Workshop:
Machine Learning e pesquisas na área contábil: status e agenda para futuras pesquisas

10 ANOS PPGCC FACIC/UFU

Palestrante:

Pós doutorando Jefferson Melo (UFU)

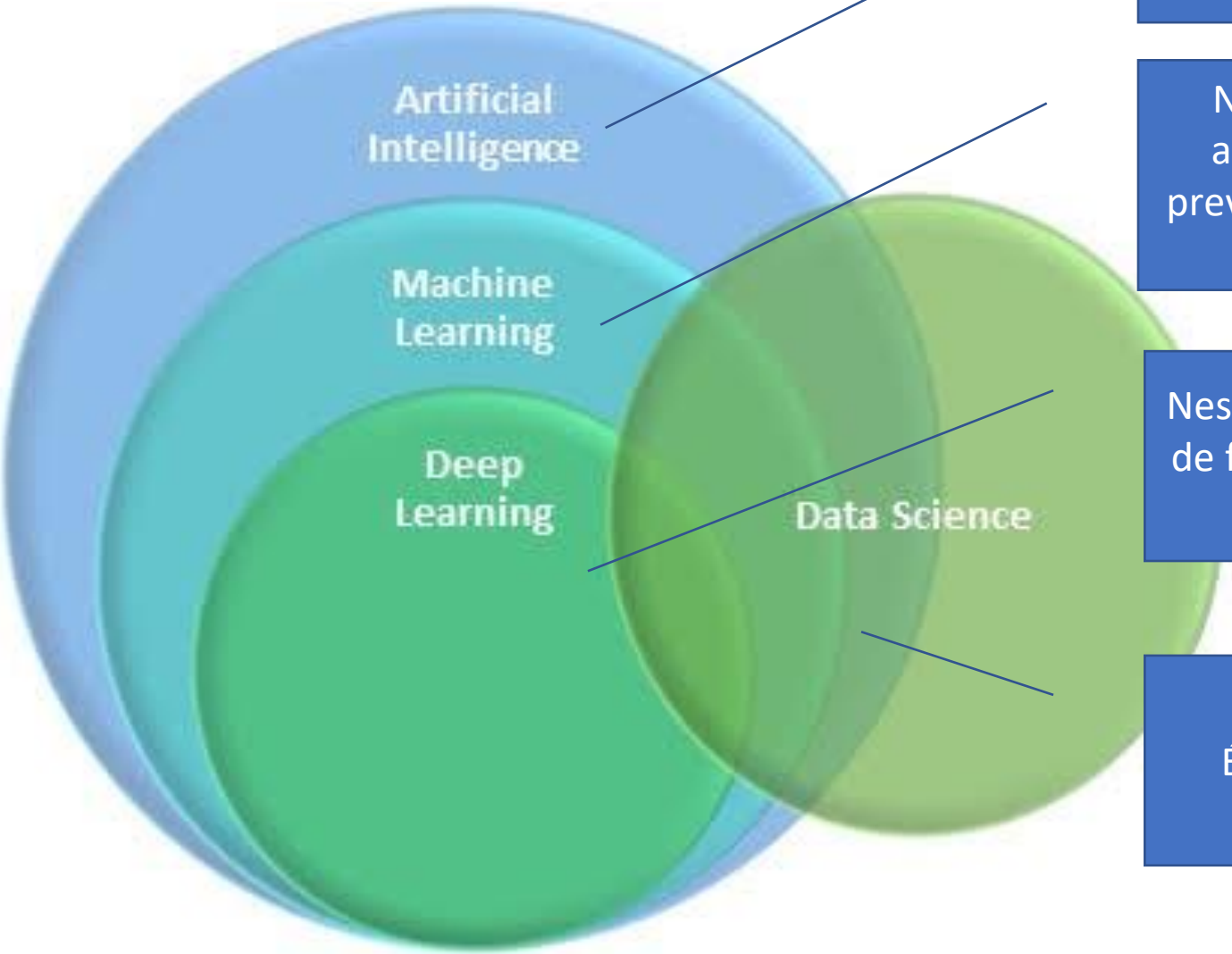
Supervisora: Dra. Sirlei Lemes (UFU)

Evento online
Data: 20/04/2023 (quinta)
Hora: 8:30 h

Sumário da apresentação

- Introdução
  - *O que é Machine Learning (ML) e inteligência artificial? De onde surgiu? e qual o contexto atual?;*
  - *Machine learning no processo de pesquisa;*
  - *Principais skills do(a) pesquisador(a)*
  - *Visão geral das técnicas de ML;*
- Métricas de pesquisas na área contábil
  - *Pesquisa bibliométrica com 289 papers na base Scopus e Web of Science;*
- Técnicas de ML: conceitos e exemplos de aplicação em pesquisa de alto impacto:
  - *Métodos supervisionados de ML*
  - *Métodos não supervisionados de ML*
  - *Aprendizagem por esforço em ML*
  - *Métodos em conjunto*
  - *Redes neurais*
  - *Deep learning*
- Principais observações em ML
- Conclusões (reflexão)

Machine learning: uma subdivisão da inteligência artificial



A maquina executa comandos pré-programados por humanos

Nesse nível a máquina aprende com dados e com esse aprendizado é capaz de classificar novos objetos, fazer previsões, separar grupos, identificar sequências etc. Ótimo para textos e números.

Nesse nível a máquina tenta simular o complexo mecanismo de funcionamento do cérebro humano. Ótimo para grandes dados (big data), fotos e vídeos.

É a parte da estatística e probabilidade utilizada na IA.

De onde surgiu o termo “inteligência artificial” e sua evolução?

1º artigo Alan Turing 1950, *Computing Machinery and Intelligence.* Questão fundamental: máquinas podem pensar?

Em 1956, o prof. *John McCarthy da* Dartmouth College(New Hampshire) convidou vários cientistas para um Brainstorm de 8 semanas sobre IA. Foi a primeira vez que a palavra apareceu.

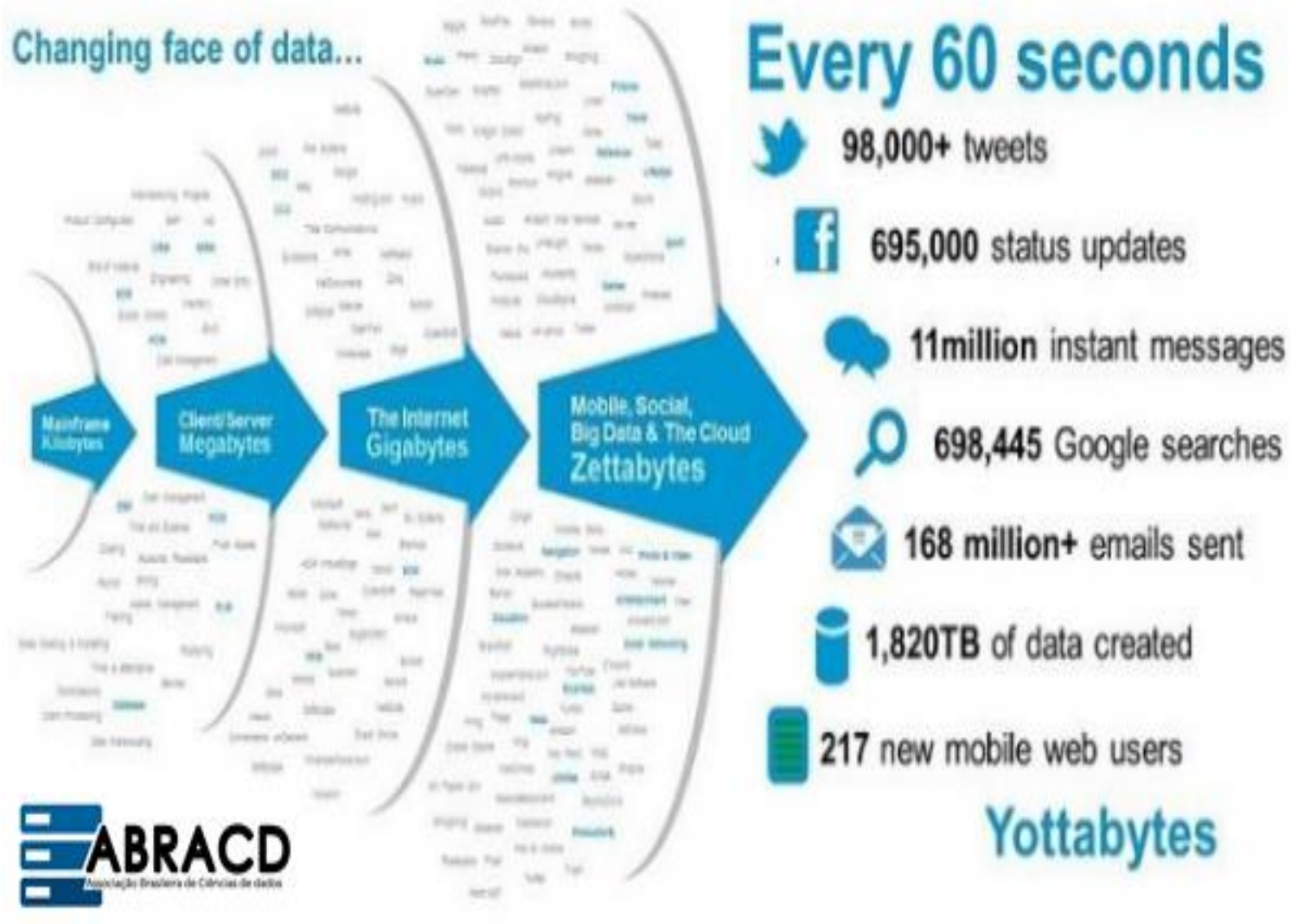
1º chatbot (robôs de conversação) criado por Joseph Weizenbaum em 1964-66 no laboratório de Inteligência Artificial do MIT

1997: O computador Deep Blue, da IBM, derrota o campeão mundial de xadrez Garry Kasparov.

2020: O chat GPT-3, desenvolvido pela OpenAI, é o mais avançado já criado, gera textos quase humano.

Já que o tema é tão antigo, por que todas as pessoas estão falando sobre isso hoje?

- Apesar de ser um tema antigo, essa área se desenvolveu nos últimos tempos, por dois motivos:
  - Barateamento do poder computacional, rapidez de cálculo e armazenamento, (principalmente em nuvem);
  - Quantidade massiva de novos dispositivos que geram dados;

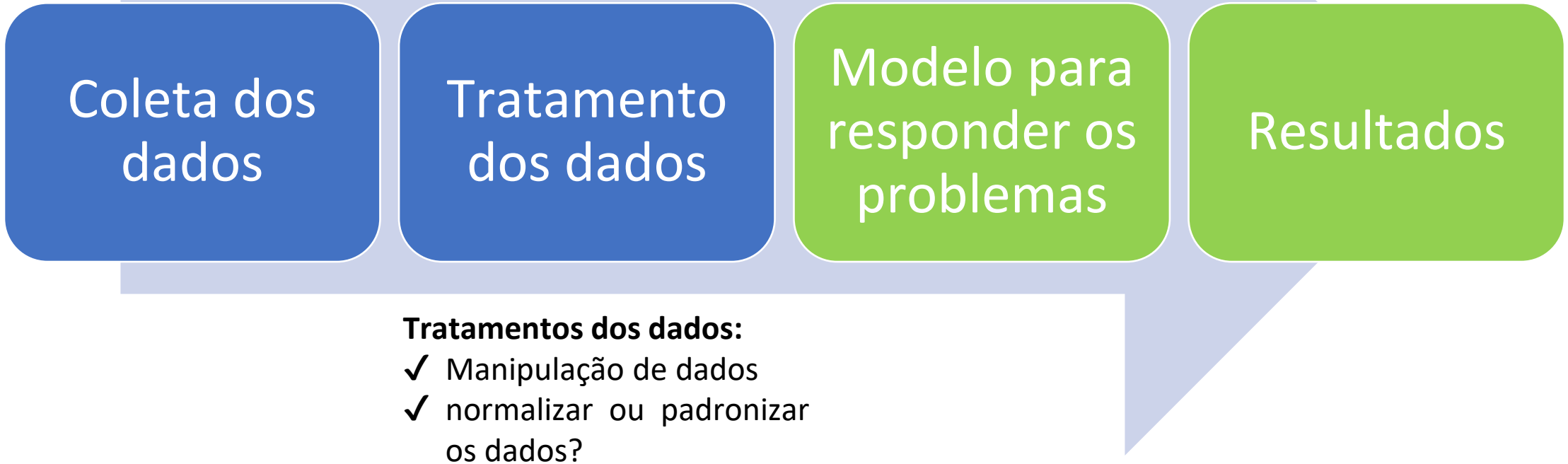


Machine learning no processo de pesquisa

Tipos de Dados:
- ✓ Estruturados
- ✓ Não estruturados
- ✓ Semi estruturados

Aqui entra o ML...

Pode ser a **técnica fim da pesquisa ou pode ser um meio para classificar variáveis e assim** rodar em outros modelos estatísticos.



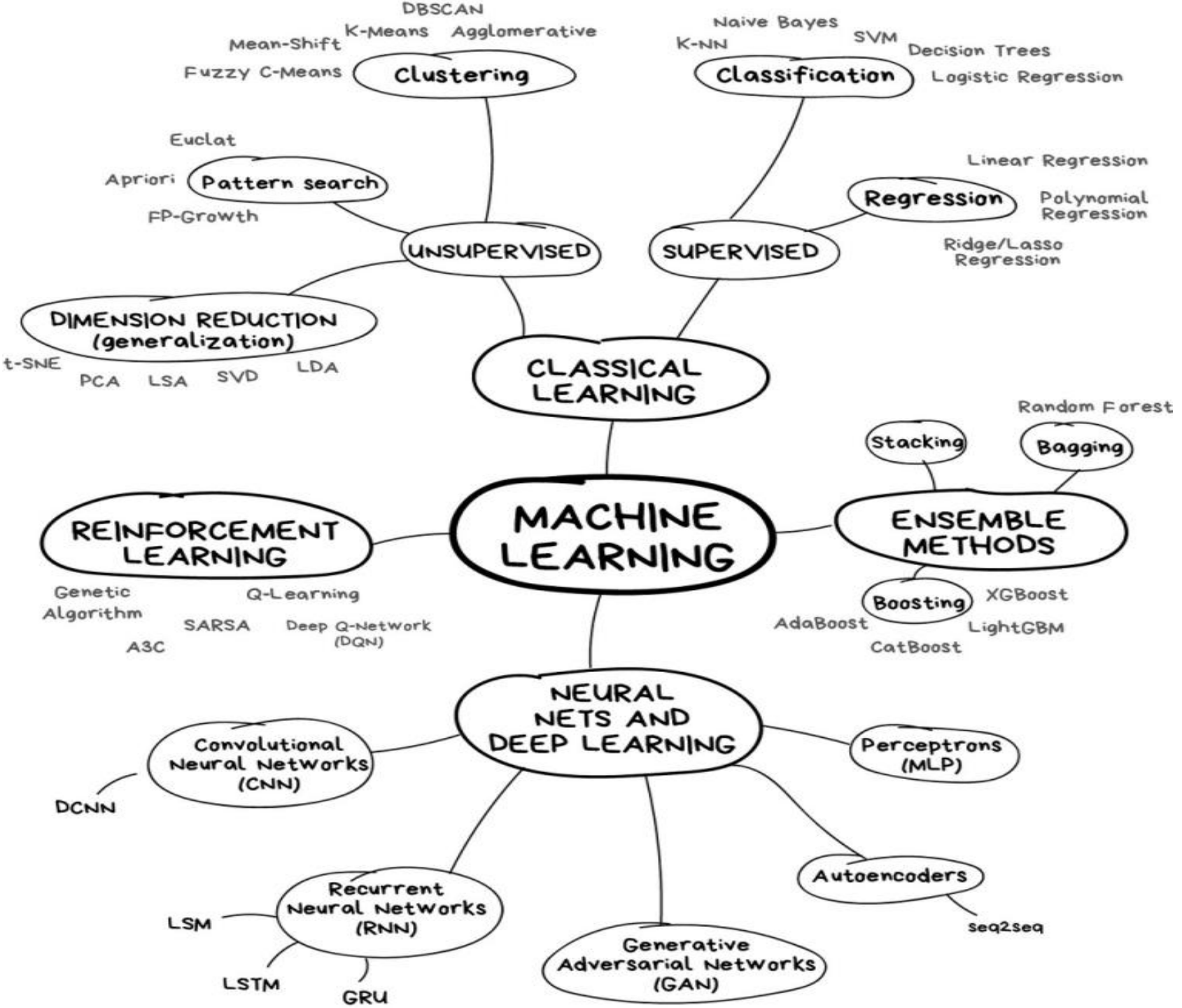
Tratamentos dos dados:
- ✓ Manipulação de dados
- ✓ normalizar ou padronizar os dados?
- ✓ Outliers
- ✓ missing

Principais *skills* do pesquisador

- Conhecimento em linguagens de programação. As mais comuns: R e Python.

- De modo geral o python é mais utilizado em *machine learning*, porém na parte de *text Mining,* o R é mais utilizado .

Visão geral das divisões do ML

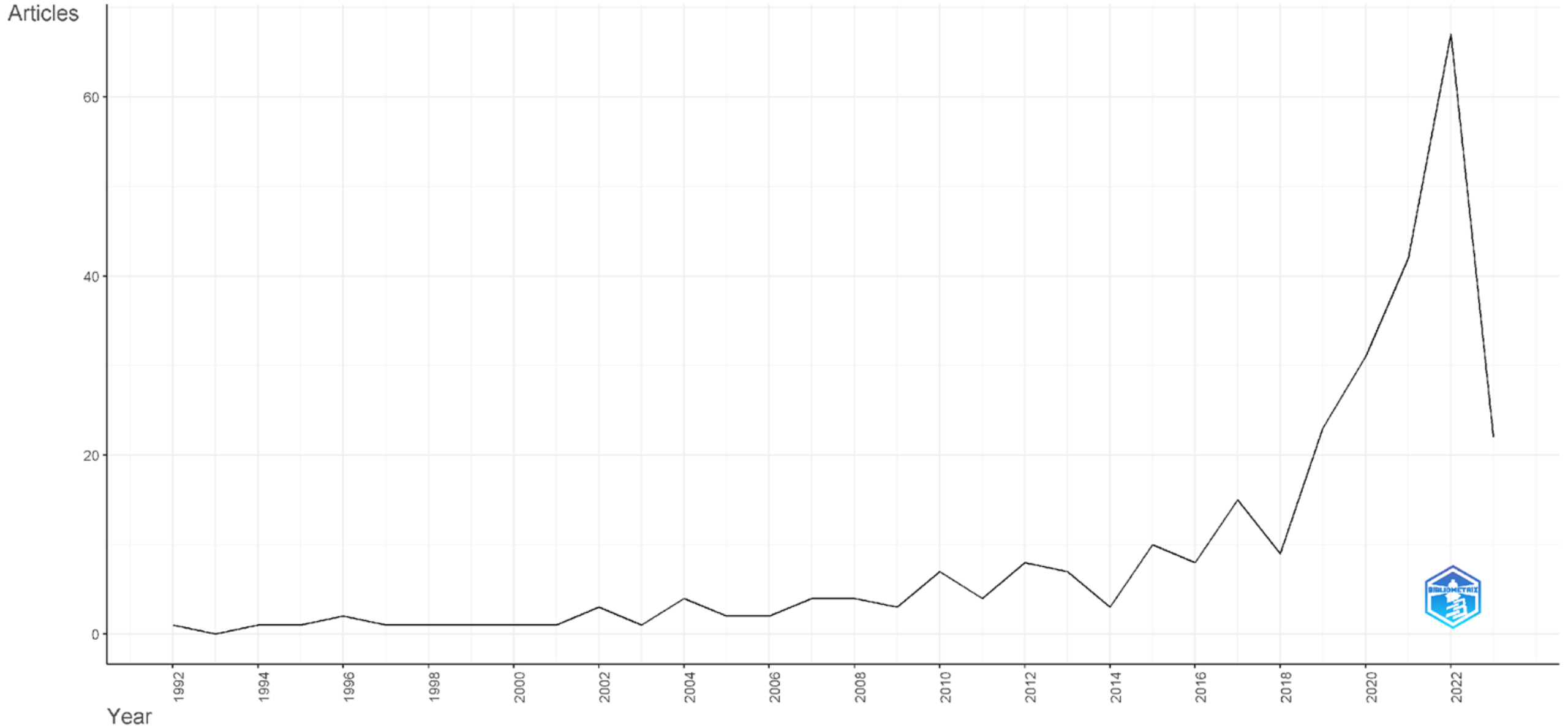


Pesquisas sobre o tema *machine learning* e pesquisas na área contábil

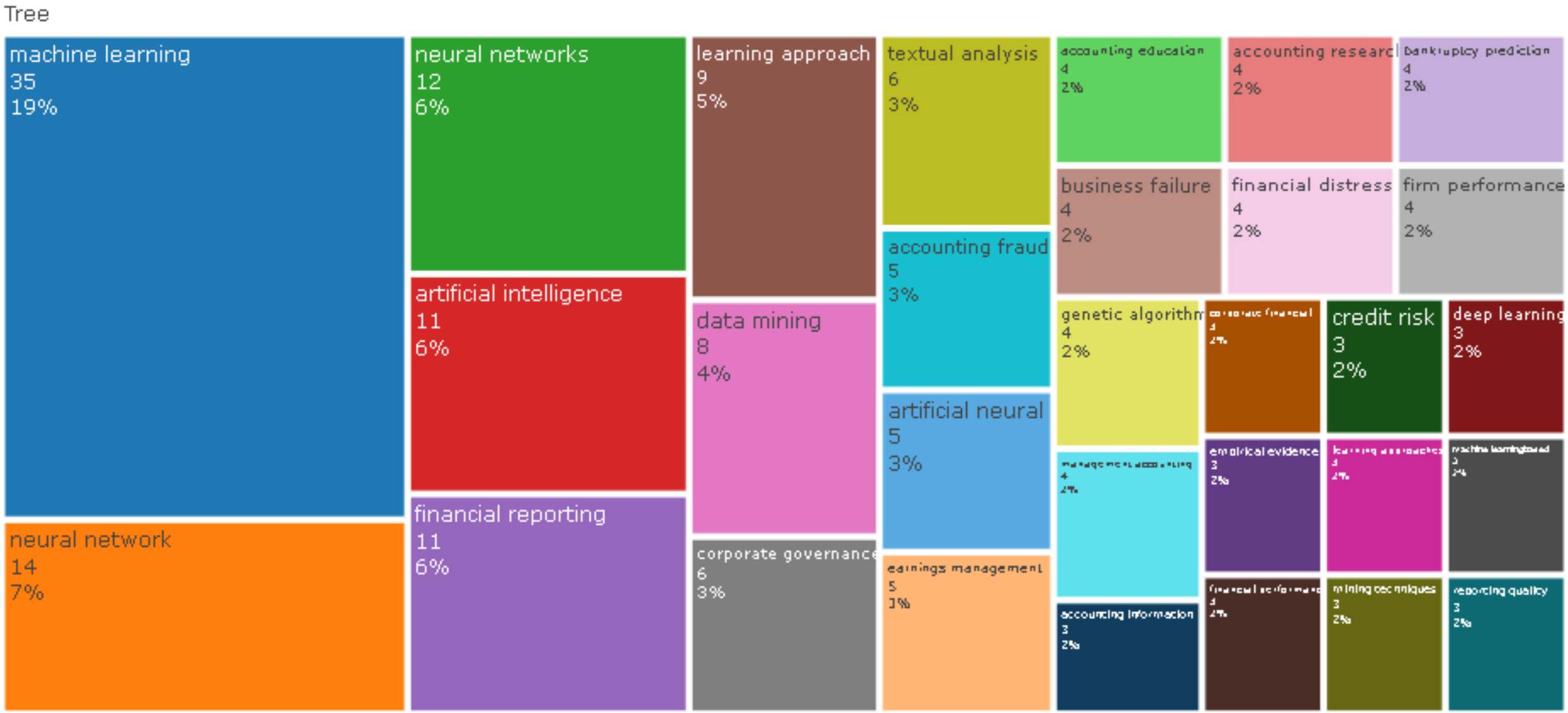
- **Base:**  Scopus® **e** Clarivate Web of Science™
- **Filtro palavras =** “machine learning” “accounting”
- **Filtro data =** não especificado
- **Ferramenta =** BIBLIOMETRIX → R
- 289 *papers* reportados

Annual Scientific Production

289 Papers sobre o tema

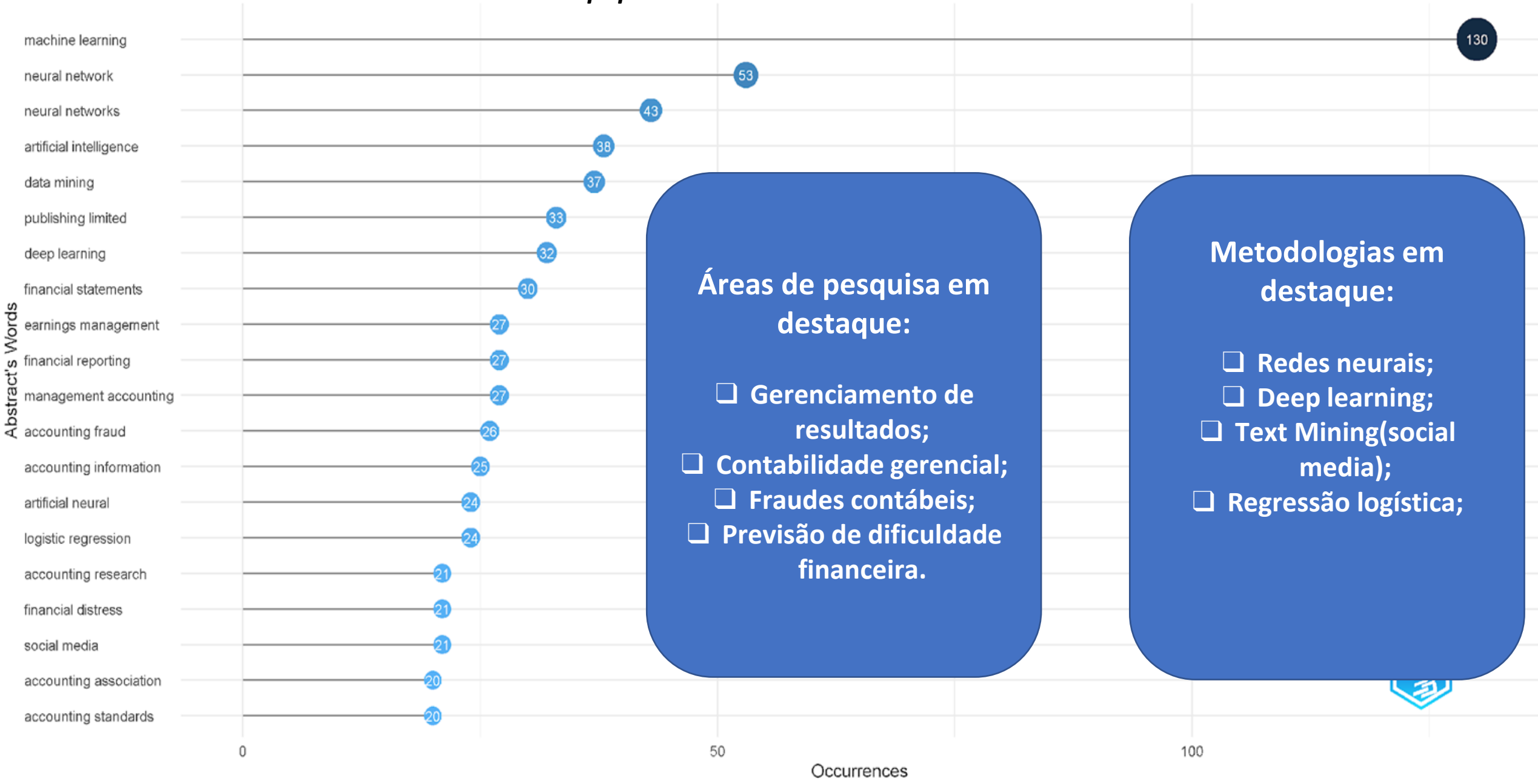


Frequência de títulos dos 289 artigos



Most Relevant Words

Abstracts – 289 papers



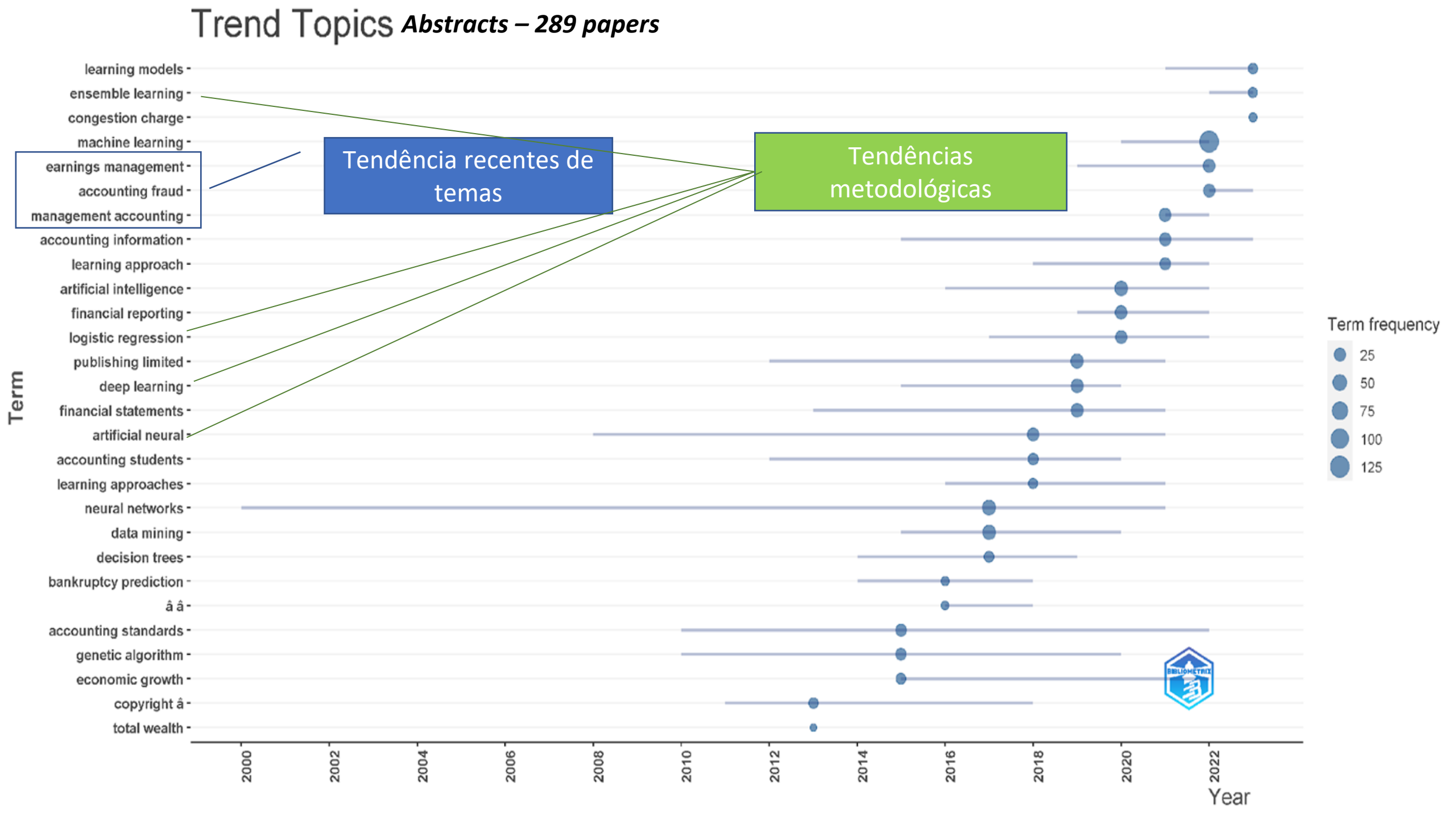
Áreas de pesquisa em destaque:

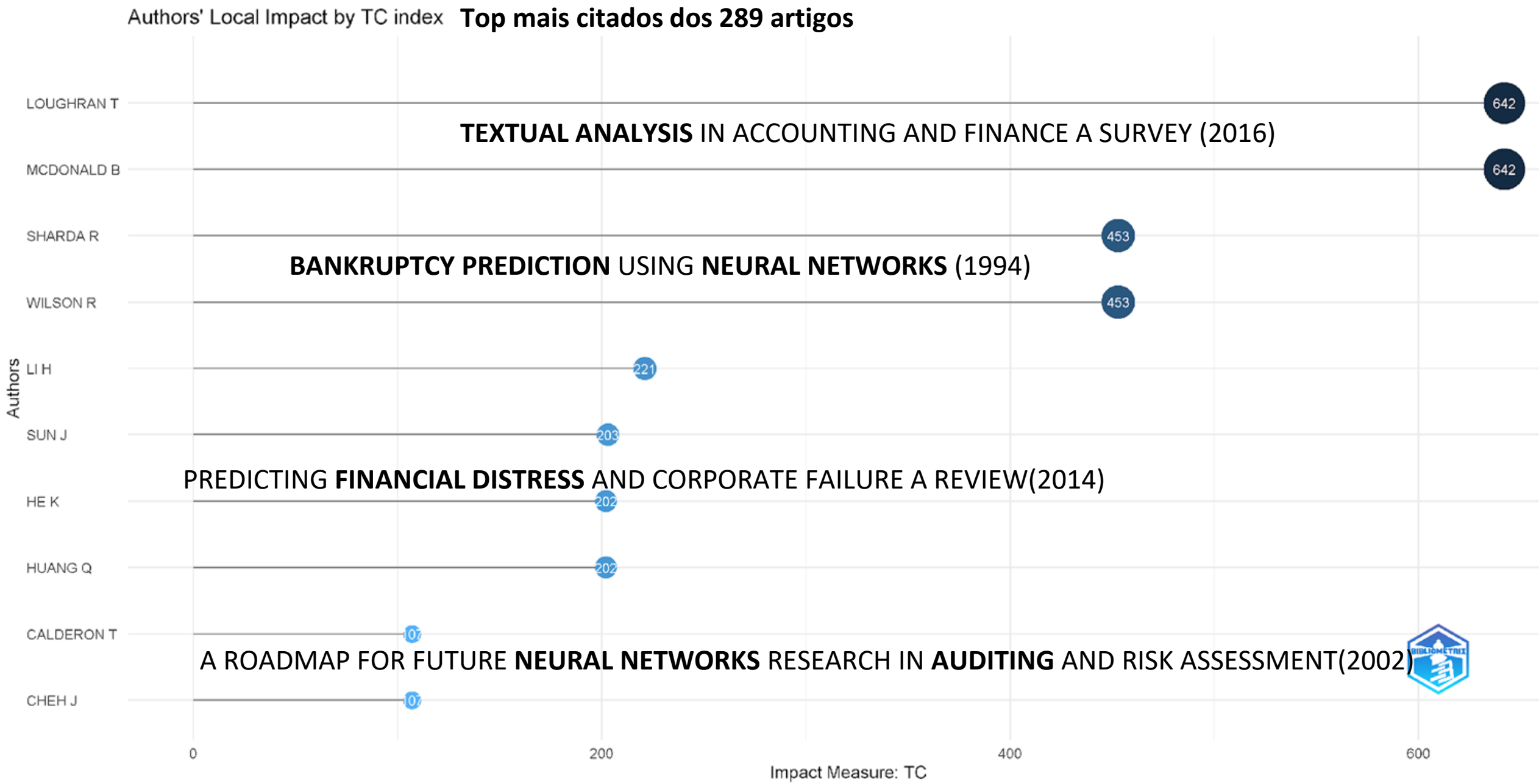
- ❑ Gerenciamento de resultados;
- ❑ Contabilidade gerencial;
- ❑ Fraudes contábeis;
- ❑ Previsão de dificuldade financeira.

Metodologias em destaque:

- ❑ Redes neurais;
- ❑ Deep learning;
- ❑ Text Mining(social media);
- ❑ Regressão logística;

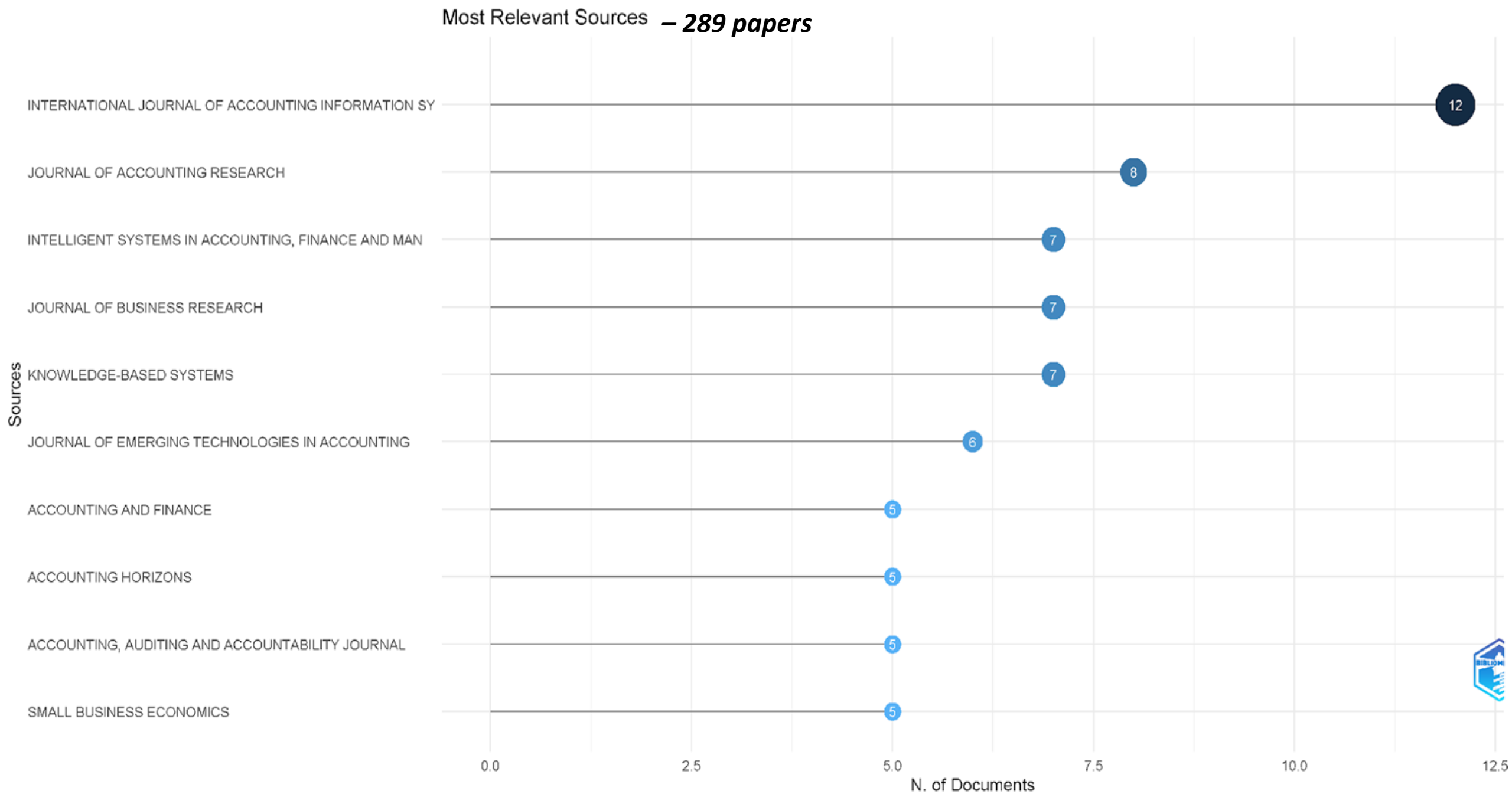
Trend Topics *Abstracts – 289 papers*



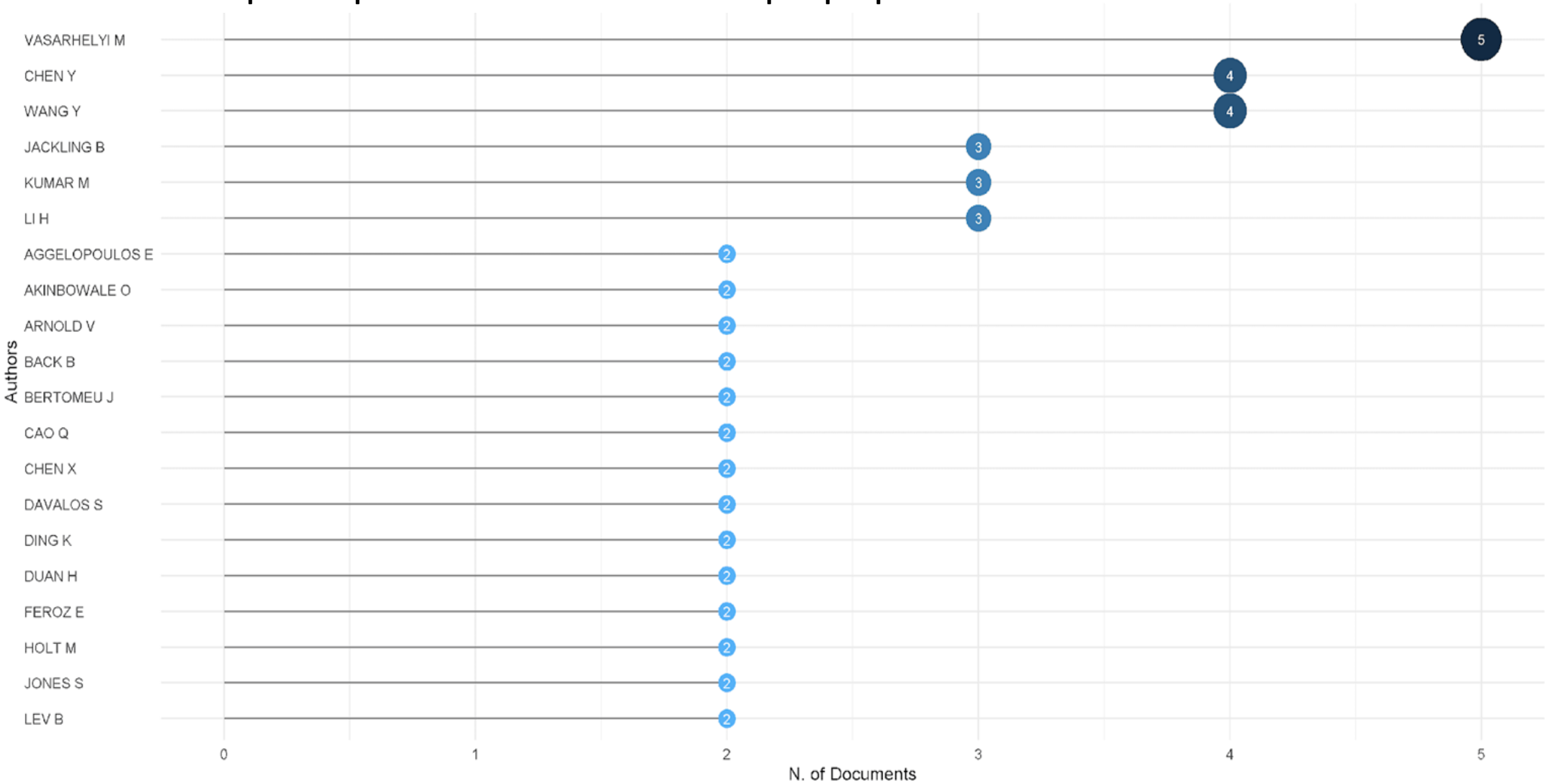

Authors' Local Impact by TC index
Top mais citados dos 289 artigos
LOUGHRAN T
642
TEXTUAL ANALYSIS IN ACCOUNTING AND FINANCE A SURVEY (2016)
MCDONALD B
642
SHARDA R
453
BANKRUPTCY PREDICTION USING NEURAL NETWORKS (1994)
WILSON R
453
LI H
221
Authors
SUN J
203
PREDICTING FINANCIAL DISTRESS AND CORPORATE FAILURE A REVIEW(2014)
HE K
202
HUANG Q
202
CALDERON T
107
A ROADMAP FOR FUTURE NEURAL NETWORKS RESEARCH IN AUDITING AND RISK ASSESSMENT(2002)
CHEH J
107
0
200
400
600
Impact Measure: TC

10 referências mais citadas pelos 289 artigos – útil para encontrar artigos seminais

Google Scholar	Cited References	Citations	
link	ALTMAN, E.I., FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY (1968) THE JOURNAL OF FINANCE, 23 (4), PP. 589-609	8	Previsão de falência
link	BENEISH, M.D., THE DETECTION OF EARNINGS MANIPULATION (1999) FINANCIAL ANALYSTS JOURNAL, 55 (5), PP. 24-36	7	Gerenciamento de resultados
link	BEAVER, W.H., FINANCIAL RATIOS AS PREDICTORS OF FAILURE (1966) JOURNAL OF ACCOUNTING RESEARCH, 4, PP. 71-111	6	Previsão de falência
link	BREIMAN, L., RANDOM FORESTS (2001) MACHINE LEARNING, 45 (1), PP. 5-32	6	
link	DECHOW, P.M., GE, W., LARSON, C.R., SLOAN, R.G., PREDICTING MATERIAL ACCOUNTING MISSTATEMENTS (2011) CONTEMPORARY ACCOUNTING RESEARCH, 28 (1), PP. 17-82	6	Fraudes contábeis
link	JENSEN MC, 1976, J FINANC ECON, V3, P305, DOI 10.1016/0304-405X(76)90026-X	6	
link	LI, F., TEXTUAL ANALYSIS OF CORPORATE DISCLOSURES: A SURVEY OF THE LITERATURE (2010) JOURNAL OF ACCOUNTING LITERATURE, 29, PP. 143-165	6	
link	PEROLS, J., FINANCIAL STATEMENT FRAUD DETECTION: AN ANALYSIS OF STATISTICAL AND MACHINE LEARNING ALGORITHMS (2011) AUDITING: A JOURNAL OF PRACTICE & THEORY, 30 (2), PP. 19-50	5	Fraudes contábeis
link	APPELBAUM, D., KOGAN, A., VASARHELYI, M., YAN, Z., IMPACT OF BUSINESS ANALYTICS AND ENTERPRISE SYSTEMS ON MANAGERIAL ACCOUNTING (2017) INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS, 25, PP. 29-44	4	Contabilidade Gerencial
link	BAO, Y., KE, B., LI, B., YU, Y.J., ZHANG, J., DETECTING ACCOUNTING FRAUD IN PUBLICLY TRADED US FIRMS USING A MACHINE LEARNING APPROACH (2020) JOURNAL OF ACCOUNTING RESEARCH, 58 (1), PP. 199-235	4	Fraudes contábeis

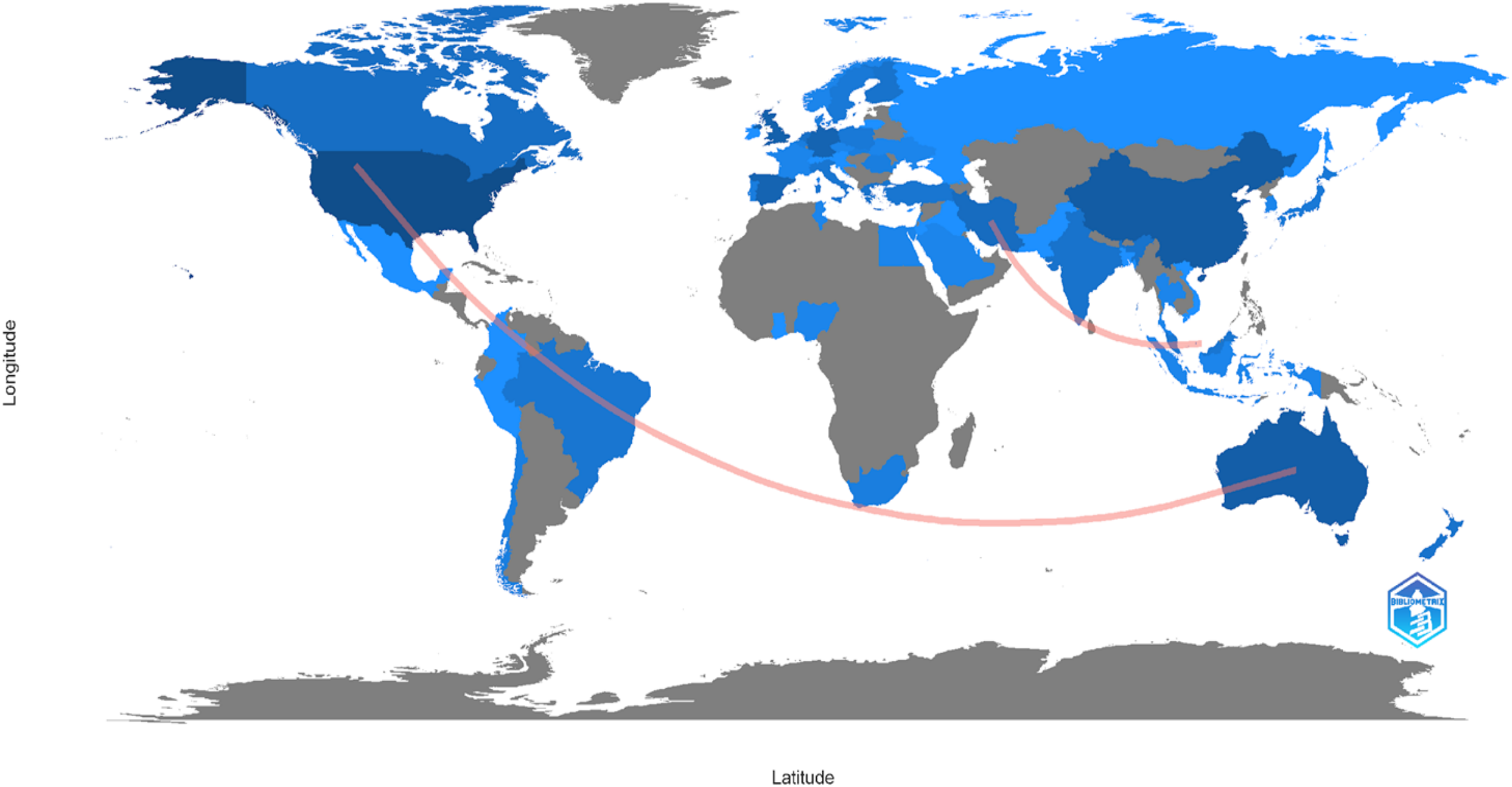

Most Relevant Sources – 289 papers
INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SY
JOURNAL OF ACCOUNTING RESEARCH
INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MAN
JOURNAL OF BUSINESS RESEARCH
KNOWLEDGE-BASED SYSTEMS
JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING
ACCOUNTING AND FINANCE
ACCOUNTING HORIZONS
ACCOUNTING, AUDITING AND ACCOUNTABILITY JOURNAL
SMALL BUSINESS ECONOMICS
Sources
12
8
7
7
7
6
5
5
5
5
0.0
2.5
5.0
7.5
10.0
12.5
N. of Documents

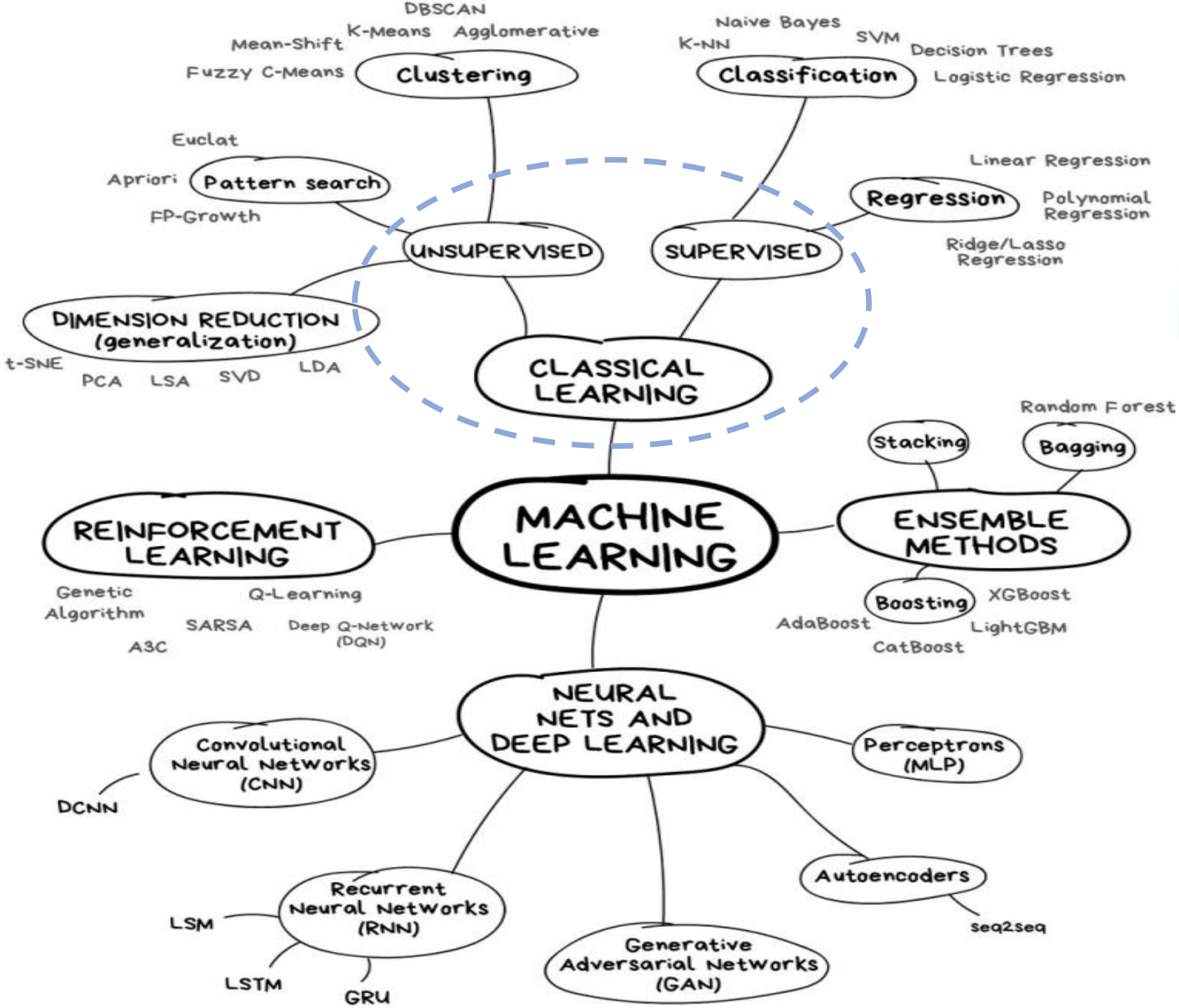
Autores que mais publicaram sobre o tema – 289 Papers pesquisados



Country Collaboration Map

289 Papers pesquisados





Vamos ver quais as diferenças entre um modelo supervisionado e não supervisionado em ML clássico

ML clássicos

- **Supervisionado:** Nos modelos supervisionados o banco de dados já possui um rótulo (para pesquisadores isso é a variável dependente y) e as *features (variáveis independentes x)**. Dessa forma o aprendizado se dá pela observação das *features* e as respostas já colocadas nos rótulos. Exemplo:

Features			***Resposta***
Ações	**Quanti no conselho**	**Free Float %**	**Governança?**
ON	5	25	Sim
PN	4	20	Não
ON	7	30	Sim
ON e PN	4	25	Não

A máquina aprende com as respostas para depois fazer previsões sem a resposta.

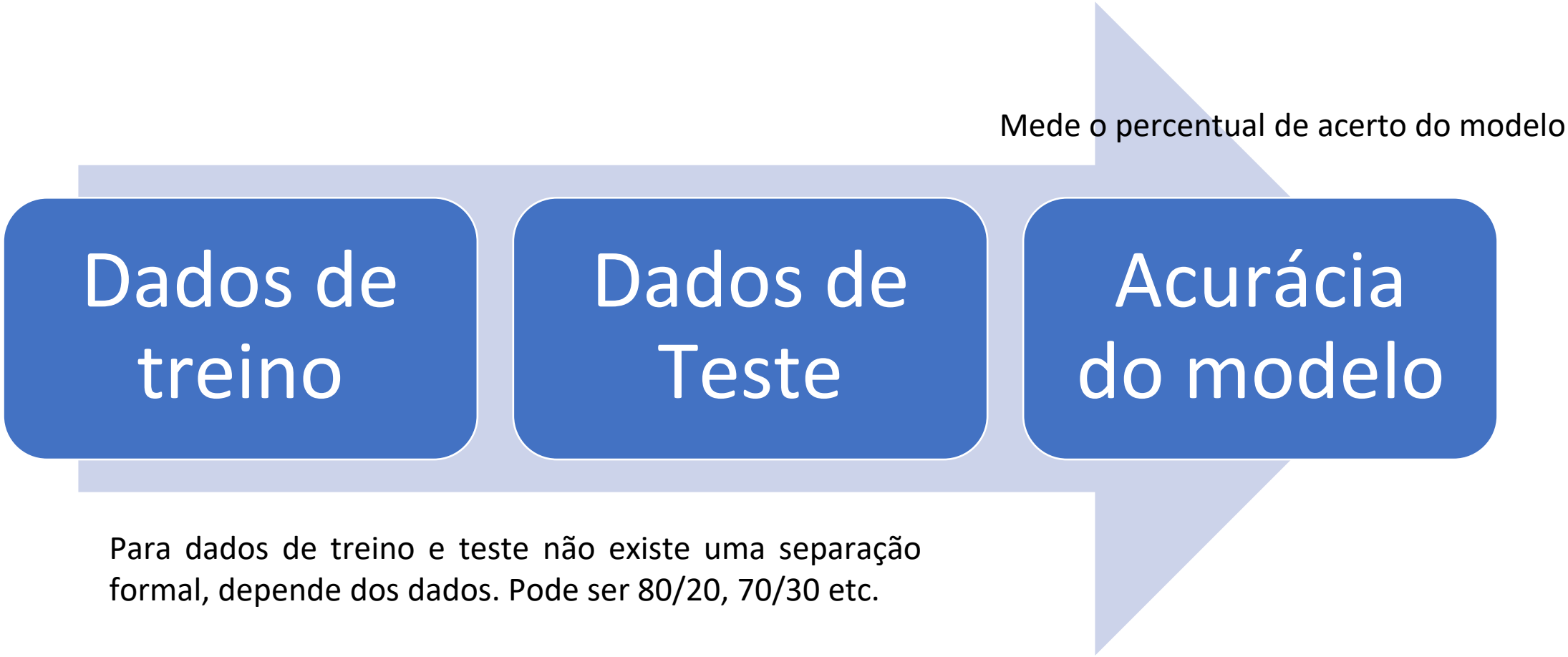
Pergunta: Máquina, se uma empresa tiver PN, 3 conselheiros e 23% *Free Float* ela tem governança?

- **Não supervisionado:** dados não possuem nenhum tipo de rótulo (respostas).

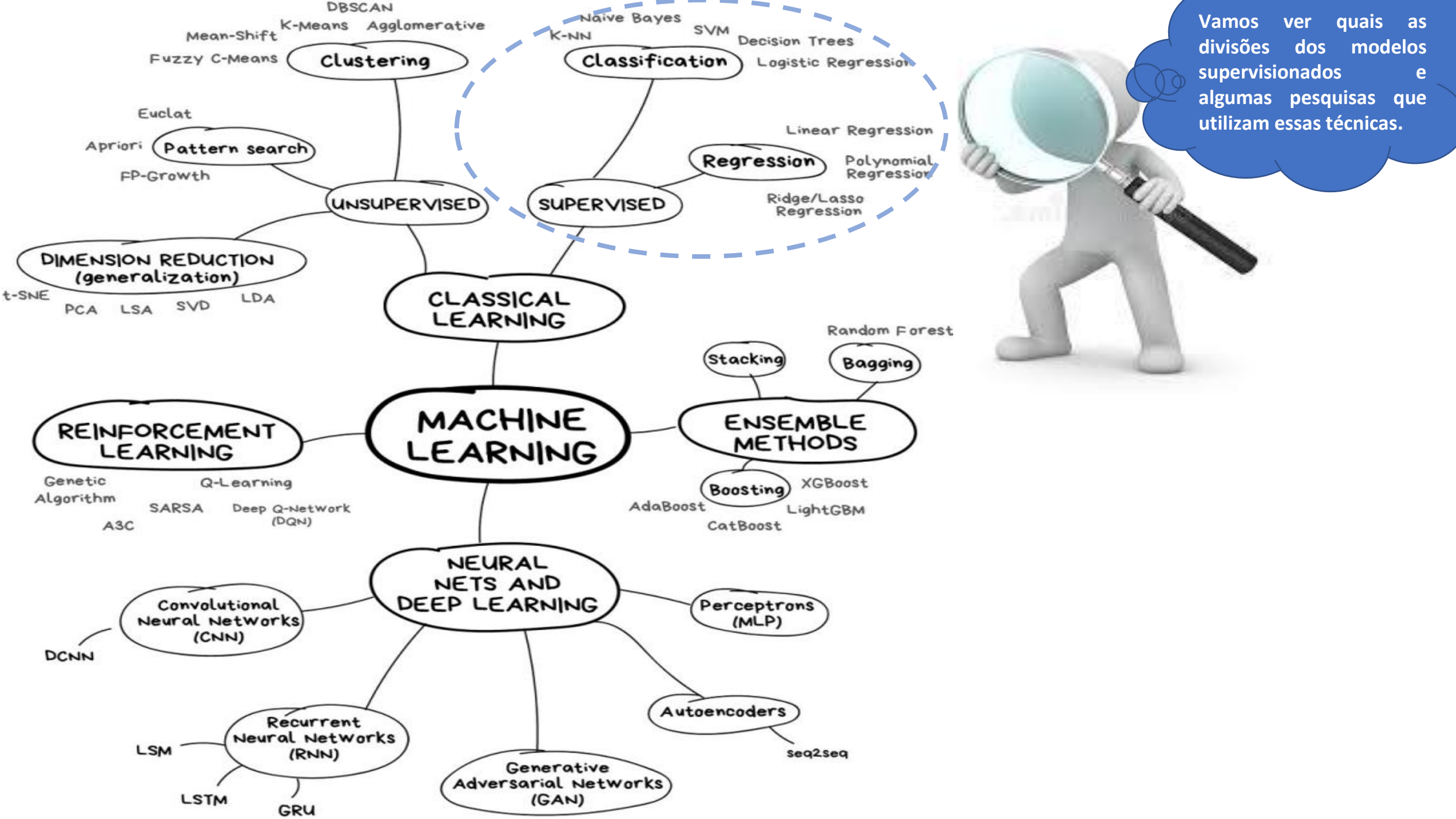
Features			***Separação***
Ações	**Quanti no conselho**	**Free Float %**	**grupo**
ON	5	25	Grupo 1
PN	4	20	Grupo 2
ON	7	30	Grupo 1
ON e PN	4	25	Grupo 2

A máquina observa os dados e os separam em grupos de acordo com as características (*features*).

Fluxo dos modelos supervisionados



Para dados de treino e teste não existe uma separação formal, depende dos dados. Pode ser 80/20, 70/30 etc.

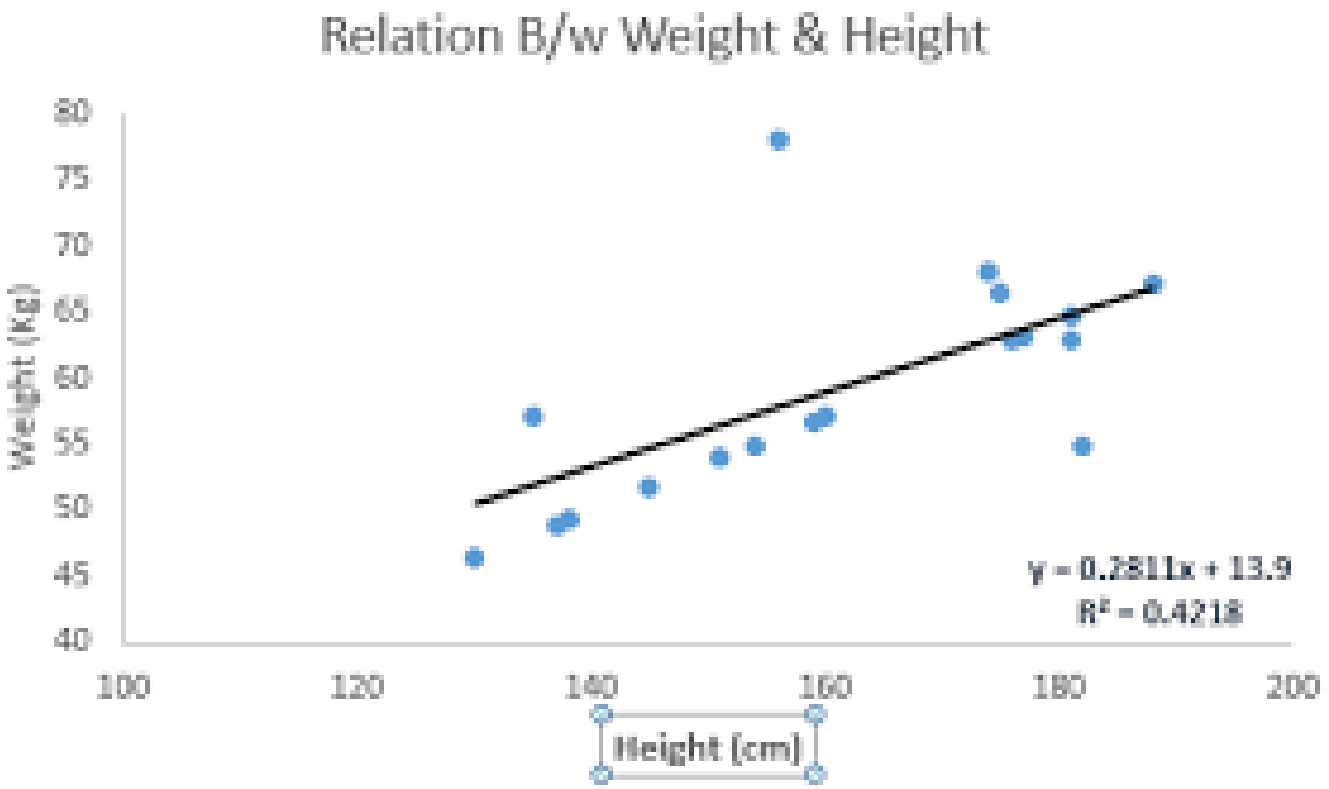


Vamos ver quais as divisões dos modelos supervisionados e algumas pesquisas que utilizam essas técnicas.

Modelos supervisionados

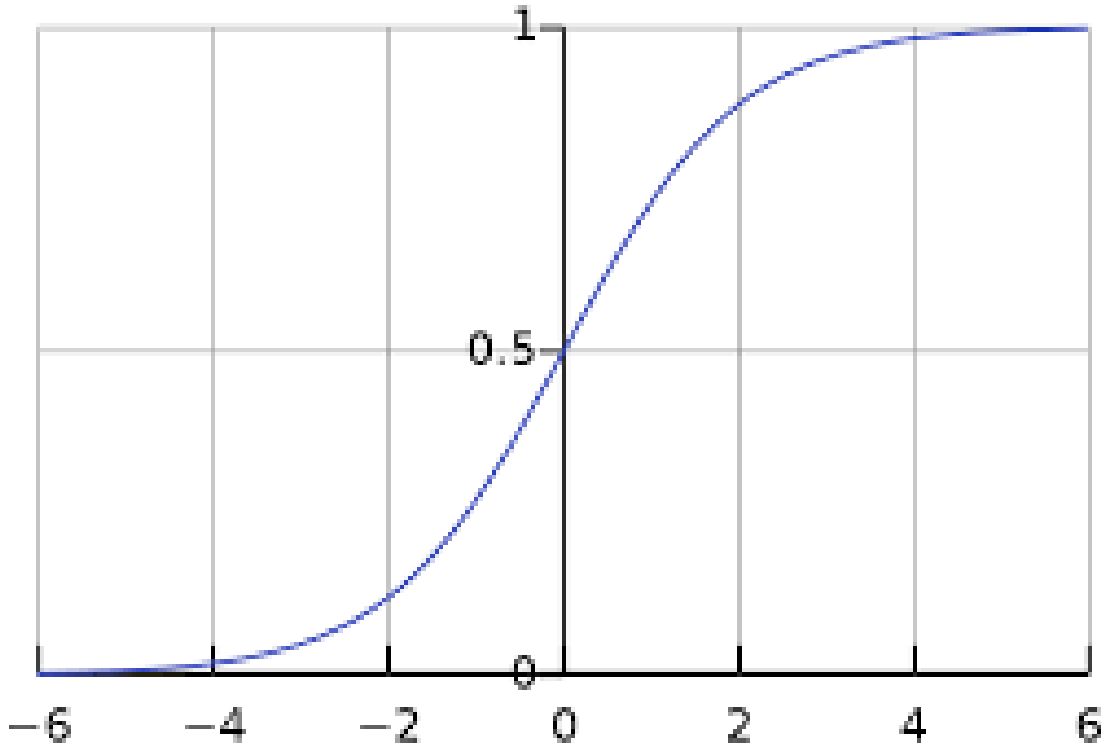
- **Regressão:** utilizada quando a variável dependente (rótulo) é uma variável quantitativa (contínua ou discreta).
  - **Principais <u>algoritmos</u>:** *regressão linear*.
  - **Utilização em pesquisas:** previsão de preço de ações, retorno etc.
- **Classificação:** utilizada quando a variável dependente (rótulo) é uma variável binária, 0 ou 1. A máquina aprende a classificar coisas com base nas *features.*
  - **Principais <u>algoritimos</u>:** *naive bayes, support vector machine, decision tree, k-nearest Neighbours, logistic regression.*
  - **Utilização em pesquisas:** análise de sentimentos textuais, detecção de fraudes etc.

Algoritmo regressão



- Muito utilizado em problemas de previsão;
- As primeiras ideias vieram de Galton no século XIX;
- Na fase de treino a máquina estima a reta que melhor ajusta os pontos;
- Depois de estimado o modelo, agora é possível fazer uma estimação do y levando em consideração a inclinação da reta e o intercepto.
- Algorítmo útil para quaisquer correlações lineares. Os mais utilizados são a regressão linear e a polinomial.
- **Objetivo:** minimizar o erro de previsão (uma das métricas utilizadas é o *Root Mean Squared Error* (RMSE));

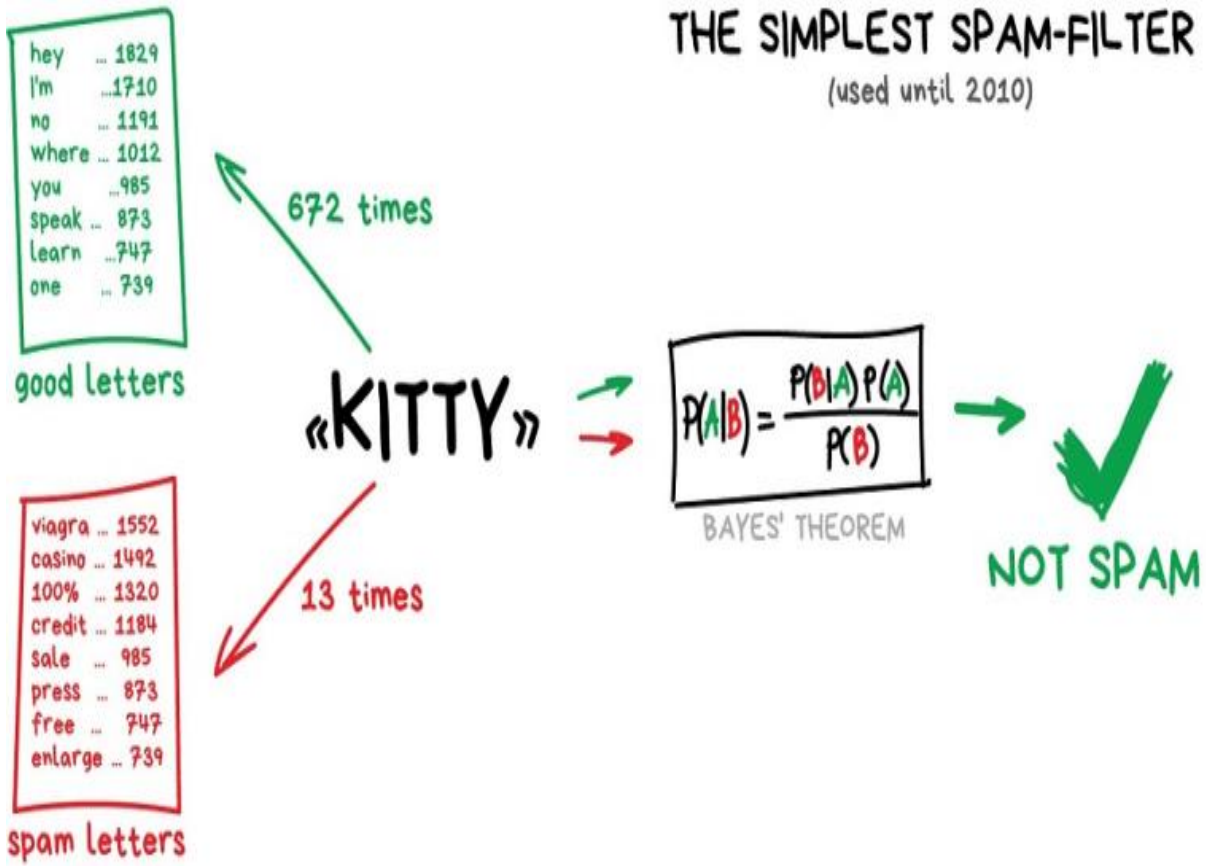
Algoritmo *logistic regression*



- Muito utilizado em problemas de **classificação**;
- Primeiras ideias, Joseph Berkson em 1944 e desenvolvida por David Cox em 1958.
- Na fase de treino a máquina estima a probabilidade de um elemento estar ou não em uma classe;
- Depois de estimar o modelo e avaliar sua acurácia, é possível classificar determinados elementos em novas amostras.

Função sigmoide: $f(x) = \frac{1}{1 + e^{-x}}$

Algoritmo Naive Bayes (NB)



- Utilizado em problemas de **classificação (principalmente textos)**;
- Algoritimo creditado a **Thomas Bayes século 18**, reformulado no trabalho de **John Naive e colaboradores em 1959**;
- **Utilidade:** É usado para classificar documentos em categorias, como spam ou não spam, notícias, esporte etc.. Muito utilizado também para determinar o sentimento de um texto, como positivo, negativo ou neutro.
- **Principais falhas:**
  - naive (ingênuo) assume que as **variáveis são independentes**;
  - Além disso o modelo **não suporta variáveis não lineares**.

Jogar ou não jogar ao ar livre? vamos resolver com o Naive Bayes!

❑ **1º passo:** montar os dados

Aparência	Temperatura(°F)	Umidade	Vento	Jogar
Limpo	Quente	Alta	Fraco	Não
Limpo	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Temperada	Alta	Fraco	Sim
Chuvoso	Fria	Normal	Fraco	Sim
Chuvoso	Fria	Normal	Forte	Não
Nublado	Fria	Normal	Forte	Sim
Limpo	Temperada	Alta	Fraco	Não
Limpo	Fria	Normal	Fraco	Sim
Chuvoso	Temperada	Normal	Fraco	Sim
Limpo	Temperada	Normal	Forte	Sim
Nublado	Temperada	Alta	Forte	Sim
Nublado	Quente	Normal	Fraco	Sim
Chuvoso	Temperada	Alta	Forte	Não

Rótulo (resposta)

Treinamento (aprendizagem)

❑ **2º passo:** frequências relativas dos atributos de cada *feature (*tabela de probabilidades*)*

Aparência	Sim	Não	Temperatura	Sim	Não	Umidade	Sim	Não	Vento	Sim	Não
Limpo	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Fraco	6/9	2/5
Nublado	4/9	0/5	Temperada	4/9	2/5	Normal	6/9	1/5	Forte	3/9	3/5
Chuvoso	3/9	2/5	Fria	3/9	1/5						

Previsão- Aplicação em qualquer situação utilizando a tabela prob.

❑ **3º passo:** depois de treinado, vamos testar o algoritmo na seguinte situação:

➢ **aparência (chuvoso); temperatura (fria); umidade (alta) e vento (forte**

❑ **4º passo:** cálculo da probabilidade e conclusão

L(“sim”) = 3/9*3/9*3/9*3/9*9/14 = 0,00794

L(“não”) = 3/5*1/5*4/5*3/5*5/14 = 0,01371

Conclusão: não haverá jogo

Fonte:Microsoft Word - Dissertação de Mestrado Completa Word - Tobias.docx (ibict.br)

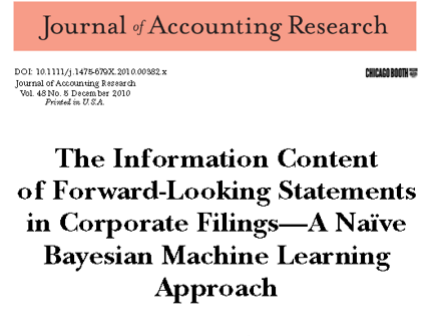
Journal of Accounting Research

DOI: 10.1111/j.1475-679X.2010.00382.x
Journal of Accounting Research
Vol. 48 No. 5 December 2010
Printed in U.S.A.

CHICAGO BOOTH

The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach

FENG LI*



objetivo

O artigo examinou o **conteúdo textual** das declarações prospectivas (FLS) da seção de Discussão e Análise Gerencial (MD&A) dos arquivos 10-K e 10-Q.

Metodologia

Passo 1 – download de todos os 10-K e 10-Q do site de SEC de 1994 a 2007 e separação da seção MD&A;

Passo 2 –Para construir os dados de **treinamento**, classificou manualmente **30.000 FLS** selecionados aleatoriamente em tons positivo(1), neutro(0) ou negativo/incerto(-1);

Passo 3 – depois de treinado o algoritmo **Naive Bayes** fez a classificação de **140.000 FLSs**;

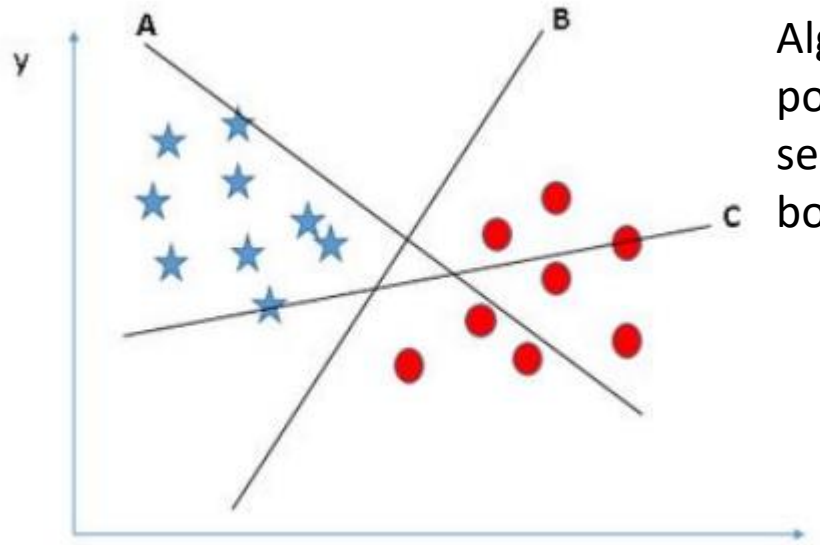
Passo 4 – rodou um modelo de **regressão linear** com a média dos tons sendo a variável dependente.

Observe que o NB não foi objetivo fim da pesquisa e sim um meio para classificar a variável tom

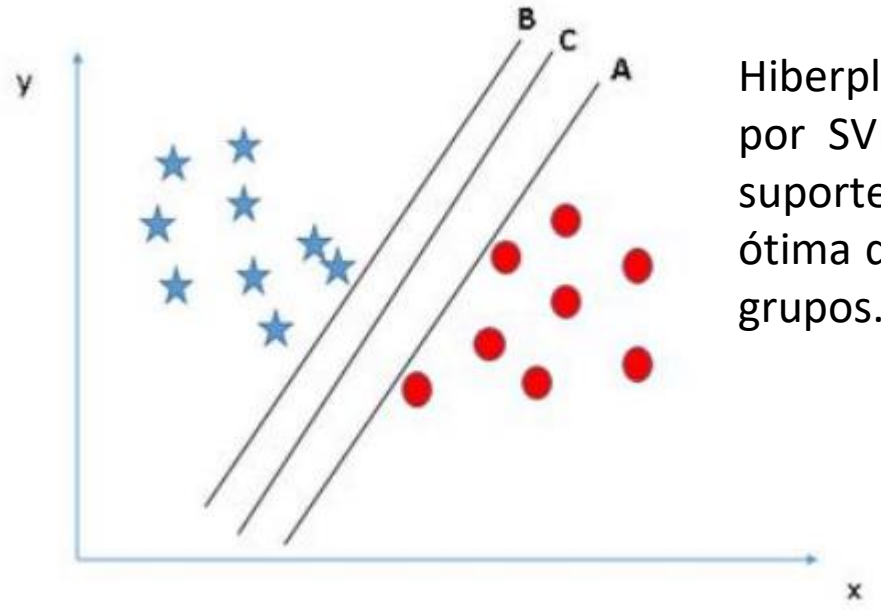
Achados

O tom médio FLS é em função do desempenho atual, *accruals*, tamanho da empresa, índice MTB, volatilidade do retorno e idade da empresa.

Algoritmo Máquina de vetores de suporte (*Supporte Vector Machine* - SVM)



Alguns hiberplanos possíveis para separar estrelas e bolas.

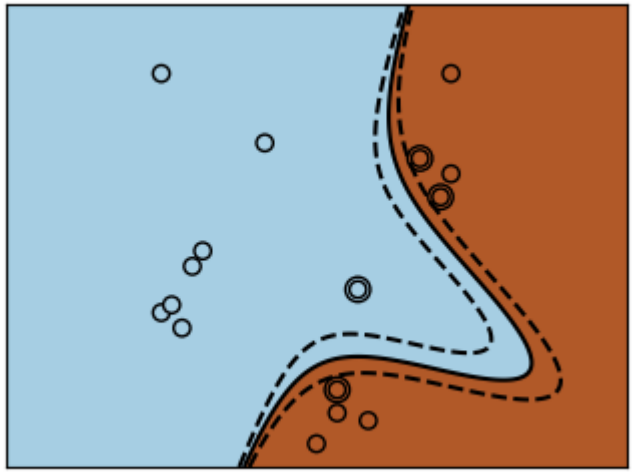


Hiberplanos organizados por SVM. A e B são os suportes e C é a linha ótima que separa os dois grupos.

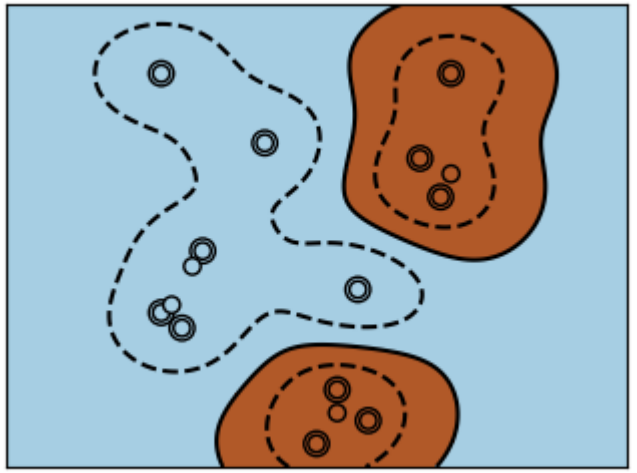
- ❑ Útil para problemas de **classificação**;
- ❑ As primeiras ideias vieram dos Russos **Vapnik e Chervonenkis (1963)**
- ❑ Separa classes por uma linha, pode ser reta ou não;
- ❑ Essa separação pode ter n classificações.
- ❑ **Utilidade:** detecção de fraudes, análise de crédito, previsão de preço das ações, previsão de falências, detecção de anomalias dentre outras.

Algoritmo Suporte Vector Machine (SVM)

SVM-Kernels polynomial



SVM-Kernels RBF

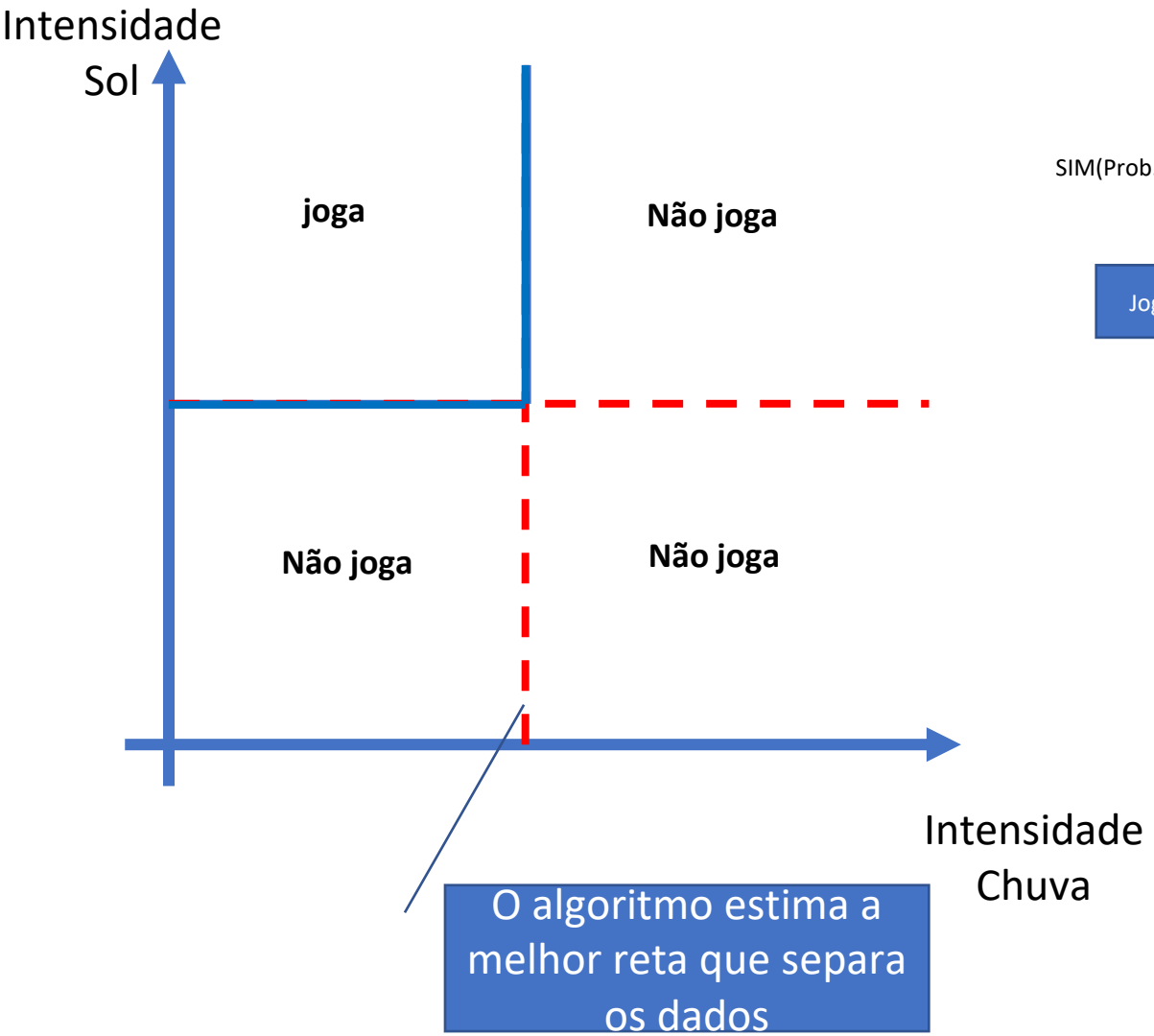


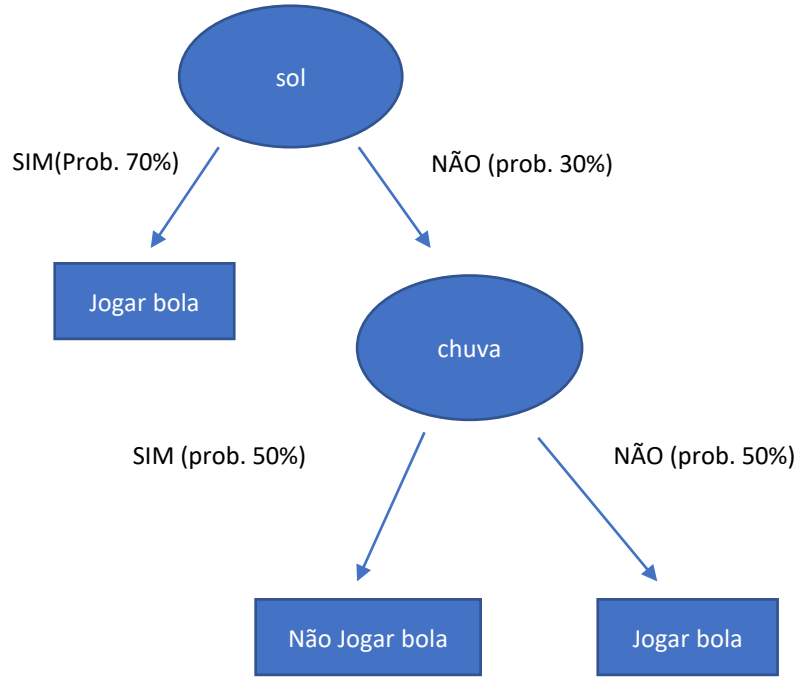
- O super poder desse algorítimo é que ele também é útil para dados **não linearmente separáveis**, podemos utilizar uma função de kernel polinomial ou RBF, como dados das figuras acima.

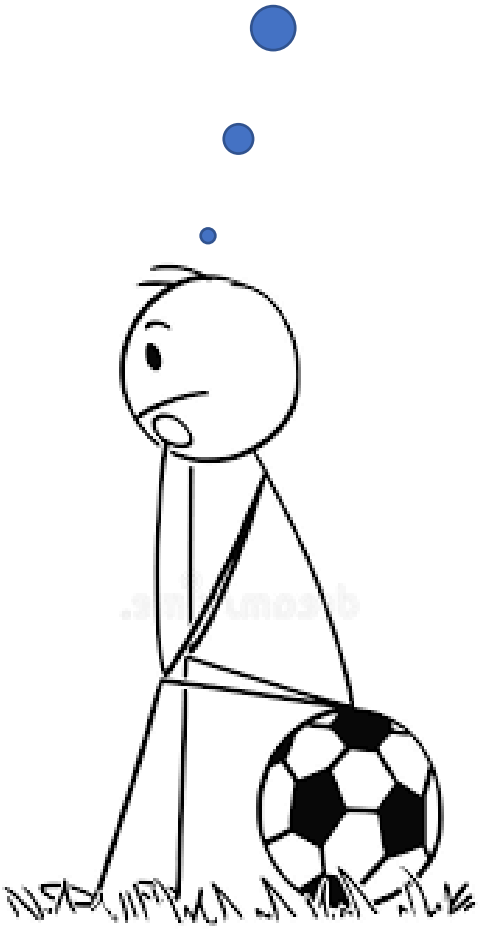
Fonte: SVM-Kernels — scikit-learn 1.2.2 documentation

Algoritmo árvore de decisão (*Decision tree*)

Jogar ou não jogar?



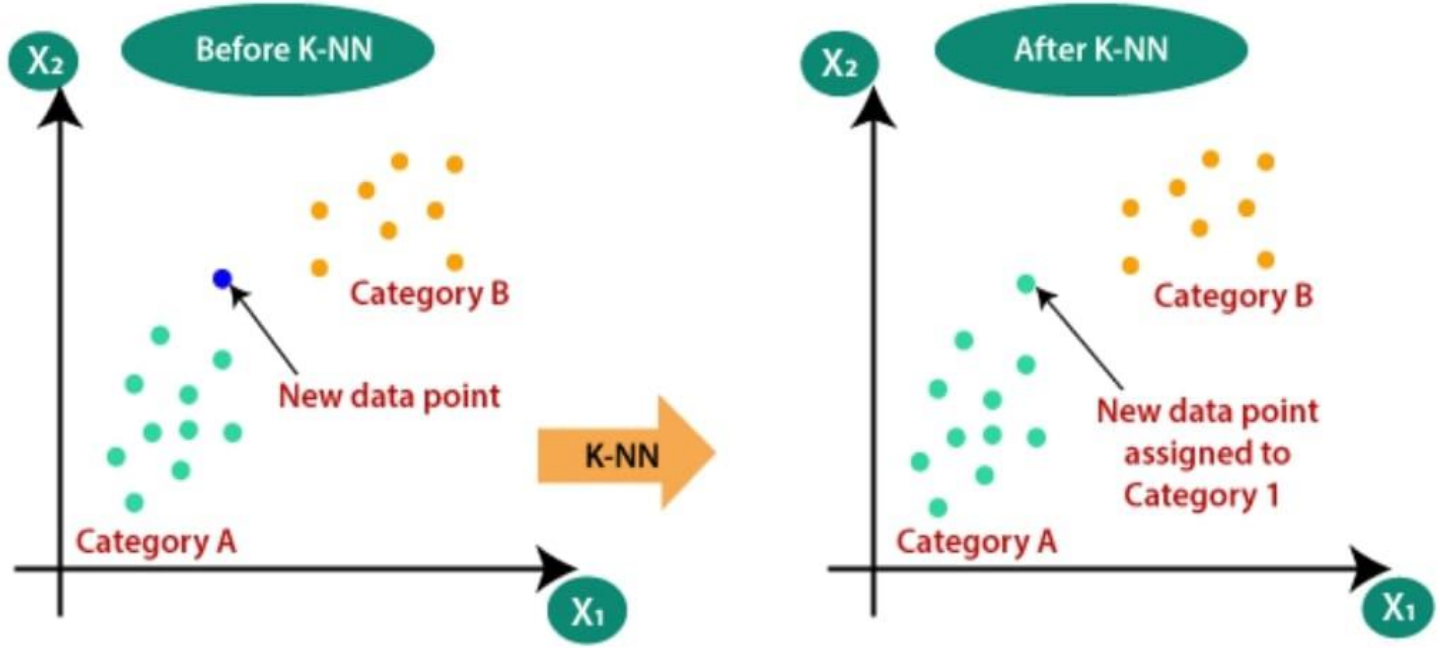




Algoritmo Decision tree

- Útil para problemas de **classificação**;
- introduzida em **1950 por Morgan and Sonquist**, mas como algoritmo foi desenvolvido em **1960 por Moore e Quinlan;**
- O primeiro nó da arvore (nó raiz) inicia com aquelas variáveis que são melhor separáveis. A máquina pode utilizar a **taxa de ganho**, **índice de Gini** ou **entropia** para fazer essa escolha.
- Na entropia a máquina escolhe a ordem dos nós de acordo com a homogeneidade e com menor grau de impureza. No caso o nó Sol, do slide anterior, tem um caminho mais fácil de seguir (menos impuro e homogêneo), prob. 70%/30%, do que o nó chuva, 50%/50%.
- As variáveis que formam a árvore podem ser qualitativa ou quantitativa.
- **Algoritmos mais populares:**
  - D3, C4.5 melhorado para C5.0 (**Quinlan, 1993**),
  - CART (**Breiman et al.,1984**),
  - CHAID (**Kass, 1980**)
  - QUEST. A maior parte deles foram criados nos **anos 80 e 90**.

Algoritmo k-vizinhos próximos (*k-nearest Neighbours)*



- Útil para problemas de **classificação**;
- Ideia inicial é atribuída ao alemão Helmholtz, (século 19), mas o algoritmo foi desenvolvido a partir da década de 50 por vários pesquisadores;
- Calcula as distancias entre os pontos no plano. Essas distancias podem ser calculadas pelo **método Euclidiano, Hamming, Manhattham e Markowski.**
- De acordo com as distâncias o KNN classifica determinada observação naquela categoria.
- O K indica a distância tolerada do novo ponto para classifica-lo em determinada categoria. O k é determinado pelo pesquisador.

Análise de artigo com SVM, Decision Tree e Regressão Logistica

An investigation of the factors influencing cost system functionality using decision trees, support vector machines and logistic regression

Cemil Kuzey
Arthur J. Bauernfeind College of Business, Murray State University, Murray, Kentucky, USA

Ali Uyar
La Rochelle Business School, Excelia Group, La Rochelle, France, and

Dursun Delen
Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA

International Journal of Accounting & Information Management
Vol. 27 No. 1, 2019
pp. 27-55

1834-7649
DOI 10.1108/IJAIM-04-2017-0052

An investigation of the factors influencing cost system functionality using decision trees, support vector machines and logistic regression

Cemil Kuzey
Arthur J. Bauernfeind College of Business, Murray State University, Murray, Kentucky, USA
Ali Uyar
La Rochelle Business School, Excelia Group, La Rochelle, France, and
Dursun Delen
Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA

objetivo

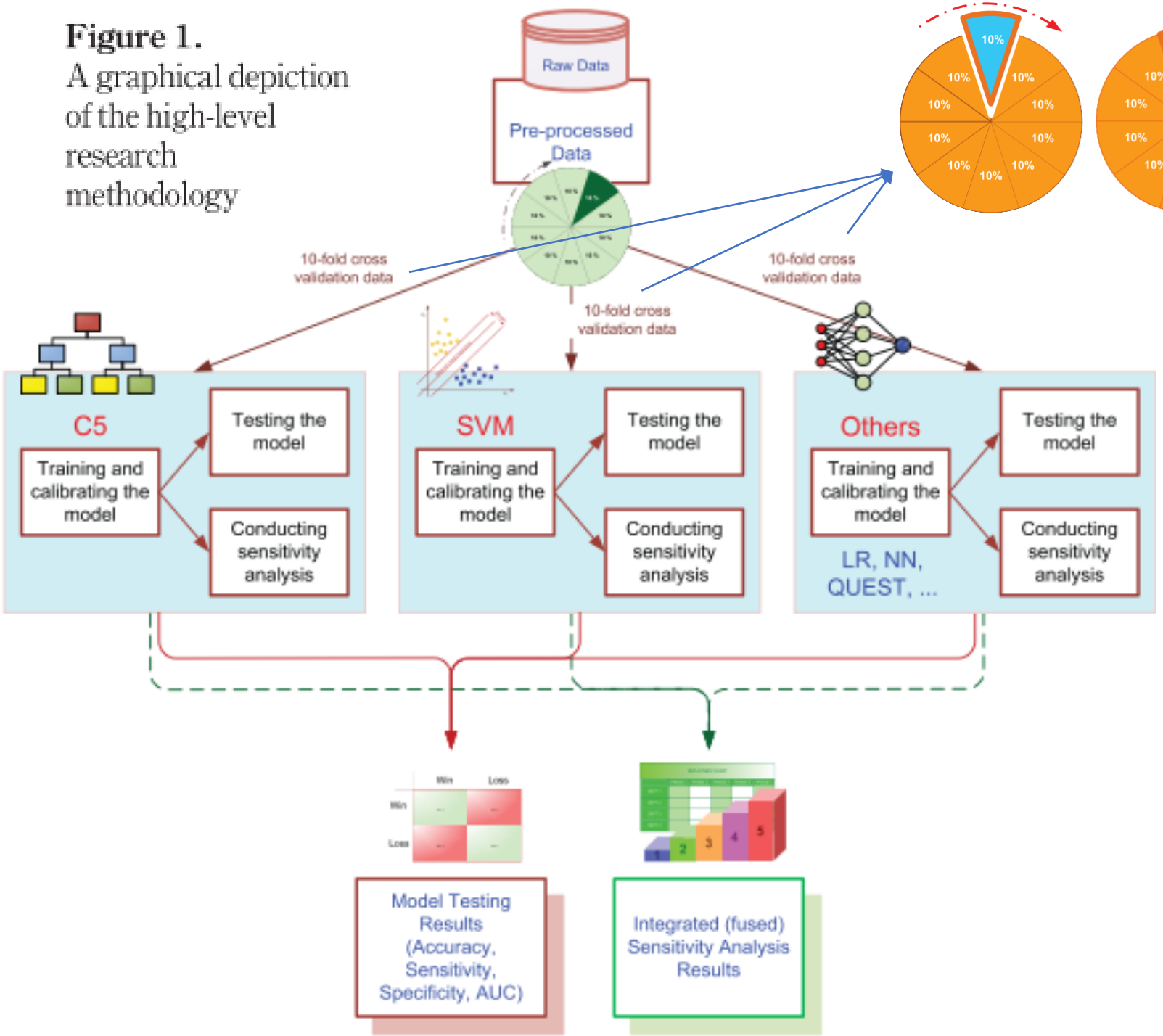
- identificar os fatores que influenciam a funcionalidade de um sistema de custos (CSF)

Metodologia

- Questionário aplicado aos gestões de 565 empresas Turcas escolhidas aleatoriamente
- Algoritmos de decision tree (DT), suporte vector machine(SVM) e regressão logística (LR);

- **Variável dependente:** Cost system functionality (1 = concorda com beneficios do sistema de custo que a firma segue, 0 = discorda). **Variáveis independentes:** Size (nº empregados), financial control, Strategy, Information technology, Complexity of production, Extent of the use of cost data, Extent of the use of budgets, Level of competition e management accounting practices (todas pela scala Likert 5 pontos).

Figure 1.
A graphical depiction of the high-level research methodology



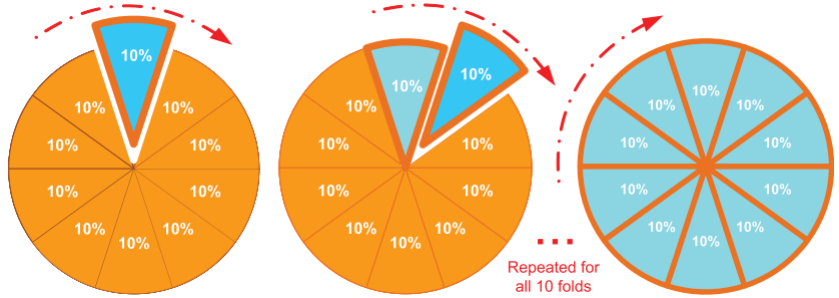


Figure 2.
A graphical representation of 10-fold cross-validation

An investigation of the factors influencing cost system functionality using decision trees, support vector machines and logistic regression

Cemil Kuzey
Arthur J. Bauernfeind College of Business, Murray State University, Murray, Kentucky, USA
Ali Uyar
La Rochelle Business School, Excelia Group, La Rochelle, France, and
Dursun Delen
Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA

Achados

- O **algoritmo C5** teve **melhor acurácia**, para **DT (91,5%),** SVM (79,5%) e LR (63,7%). Os *scores* abaixo mostram a importância dessas variáveis de acordo com o algoritmo DT.

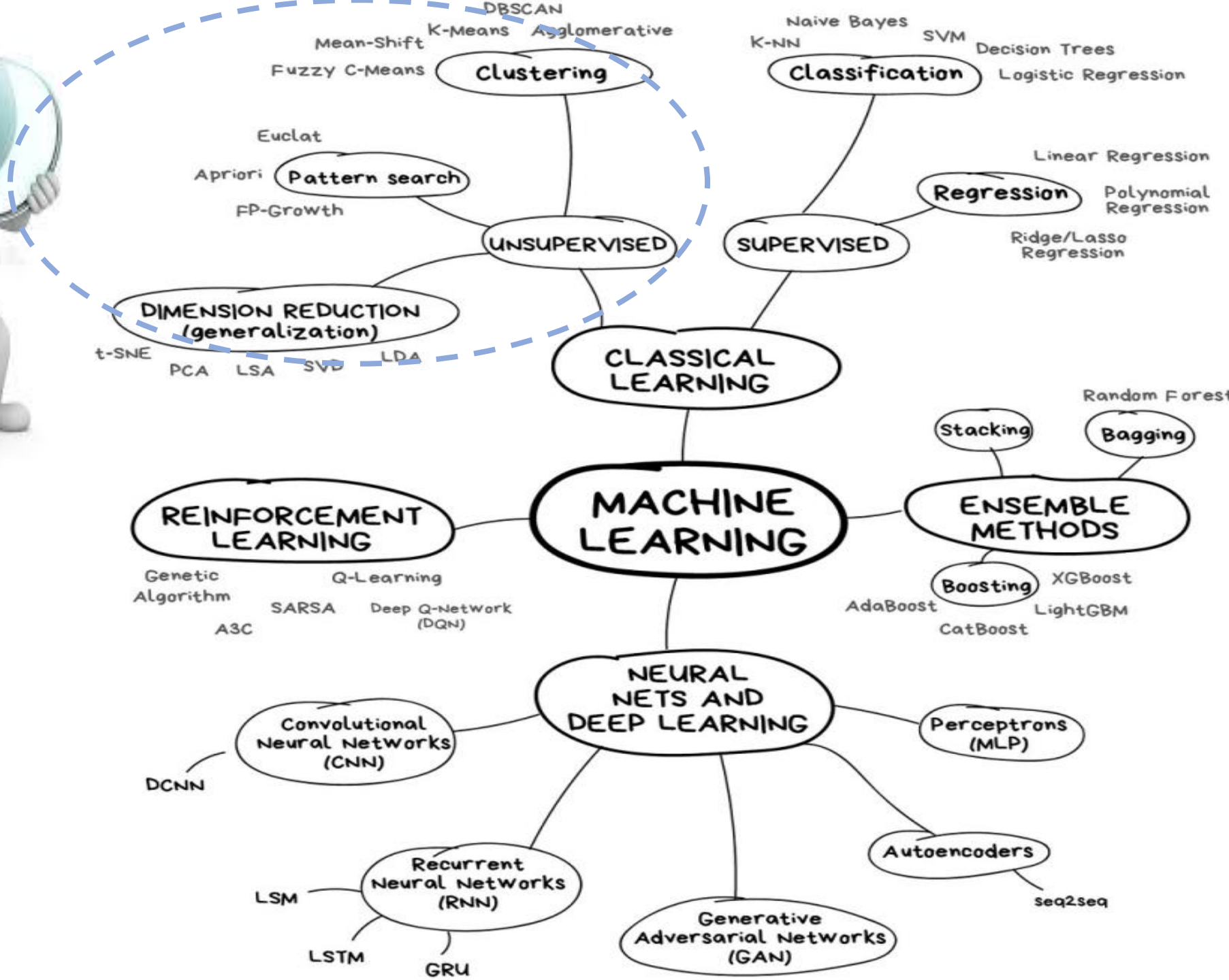
MAP(FC) variável que mais contribui com o sistema de custo

Variables	Fused scores
MAP (Financial control)	2.1292
Information technology (IT)	1.4383
MAP (Planning and control)	1.4296
Extent use of cost data	1.4145
MAP (Effective use of resource)	1.2943
MAP (Reduction of waste)	1.0342
Extend use of budget (2)	0.5940
Size	0.5413
Extend use of budget (1)	0.2396
Level of competition	0.2142
Complexity of production	0.1558
Strategy	0.1499

An investigation of the factors influencing cost system functionality using decision trees, support vector machines and logistic regression

Cemil Kuzey
Arthur J. Bauernfeind College of Business, Murray State University, Murray, Kentucky, USA
Ali Uyar
La Rochelle Business School, Excelia Group, La Rochelle, France, and
Dursun Delen
Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA

Vamos ver quais as divisões dos modelos NÃO supervisionados e algumas pesquisas que utilizam essas técnicas.



Aprendizagem não supervisionada

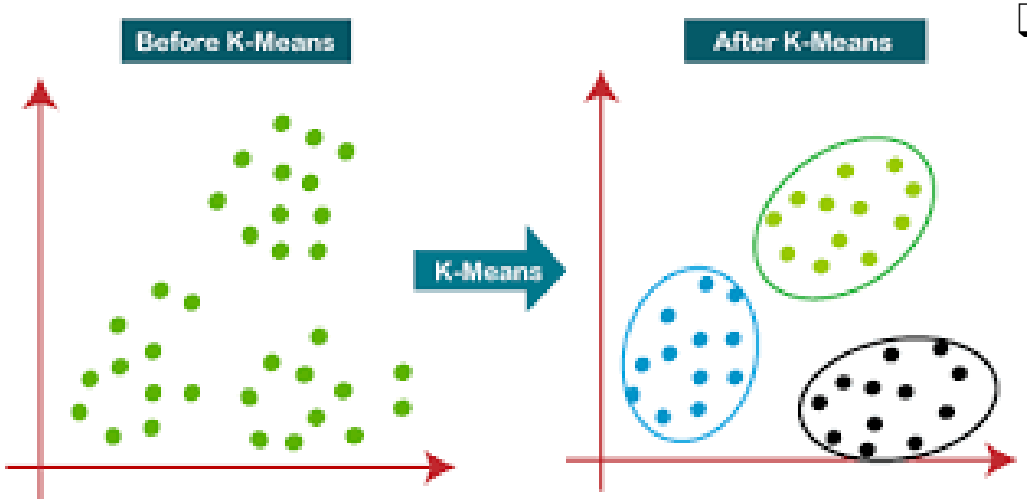
Pode ser dividido em 3 partes:

- **Clustering (agrupamento):** ajuda a agrupar elementos por similaridade. Ex. tipos de roupas, animais, empresas etc;

- **Pattern Search (pesquisa de padrões):** Exemplos: **associação** entre comprar leite e pão juntos. Quem compra carros na **sequência** compra pneus.

- **Redução dimensional(generalização):** é o processo de reduzir o número de features, geralmente aquelas que são correlacionadas. Ex. desempenho no jogo por dia e desempenho por mês. Pode ser generalização de texto também.

Clustering (agrupamento) – algoritimo k-means



- A ideia do agrupamento é juntar dados similares.
- Utilidade da área de contabilidade e finanças:
  - **Análise de portfólio:** útil para agrupar ativos financeiros em diferentes categorias com base em seus perfis de risco e retorno.
  - **Detecção de fraudes financeiras:** útil para identificar grupos de transações financeiras que podem indicar atividades fraudulentas.
  - **Análise de mercado:** útil para identificar grupos de empresas em diferentes setores com base em seus desempenhos financeiros.
  - **Previsão de tendências:** útil para identificar padrões e tendências em séries temporais financeiras, como preços de ações e taxas de juros.
- Além do k-means, existem várias técnicas para fazer agrupamentos, tais como: **Agglomerative** (Sneath, 1957), **DBSCAN (**Ester et al., 1996**), Mean-Shift** (Fukunaga e Hostetler, 1975**), Fuzzy C-Means (**Dunn, 1973 e aprimorado por Bezdek, 1981**) e o mais utilizado deles o K-Means (**MacQueen, 1967**) ;**

Análise de artigo com K-means

The International Journal of Accounting

Earnings attributes and investor-protection: International evidence ☆

engkrai Boonlert-U-Thai [a,*], Gary K. Meek [b], Sandeep Nabar [b]

[a] *Department of Accountancy, Faculty of Commerce and Accountancy, Chulalongkorn University, Phayathai Road, Bangkok 10330, Thailand*

[b] *School of Accounting, Oklahoma State University, Stillwater, OK 74078, USA*

hipótese Maior proteção do investidor (utilizaram 8 métricas) significa maior qualidade dos lucros (4 métricas)

amostra 31 países de 1994 a 2003

método Clusterização com algoritimo **k-means** e regressão linear

achados A **suavização dos lucros** é menor em ambientes de **forte proteção legal**.

Amplos direitos;
Forte Enforcement;
Amplo mercado de capitais;
Baixa concentração de propriedade.

poucos direitos;
Fraco Enforcement;
pequeno mercado de capitais;
alta concentração de propriedade.

Panel B: Cluster membership of countries (*sorted in alphabet order*)

Cluster 1	Cluster 2	Cluster 3
Australia (CM)	Austria (CD)	Brazil (CD)
Canada (CM)	Belgium (CD)	Greece (CD)
Hong Kong (CM)	Chile (CD)	India (CM)
Singapore (CM)	Denmark (CD)	Indonesia (CD)
Sweden (CD)	Finland (CD)	Italy (CD)
UK (CM)	France (CD)	Malaysia (CM)
USA (CM)	Germany (CD)	Mexico (CD)
	Japan (CD)	Philippines (CD)
	Netherlands (CD)	South Africa (CM)
	Norway (CD)	Thailand (CM)
	South Korea (CD)	
	Spain (CD)	
	Switzerland (CD)	
	Taiwan (CD)	

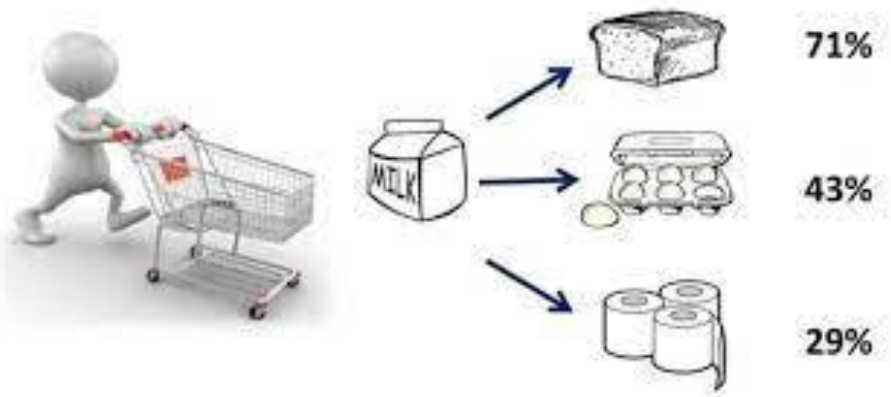
Panel D: Mean earnings-smoothness rank by investor-protection cluster [mean country ranks are presented in parentheses]

	USA (29.00)	Norway (23.14)	Philippines (19.57)
	UK (26.14)	Netherlands (19.86)	Malaysia (18.29)
	Canada (23.86)	Taiwan (17.43)	Indonesia (16.43)
	Hong Kong (23.57)	Finland (17.57)	Mexico (14.29)
	Australia (21.43)	Switzerland (16.29)	Brazil (14.14)
	Sweden (21.14)	Germany (15.86)	Thailand (14.00)
	Singapore (13.00)	Denmark (15.00)	India (13.29)
		France (13.57)	South Africa (12.57)
		Belgium (13.00)	Italy (7.57)
		Austria (12.86)	Greece (6.14)
		Chile (12.43)	
		South Korea (9.43)	
		Japan (9.00)	
		Spain (6.14)	
Mean-rank values	22.59	14.40[@]	13.63[@]
Cluster rank	1st	2nd	3rd

Maior a proteção legal, menor a suavização.

Pattern Search

associação



sequência

- Útil para verificar variáveis que estão associadas a outra ou que seguem uma sequência;
- Um exemplo clássico de **associação** é o caso das compras em supermercados, por exemplo: quem leva leite sempre leva pão. Quem leva cerveja pode levar amendoim também.
- Um exemplo clássico de **sequência** é a questão de quem vem primeiro, o ovo ou a galinha?.
- **Principais algoritmos:** Apriori(Agrawal e Srikant, 1994), Euclat (Borgelt, 1997), FP-growth(Han et al, 2000).

Redução dimensional(generalização)

Exemplo de aplicação da LSA:

"Manipulating facial expressions and body movements in videos has become so advanced that most people struggle to tell the difference between fake and real. A fake video of Barack Obama went viral last year where you see the former President addressing the camera. If you turn off the sound, you will not even realize it's a fake video!"

	Topic 1
	Topic 2
	Topic 3

Pelo tópico 2 é fácil notar que o texto fala de **vídeos fake**. A máquina poderia generalizar dessa forma.

- Útil para reduzir várias características similares entre um conjunto de *features*. Ex. cachorros de orelhas triangulares, focinhos longos e rabos grandes e pelo médio, resumindo tudo isso a abstração "pastor alemão". Ganhamos assim rapidez e menos possibilidade de erros.

- **Algoritmos populares:**
  - Principal Component Analysis (PCA) (ideia inicial Pearson, 1901);
  - Singular Valu Decomposition (SVD) (ideia inicial Beltrami, 1873);
  - Latent Dirichlet allocation (LDA)(Blei et al., 2003);
  - Latent Semantic Analysis (LSA, pLSA, GLSA) (Deerwester, 1990);
  - t-SNE (Hinton, Maaten, 2008).

- Na área de finanças e contabilidade as aplicações mais comuns são a PCA e LSA. No slide seguinte veremos as aplicações da LSA.

Redução dimensional(generalização) - LSA

- São exemplos de aplicações da Análise semântica latente (LSA) na área de contabilidade e finanças:
  - **Análise de relatórios financeiros:** identifica palavras ou expressões que indicam sentimentos, riscos financeiros, oportunidades de investimento ou tendências de mercado.
  - **Análise de risco de crédito:** identifica palavras ou expressões que indiquem possíveis problemas financeiros.
  - **Análise de notícias financeiras:** identifica eventos que possam impactar o mercado financeiro, como crises econômicas, mudanças na política fiscal ou mudanças regulatórias.
  - **Análise de redes sociais:** analisa as conversas nas redes sociais e identifica tendências financeiras e de mercado, bem como sentimentos e opiniões expressas pelos usuários.

Key audit risks and audit procedures during the initial year of the COVID-19 pandemic: an analysis of audit reports 2019-2020

Michael Kend and Lan Anh Nguyen
School of Accounting, Information Systems and Supply Chain, RMIT University, Melbourne, Australia

Objetivo: verificar os **principais riscos de auditoria**, durante o ano anterior(2019) e o ano inicial(2020) do surto de COVID-19;

Amostra: 3073 relatórios de auditoria de empresas australianas

Metodologia: análise semântica latente

Conclusões:

- ❑ apenas 3% dos relatórios de auditoria de 2020 relataram os riscos de auditoria associados à pandemia de COVID-19;
- ❑ Pequenas empresas de auditoria relatam menos os riscos, ao contrário das grandes;
- ❑ a análise textual encontrou ainda diferenças no tom das palavras usadas por diferentes empresas de auditoria em 2020, mas não foram encontradas diferenças no tom de 2020 em relação a 2019.

Table 4.
Audit procedures referencing COVID-19 within the Top 7 KAM topics (2020) per auditor

KAM by topic 2020 (Top 7)
Goodwill and intangibles
PPE
Revenue recognition
Acquisitions
Impairments
Asset valuation
Exploration and evaluation
(COVID-19 references within the Top 7 KAM topics)
Total COVID-19 references for each auditor/s

Note: *Auditor switches not inclu

Table 5.
Sentiment analysis and readability tests

			Sentiment analysis*			
Year	FOG index	Flesch Kincaid grade level	Positive words	(%)	Negative words	(%)
2019			118	8.8	1,209	91.2
2020			382	13.2	2,515	86.8
Auditor in 2020						
Deloitte	27.32	48.41	38	12.4	269	87.6
EY	27.68	30.06	57	16.8	283	83.2
KPMG	26.88	35.84	88	18.6	384	81.4
PwC	29.77	32.15	49	11.5	376	88.5
NB4	28.44	45.73	120	10.8	982	89.2

Vamos estudar um pouco sobre aprendizagem por esforço.

MACHINE LEARNING

CLASSICAL LEARNING

UNSUPERVISED

Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means

Pattern search
Euclat
Apriori
FP-Growth

DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA

SUPERVISED

Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression

Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression

REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost

NEURAL NETS AND DEEP LEARNING

Convolutional Neural Networks (CNN)
DCNN

Perceptrons (MLP)

Recurrent Neural Networks (RNN)
LSM
LSTM
GRU

Generative Adversarial Networks (GAN)

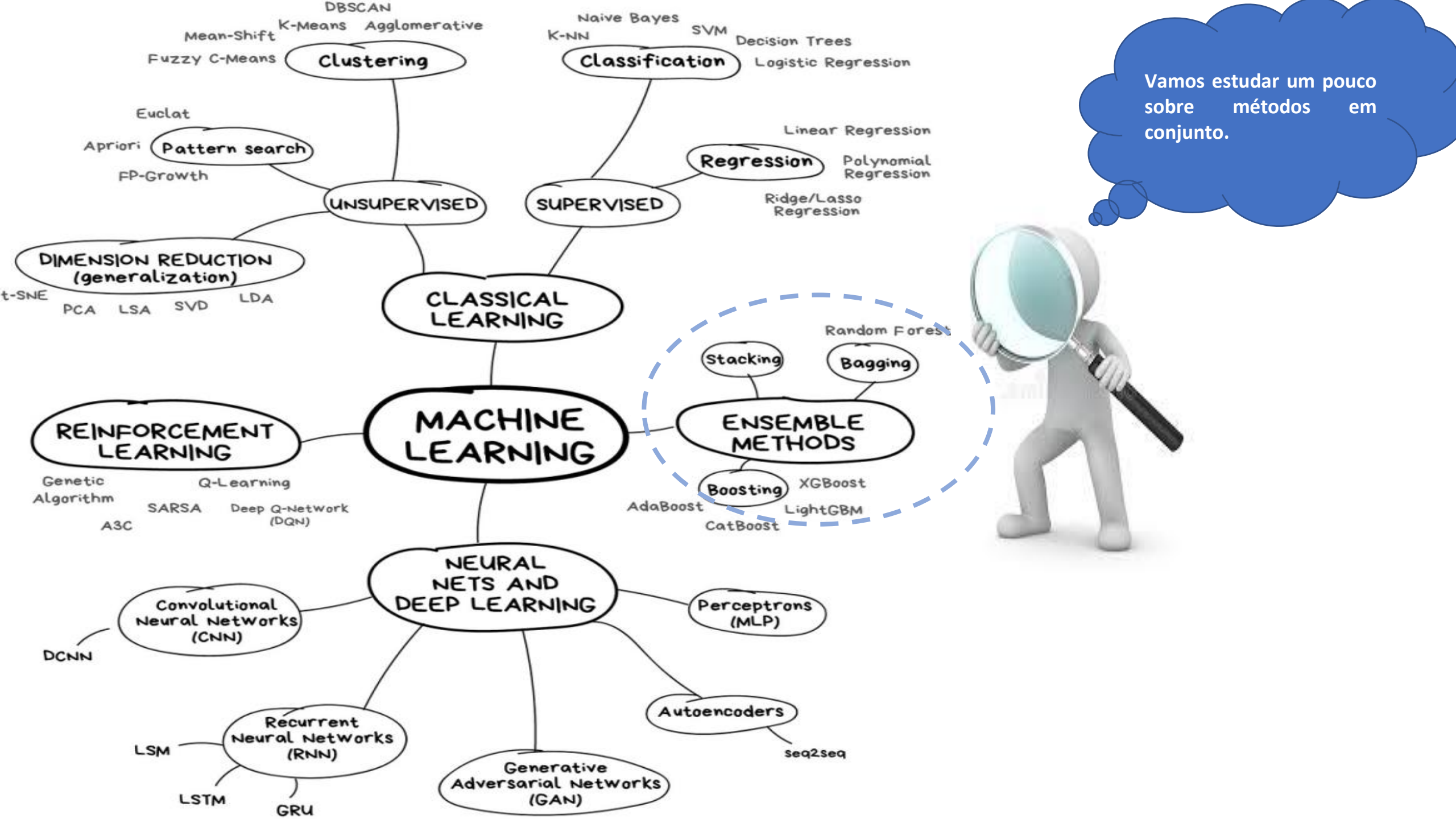
Autoencoders
seq2seq

Aprendizagem por esforço

- É um tipo de aprendizagem por **tentativa e erro**. A aprendizagem **inicia sem nenhum tipo de instrução** de como fazer. É como você tivesse que aprender um assunto do zero sozinho.
- Útil para deixar o robô aprender com os erros e não ter que prever todas as situações. Por exemplo: na programação de carros autônomos, a máquina aprende no ambiente virtual, sem instrução prévia, mata muito e aprende com estes erros. **A máquina não precisa receber nenhum tipo de instrução de como fazer**.
- **Principais algoritmos:** Q-Learning, SARSA, DQN, A3C, Genetic algorithm (Holland, 1960).
- O algoritmo genético (AG) é muito útil para problemas de otimização na área de finanças.
- O trabalho de **Kalayci et al (2017)** faz uma revisão sobre a aplicação do AG em problemas de **otimização de carteiras**. PAJES_23_4_470_476.pdf (journalagent.com)
- De acordo com **oliveira e Maciel(2021),** as **carteiras** otimizadas por meio de AG é superior em relação a outros métodos. (7c82fab8c8f89124e2ce92984e04fb40.pdf (anpad.com.br))

Aprendizagem por esforço – algoritmo genético (AG)

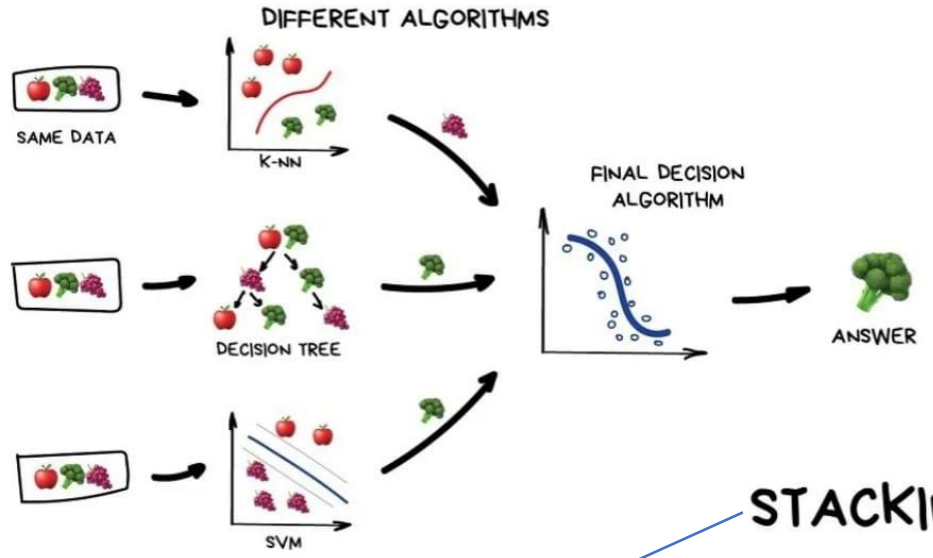
- AG é uma técnica que simula o processo de **evolução natural** para encontrar **soluções otimizadas**;
- Inspirada na **teoria da evolução de Charles Darwin**, a seleção natural dos indivíduos mais aptos;
- Esse algoritmo segue os seguintes passos:
  - uma **população** de **soluções candidatas** é **gerada aleatoriamente** para resolver um problema;
  - Cada **solução é cromossomo**, uma sequência de **genes** que codificam características da solução;
  - Os **genes** são avaliados com uma função de **fitness** que mede o quão boa é cada solução;
  - As **melhores soluções** são selecionadas para se **reproduzir** por meio de **crossover** e **mutação**;
  - Esse processo é repetido por várias gerações, a **população evolui** para soluções otimizadas.

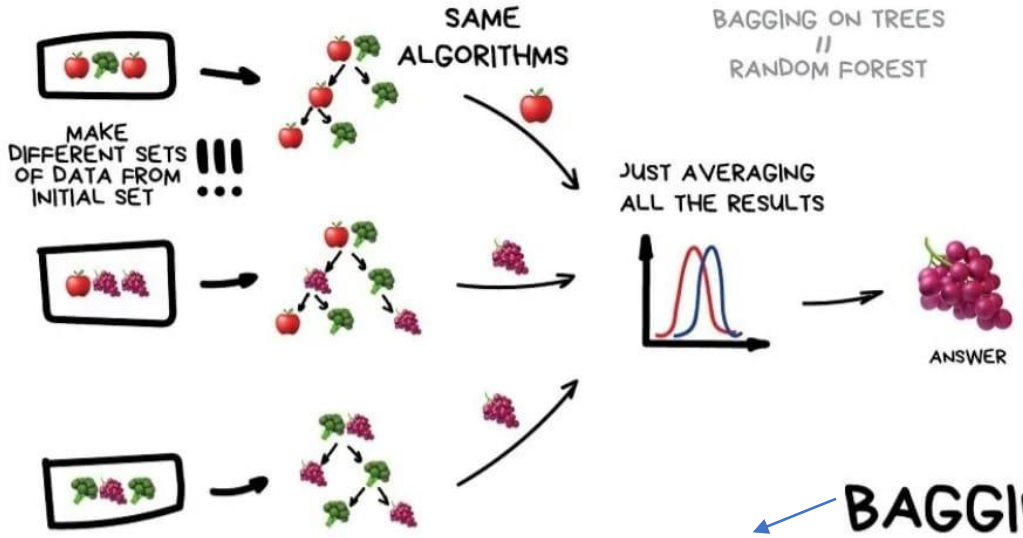

MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Random Forest
Bagging
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq
Vamos estudar um pouco sobre métodos em conjunto.

Ensemble Methods (modelos em conjunto)

- Métodos Ensemble e redes neurais (próximo tópico) são os mais próximos da singularidade;
- A diferença entre os dois é que este por ser mais simples exige menos capacidade computacional e é mais rápido que aquele. Podendo ser trabalhado em tempo real;
- Na essência o método **aprende com os erros de outros algoritmos ou com algoritmos semelhantes**;
- **Principais algoritimos:** random forest (Breiman e Cutler, 2001) (várias decision tree, uma corrigindo o erro da outra), RUSboosting (Seiffert et al., 2010) entre outros.
- Existem três métodos para criar ensemble: Stacking ou “empilhamento”, bagging e boosting, veja o slide seguinte:

Ensemble Methods (modelos em conjunto)



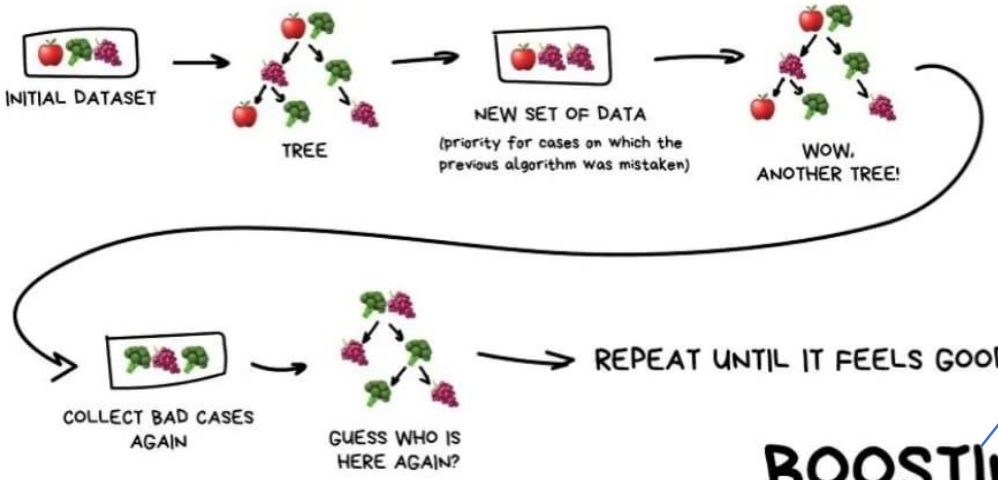


STACKING

Combina várias previsões de modelos diferentes em um meta-modelo

BAGGING

usa diferentes subconjuntos de dados para criar uma previsão geral



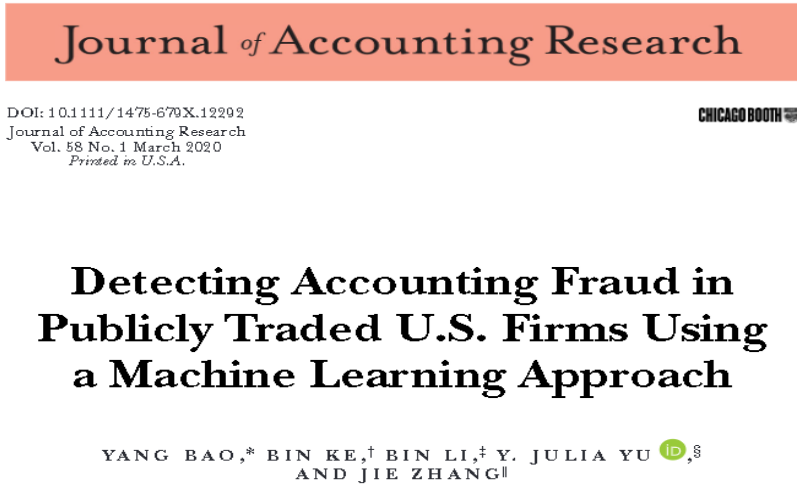
BOOSTING

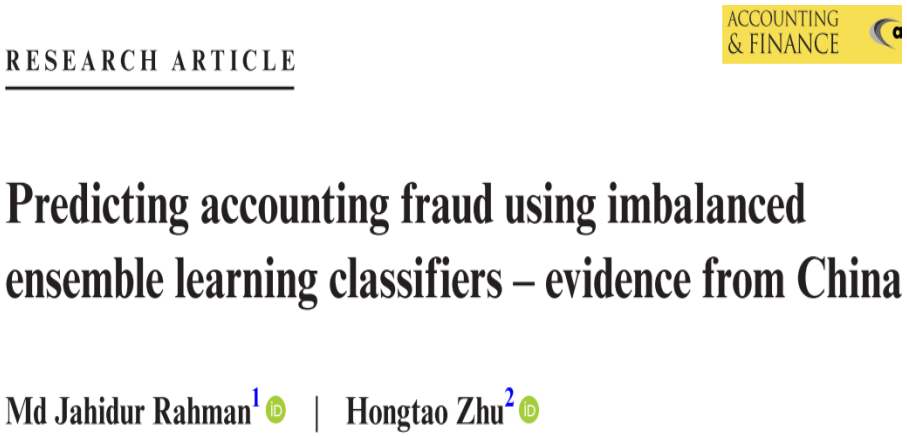
Cada modelo é treinado em cima dos erros do modelo anterior, com o objetivo de corrigir as previsões anteriores e melhorar a precisão geral do

Amostra e variáveis: de 1991 a 2008, 28 variáveis contábeis com base na literatura.

Principais Conclusões:

- **Ensemble (RUSboost)** melhor que métodos anteriores (Dechow et al. (2011) **regressão logística**, e Cecchini et al. (2010) **SVM**- Suport Vector Machine).
- Utilizar variáveis teoricamente testadas é melhor do que empregar todas as contas das demonstrações contábeis.



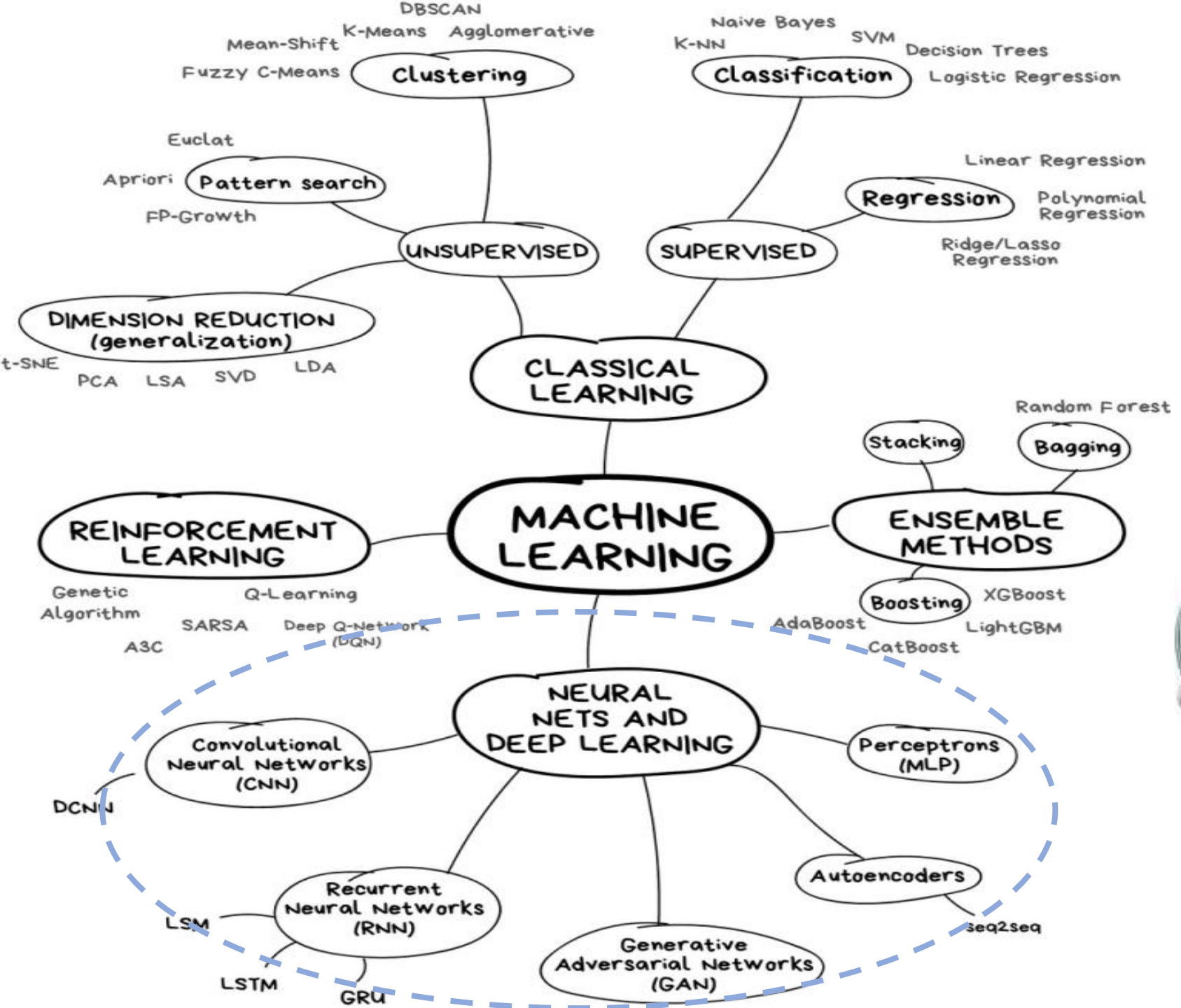


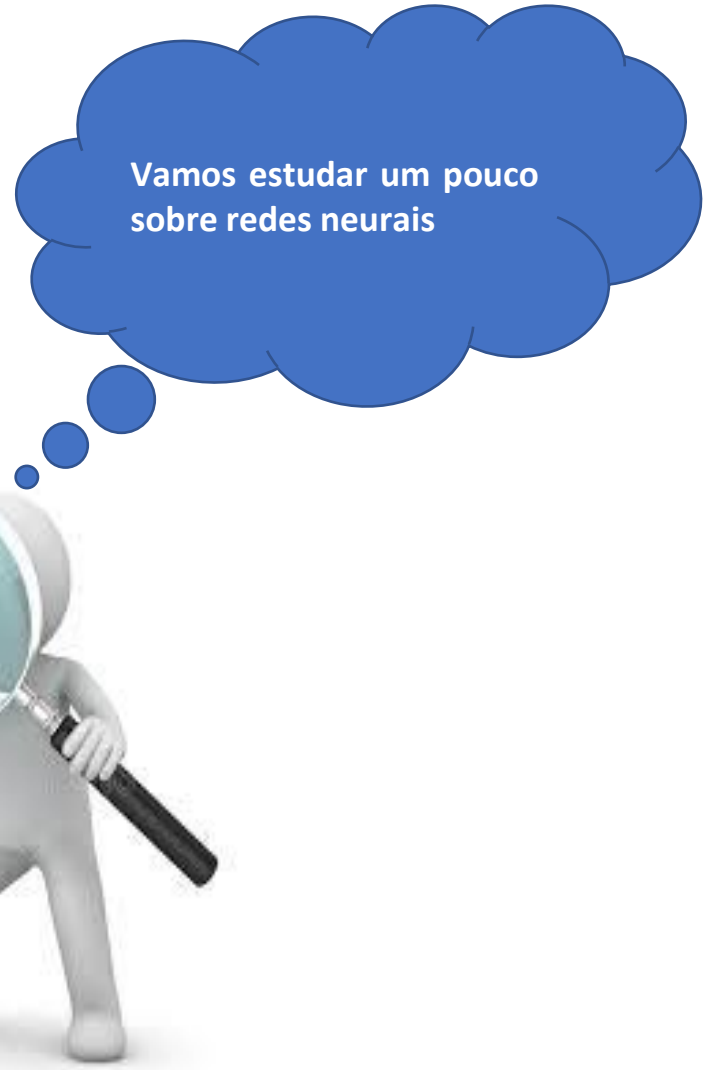
Amostra e variáveis: de 1998 a 2017, 12 índices e 28 dados financeiros.

Principais conclusões:

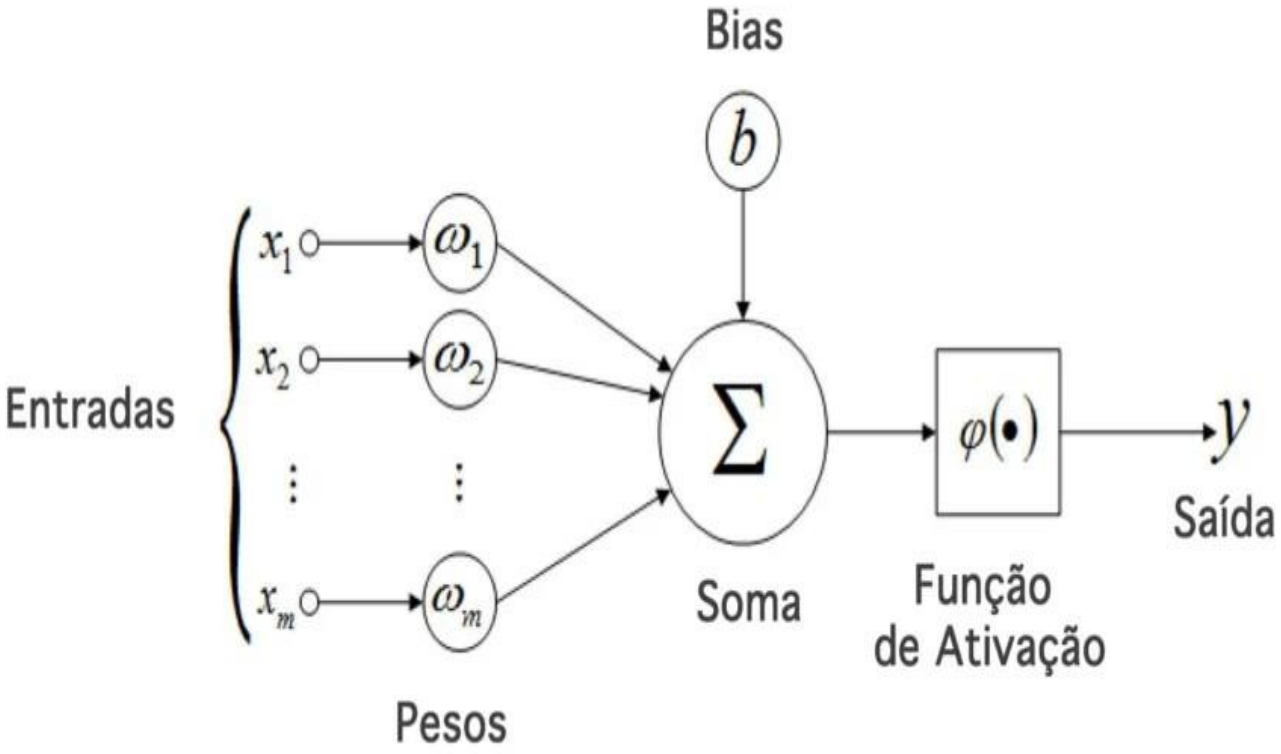
Ensemble é melhor para detectar fraude que a **regressão logística**;

CUSBoost e RUSBoost performou melhor que **AdaBoost e XGBoost**.


DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Clustering
Naive Bayes
K-NN
SVM
Decision Trees
Classification
Logistic Regression
Euclat
Apriori
Pattern search
FP-Growth
Linear Regression
Regression
Polynomial Regression
Ridge/Lasso Regression
UNSUPERVISED
SUPERVISED
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
CLASSICAL LEARNING
Random Forest
Stacking
Bagging
REINFORCEMENT LEARNING
MACHINE LEARNING
ENSEMBLE METHODS
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Autoencoders
seq2seq
Generative Adversarial Networks (GAN)


Vamos estudar um pouco sobre redes neurais

Redes neurais



Exemplo de rede neural *perceptron*

- Pode substituir todos os algoritmos de aprendizagem supervisionada e não–supervisionada já citados;

- Tenta reproduzir o cérebro humano. Pega os dados de entrada, esses dados são processados por uma rede de neurônios e conexões, são atribuídos pesos e depois é gerado uma saída;

- A literatura aponta que as RNs são superiores a modelos econométricos tradicionais, no que se refere a acurácia de previsão e classificação. Principalmente em relações não lineares.

- **Arquiteturas Populares:**
  - Perceptron (Rosenblatt, 1957),
  - Convolutional Network (CNN) (Fukushima, 1980 e LeCun, 1998) ,
  - Recurrent Networks (RNN) (vários autores na década de 80),
  - Autoencoders (ideia inicial Werbos em 1974 e desenvolvido por Hinton et al, 1991).

Níveis de governança corporativa da B3: interesse e desempenho das empresas – uma análise por meio de redes neurais artificiais

B3´s levels of corporate governance: company interest and performance – an artificial neural network analysis

Vitor Borges Tavares[1] e Antônio Sérgio Torres Penedo[2]

Níveis de governança corporativa da B3: interesse e desempenho das empresas – uma análise por meio de redes neurais artificiais | Contabilidade Gestão e Governança (revistacgg.org)

Níveis de governança corporativa da B3: interesse e desempenho das empresas – uma análise por meio de redes neurais artificiais

B3's levels of corporate governance: company interest and performance – an artificial neural network analysis

Vitor Borges Tavares[1] e Antônio Sérgio Torres Penedo[2]

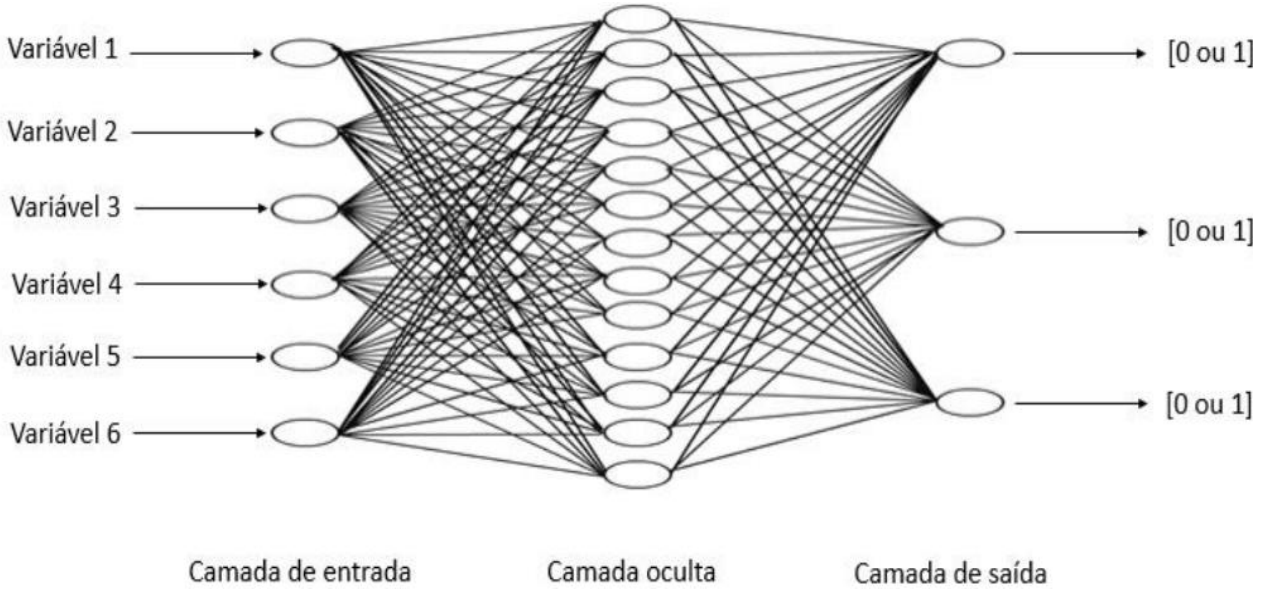
Objetivo: Classificar as empresas não listadas nos níveis de governança de acordo com variáveis das empresas optantes aos níveis de governança corporativa.

Amostra: empresas listadas na B3 no final de 2015. Treino com 179 empresas.

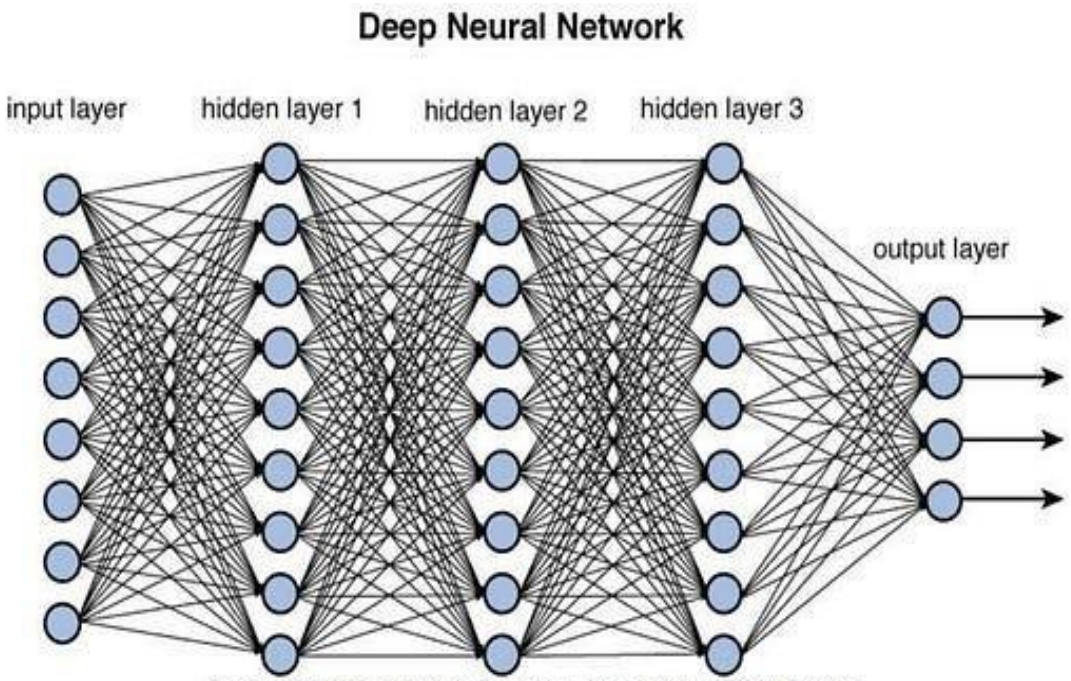
Metodologia: 6 variáveis binárias de **entrada** de governança: **1**-emite apenas ações ON? **2**-Ações preferenciais, menos de 50%? **3**-Controladores possuem menos de 70% das ON? **4 –** Conselho de administração, mínimo 3 membros? **5**-Mínimo 5 membros? **6**-Mandato de 2 anos? E três variáveis de **saída** (N1, N2 e NM)

Resultados e conclusões:
A rede classificou 200 empresas não pertencentes aos níveis de governança em N1 (84), N2 (23), NM (17) e nenhum.

Os autores compararam a acurácia da rede com um modelo de regressão logística multinomial e verificaram a maior acurácia, 92,2% contra 91,1%.



Deep learning



- É uma rede neural com um maior número de camadas ocultas (profundidade das camadas);
- Não existe um número ideal de camadas ocultas, para verificar isso pode ser utilizada técnicas de validação cruzada para avaliar qual a melhor rede (mais acurácia), sempre buscando aquela mais simples (menos neurônios e camadas).
- quanto ao número de neurônios dessas camadas, segundo Hecth-Nielson (1990) podem ser definidos como 2 x variáveis de entrada +1, mas isso não é um consenso.

Artigos que comparam vários métodos

Journal of Management Information Systems

ISSN: 0742-1222 (Print) 1557-928X (Online) Journal homepage: http://www.tandfonline.com/loi/mmis20

Leveraging Financial Social Media Data for Corporate Fraud Detection

Wei Dong, Shaoyi Liao & Zhongju Zhang

of classifiers and avoids potential overfitting problems. We use support vector machine (SVM) for document classification. SVM has been shown to be successful in working with large feature space and small sample set [16], and is capable of handling large sparse data [33]. For comparison purposes, we also implement logistic regression (LR), neural networks (NN), and decision tree (DT) in this study. We use accuracy, recall, F1 score, and the area under the receiver operating

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

journal homepage: www.elsevier.com/locate/accinf

Detecting accounting fraud in companies reporting under US GAAP through data mining

Mário Papík [a,*], Lenka Papíková [b]

measures for eight out of ten data mining techniques. The eight data mining techniques used in this study were discriminant analysis, decision tree, k-nearest neighbour, logistic regression, neural network, random forest, repeated incremental pruning to produce error reduction, and support vector machine.

Auditing: A Journal of Practice & Theory
Vol. 30, No. 2
May 2011
pp. 19–50

American Accounting Association
DOI: 10.2308/ajpt-50009

Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms

Johan Perols

SUMMARY: This study compares the performance of six popular statistical and machine learning models in detecting financial statement fraud under different assumptions of misclassification costs and ratios of fraud firms to nonfraud firms. The results show, somewhat surprisingly, that logistic regression and support vector machines perform well relative to an artificial neural network, bagging, C4.5, and stacking. The results also

Principais observações em ML

- Amostras pequenas e não diversificadas. Dados ruins, até mesmo o melhor algoritmo não ajudará;
- Há sempre vários algoritmos adequados ao problema e você precisa escolher qual deles se encaixa melhor.
- Tudo pode ser resolvido com uma rede neural, é claro, mas quem pagará por toda essa força computacional? (**no caso de big data**)
- A regra geral é quanto mais complexos os dados, mais complexo é o algoritmo.
- Para textos, números e tabelas, é melhor a abordagem clássica.
- Para fotos, vídeos e todas as outras coisas complicadas de Big Data, é melhor redes neurais.

Para reflexão...

Contato:

Email:

jeffersonramelo@hotmail.com