blogspot.com
-

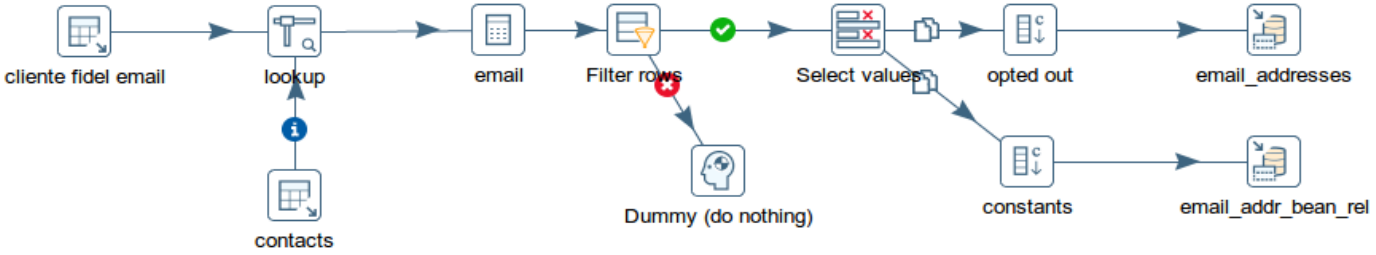
Interface Weka

- **Explorer**: Uso geral (Pré-processamento, clusterização Classificação, visualização)
- **Experimenter**: controle de treinamento (divisão do conjunto teste/treinamento, cross-validaritio
- **KnowledgeFlow**: Tarefas de ETL como fluxo dados
- **Workbench:** GUI antiga.
- **Simple CLI**: Console para uso por linha de comando

Pentaho Data Integration (PDI)

- Processa em Paralelo (também roda em Cluster Spark)
- Acessar dados diretamente (se necessário sem DW)
- Permite publicar dados diretamente em Reports, Ad-Hoc Reports e Dasboards.
- “Programação e Fluxo Visual” com aproximadamente 350 steps





Funcionalidades do Pentaho Data Integration

Funcionalidades Tradicionais

- Usadas em projetos de data warehouse

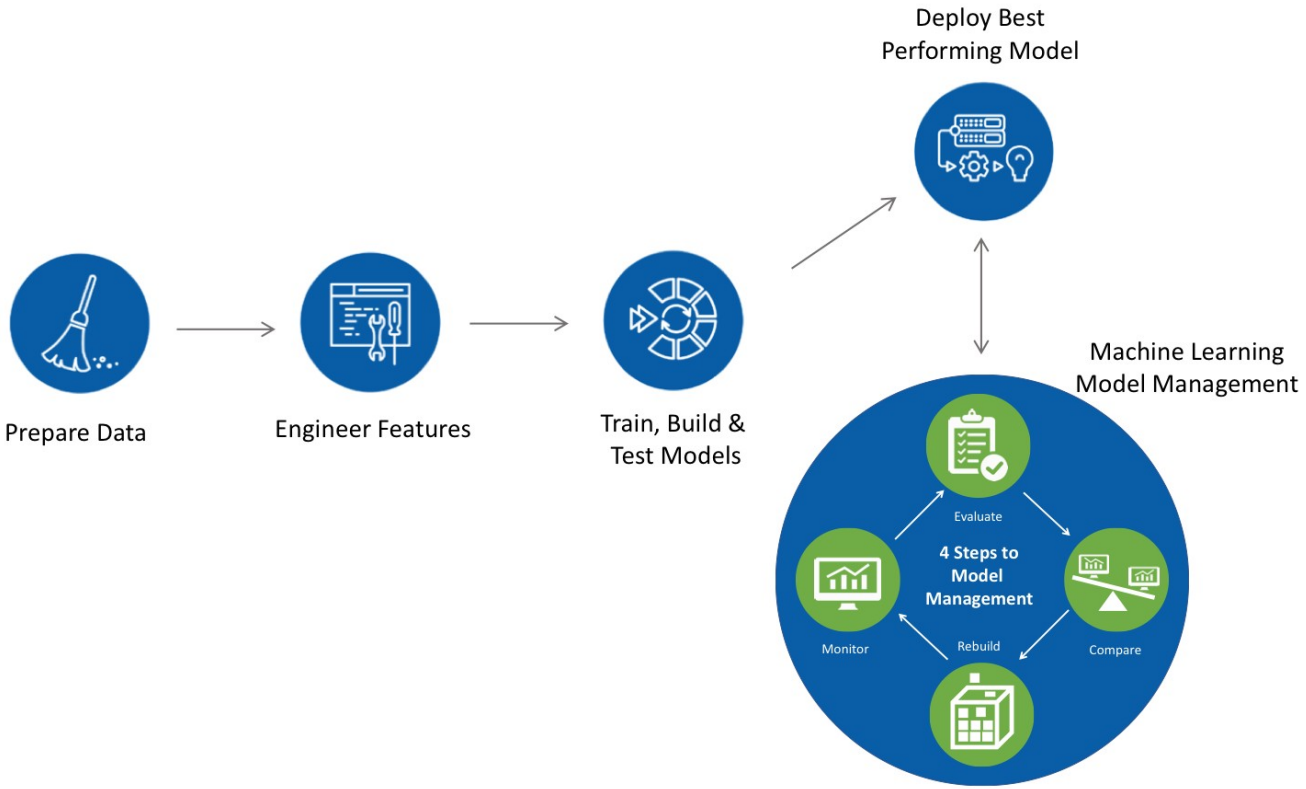
Funcionalidades Adicionais

- Migração de dados entre aplicações/banco de dados
- Exportar dados de banco de dados para arquivos texto
- Carregar massivamente dados em banco de dados
- Data Cleansing – disciplina de qualidade/limpeza de dados de data warehouse
- Integração de aplicações.
- Gerenciamento de Filesystem (File management)

Pentaho Machine Intelligence

PMI

- Plugin do PDI.



Pentaho Machine Intelligence

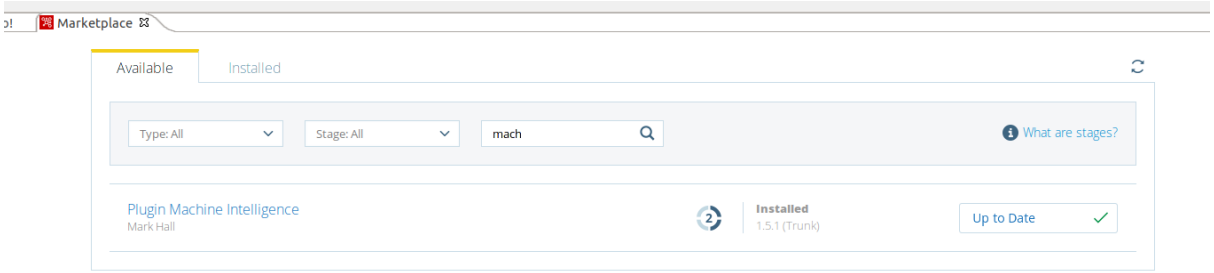
Motores de ML

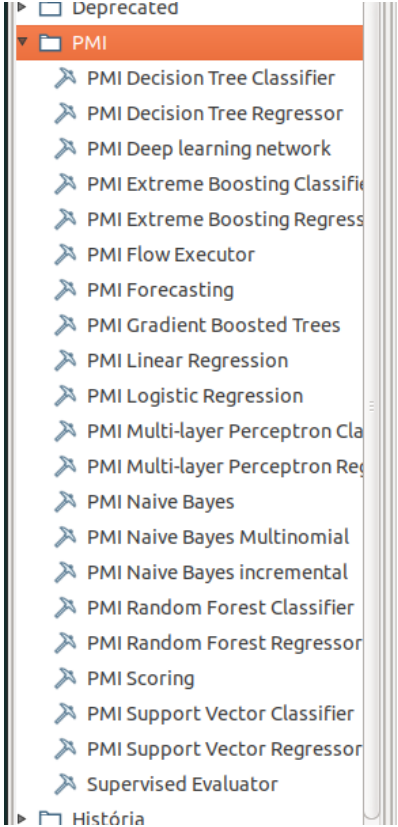
- Weka.
- Python.
- R.
- Spark MLib.

Instalando o PMI

Marketplace

- Pode acessar o menu do Marktplace no próprio PDI e instalar.





Tipos de Aprendizado de Máquina

Ambiente Livre
Big Data & Data Science

Supervisionada

- Classificação (Naive Bayes, SVM,Randon Decision Forest)
- Regressão (Linear, Logistic)

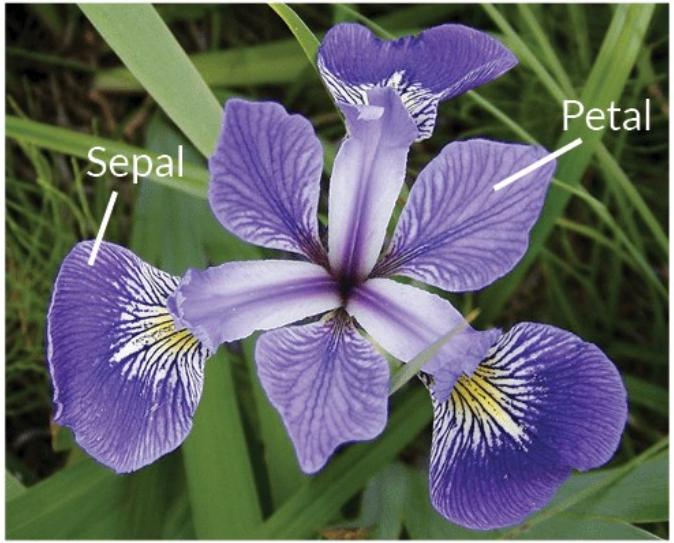
Não Supervisionada

- Associação
- Agrupamento/Clustering (K-Means)
- Detecção de Desvios
- Padrões Sequênciais
- Sumarização

Reforço

ConFLOSS
Conferência de Free/Libre e Open Source Software

Supervisionado - Extração de Características



Iris Versicolor

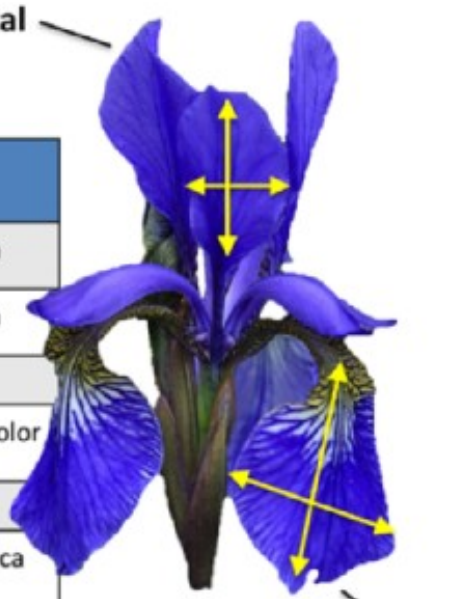
Iris Setosa

Iris Virginica

Supervisionado - Extração de Características

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Class labels
(targets)

Features
(attributes, measurements, dimensions)

Aprendizagem Supervisionada

Supervisor

Fase 1

Extração de Características → Algoritmo de Aprendizagem → Modelo Preditivo

Fase 2

Extração de Características → Modelo Preditivo → Iris Versicolor

Caso Prático.

Ambiente Livre
Big Data & Data Science

Dataset

- Desatre do Titanic.

O que queremos prever

- Se você congressista sobreviveria se tivesse embarcado.

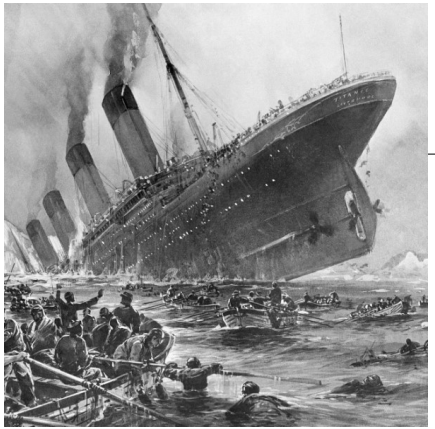
- Vamos fazer na prática!

ConFLOSS
Conferência de Free/Libre e Open Source Software

Caso Prático - Analise do Titanic.

Supervisor

Fase 1



Extração de Características → Algoritmo de Aprendizagem → Modelo Preditivo

Fase 2

Extração de Características → Modelo Preditivo → Sobreviveria?

Agradecimentos

Ambiente Livre
Big Data & Data Science

- Agradeço ao ConFLOSS pelo convite!
- Obrigado ao Anderson e Galvão!
- Obrigado aos palestrantes e congressistas!

ConFLOSS
Conferência de Free/Libre e Open Source Software

Referências

Ambiente Livre
Big Data & Data Science

Web

- Você sabe o que é Low Code?
  https://www.blendit.com/2020/11/25/voce-sabe-o-que-e-low-code/
- Artificial Intelligence with Pentaho -
  https://community.hitachivantara.com/s/article/artificial-intelligence-with-pentaho-1
- **Slides da Apresentação**
  https://www.slideshare.net/ambientelivre

Códigos Fontes das Transformação e Jobs

@ambientelivre no Github
https://github.com/ambientelivre/samples-pentaho/tree/master/data-integration/pentaho-machine-intelligence

ConFLOSS
Conferência de Free/Libre e Open Source Software

Obrigado

Ambiente Livre
Big Data & Data Science

Marcio Junior Vieira
marcio@ambientelivre.com.br
@marviojvieira @ambientelivre
@ambientelivreopensoftware
https://www.linkedin.com/in/mvieira1/

Slide da Palestra será publicada em:
Linkedin….: https://www.linkedin.com/in/mvieira1/
SlideShare: http://slideshare.net/ambientelivre/
Blog……...: http://blogs.ambientelivre.com.br/marcio/

ConFLOSS