A Literature Review of Using Machine Learning in Software Development Life Cycle Stages

Carolina Dias
Claudio Fortier

Introdução

- **Eterno *trade-off***: as práticas recomendadas de engenharia de software na indústria geralmente são deixadas de lado devido ao tempo para a solução ser entregue.
- Nisso, os sistemas de software podem se **beneficiar de ferramentas de IA** para automatizar alguns processos, como no design de requisitos.
- Aplicações de ML para ES variam desde a resolução de requisitos ambíguos até a predição de defeitos em softwares.

Escopo e Trabalhos Relacionados

- **Revisão da literatura** relacionada a aplicação de técnicas de ML ao Ciclo de Vida do Desenvolvimento de Sistemas (SDLC).
- **Visão ampla** do estado-da-arte.
- **Sugestão de áreas de interesse** de pesquisa onde estudos mais aprofundados são necessários.
- Diferentemente de outras revisões da literatura, essa abrange **todos os passos do ciclo de vida**, ao invés de só uma etapa, como a de testes, por exemplo, já realizada por outros artigos.

Metodologia

1. Identificar a suscetibilidade de vários tipos de técnicas de ML nas etapas do ciclo de vida de desenvolvimento de software
   1.1. Quais etapas do SDLC os pesquisadores da indústria e da academia mais focam?
   1.2. Quais as aplicações do ML para a ES?
   1.3. Quais tipos e técnicas de ML são utilizadas em ES?
2. Entender a maturidade da pesquisa nessa área
   2.1. Qual a contribuição real dos artigos publicados?
   2.2. Quais as evidências empíricas dos artigos?
   2.3. Quais conjuntos de dados são comumente utilizados?
3. Identificar a diversidade demográfica dessa área
   3.1. Quais são as tendências por ano das publicações da área?
   3.2. Quais são as conferências com maior quantidade de publicações da área?
4. Entender as implicações, desafios, limitações e direções futuras de pesquisa na área

Metodologia

1. Identificar a suscetibilidade de vários tipos de técnicas de ML nas etapas do ciclo de vida de desenvolvimento de software
   1.1. Quais etapas do SDLC os pesquisadores da indústria e da academia mais focam?
   1.2. Quais as aplicações do ML para a ES?
   1.3. Quais tipos e técnicas de ML são utilizadas em ES?
2. Entender a maturidade da pesquisa nessa área
   2.1. Qual a contribuição real dos artigos publicados?
   2.2. Quais as evidências empíricas dos artigos?
   2.3. Quais conjuntos de dados são comumente utilizados?
3. Identificar a diversidade demográfica dessa área
   3.1. Quais são as tendências por ano das publicações da área?
   3.2. Quais são as conferências com maior quantidade de publicações da área?
4. Entender as implicações, desafios, limitações e direções futuras de pesquisa na área

Metodologia

1. Identificar a suscetibilidade de vários tipos de técnicas de ML nas etapas do ciclo de vida de desenvolvimento de software
   1.1. Quais etapas do SDLC os pesquisadores da indústria e da academia mais focam?
   1.2. Quais as aplicações do ML para a ES?
   1.3. Quais tipos e técnicas de ML são utilizadas em ES?
2. Entender a maturidade da pesquisa nessa área
   2.1. Qual a contribuição real dos artigos publicados?
   2.2. Quais as evidências empíricas dos artigos?
   2.3. Quais conjuntos de dados são comumente utilizados?
3. Identificar a diversidade demográfica dessa área
   3.1. Quais são as tendências por ano das publicações da área?
   3.2. Quais são as conferências com maior quantidade de publicações da área?
4. Entender as implicações, desafios, limitações e direções futuras de pesquisa na área

Metodologia

1. Identificar a suscetibilidade de vários tipos de técnicas de ML nas etapas do ciclo de vida de desenvolvimento de software
    1.1. Quais etapas do SDLC os pesquisadores da indústria e da academia mais focam?
    1.2. Quais as aplicações do ML para a ES?
    1.3. Quais tipos e técnicas de ML são utilizadas em ES?
2. Entender a maturidade da pesquisa nessa área
    2.1. Qual a contribuição real dos artigos publicados?
    2.2. Quais as evidências empíricas dos artigos?
    2.3. Quais conjuntos de dados são comumente utilizados?
3. Identificar a diversidade demográfica dessa área
    3.1. Quais são as tendências por ano das publicações da área?
    3.2. Quais são as conferências com maior quantidade de publicações da área?
4. Entender as implicações, desafios, limitações e direções futuras de pesquisa na área

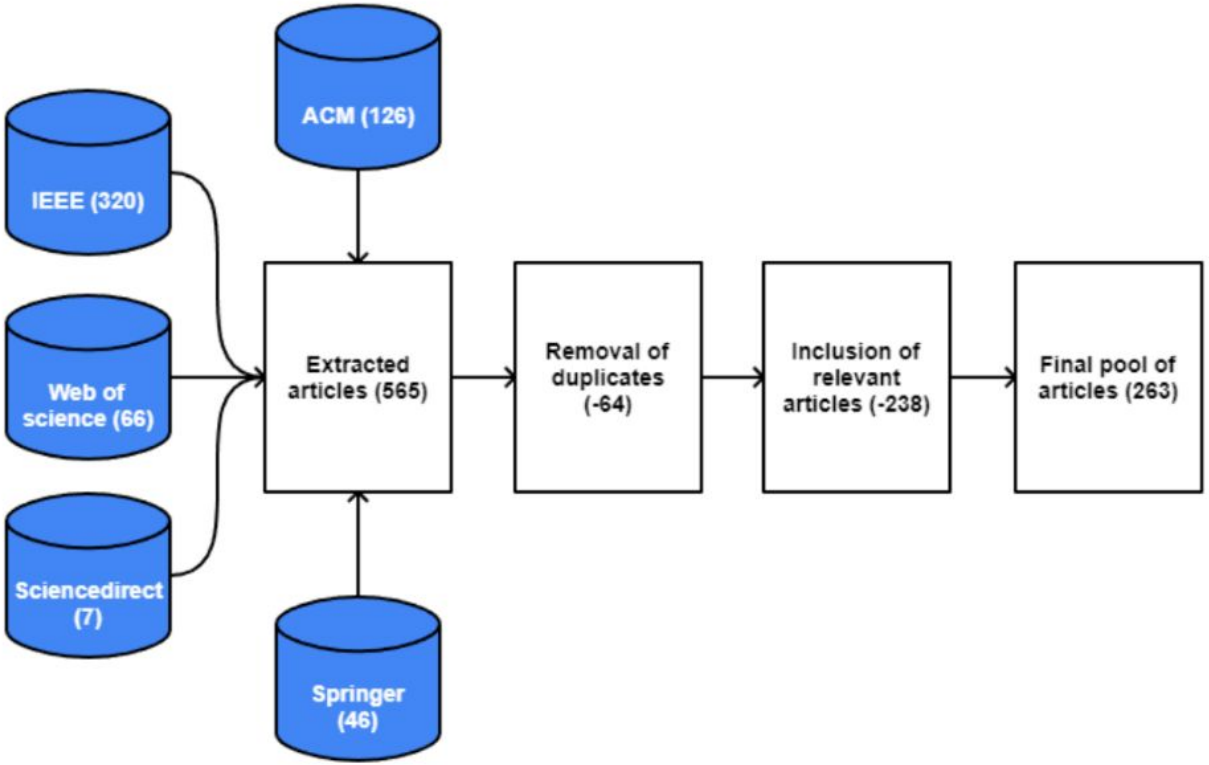
Busca pelos Artigos

String de Busca:

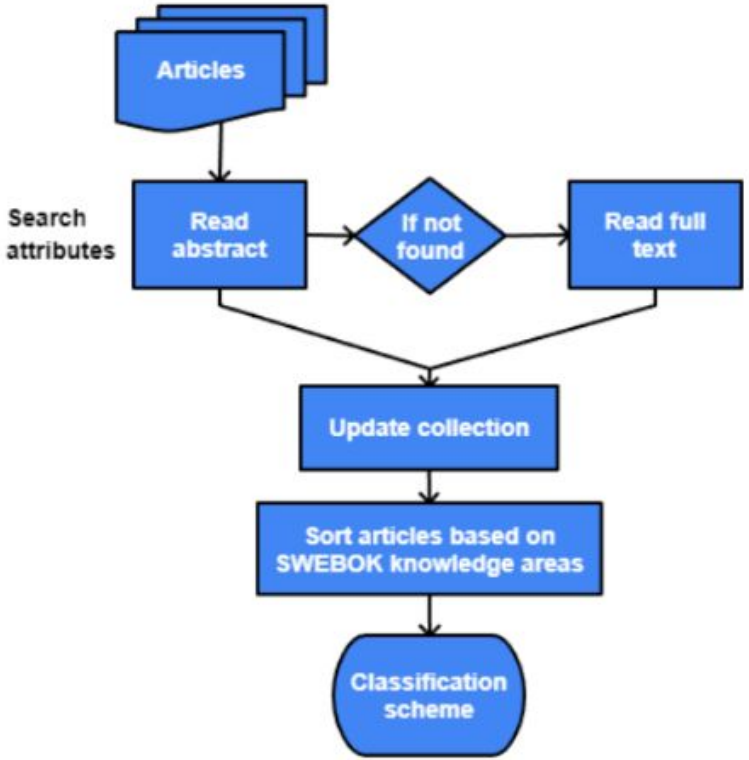
```
(‘‘machine learning’’ OR ‘‘deep learning’’) AND software AND requirement* OR
specification* OR design OR model OR analysis OR architecture OR
implementation OR code OR test* OR verification OR validation OR maintenance
```

- ACM Digital Library
- IEEEXplore
- ScienceDirect
- Springer
- Web of Science

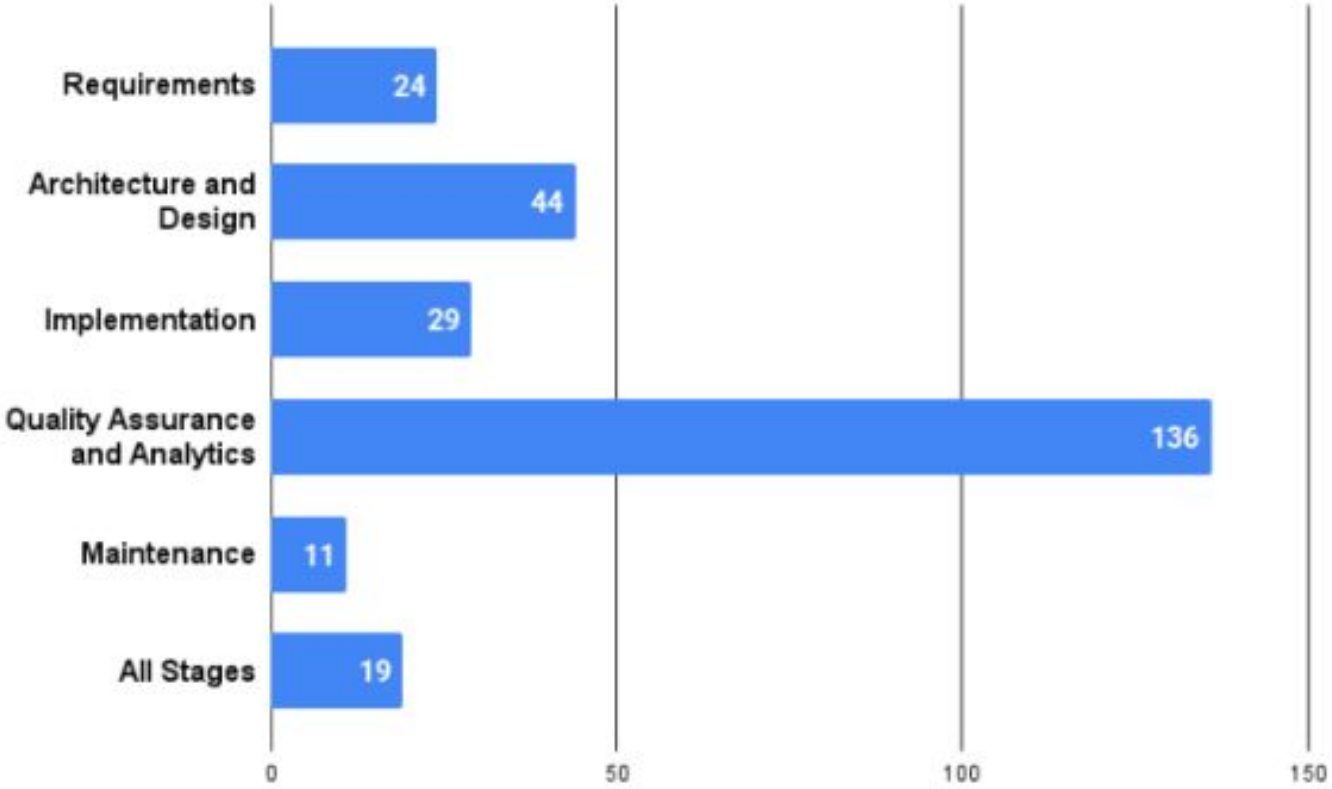
Escolha dos Artigos



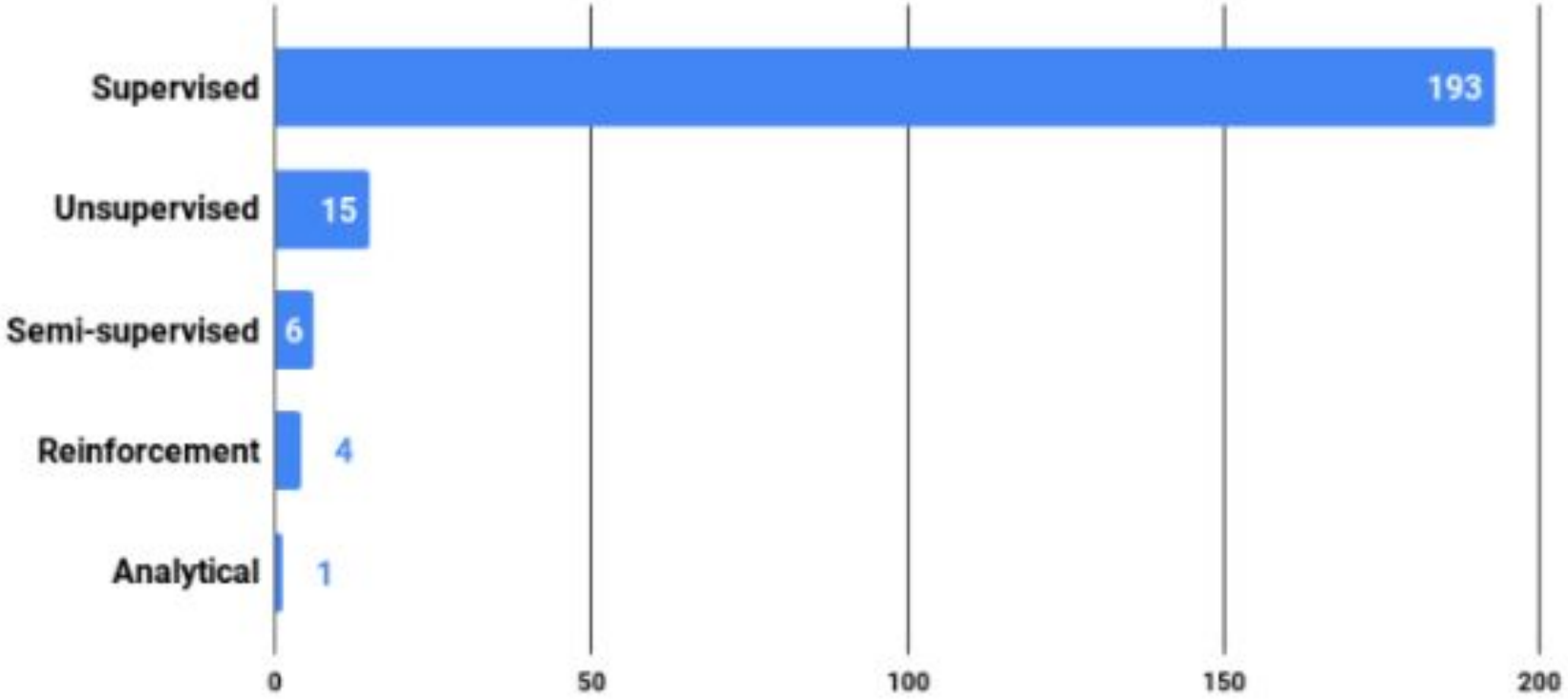
Classificação dos Artigos



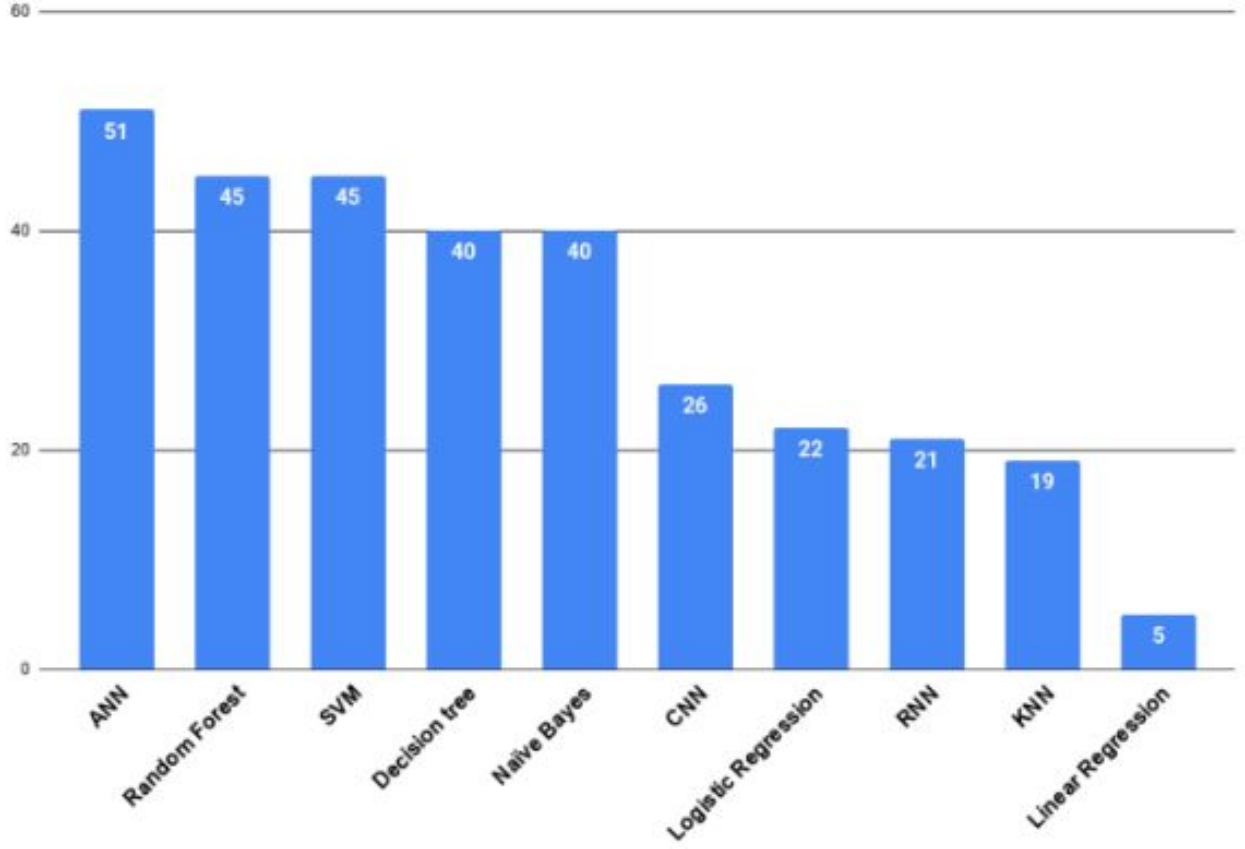
Etapa do SDLC com mais Artigos com ML



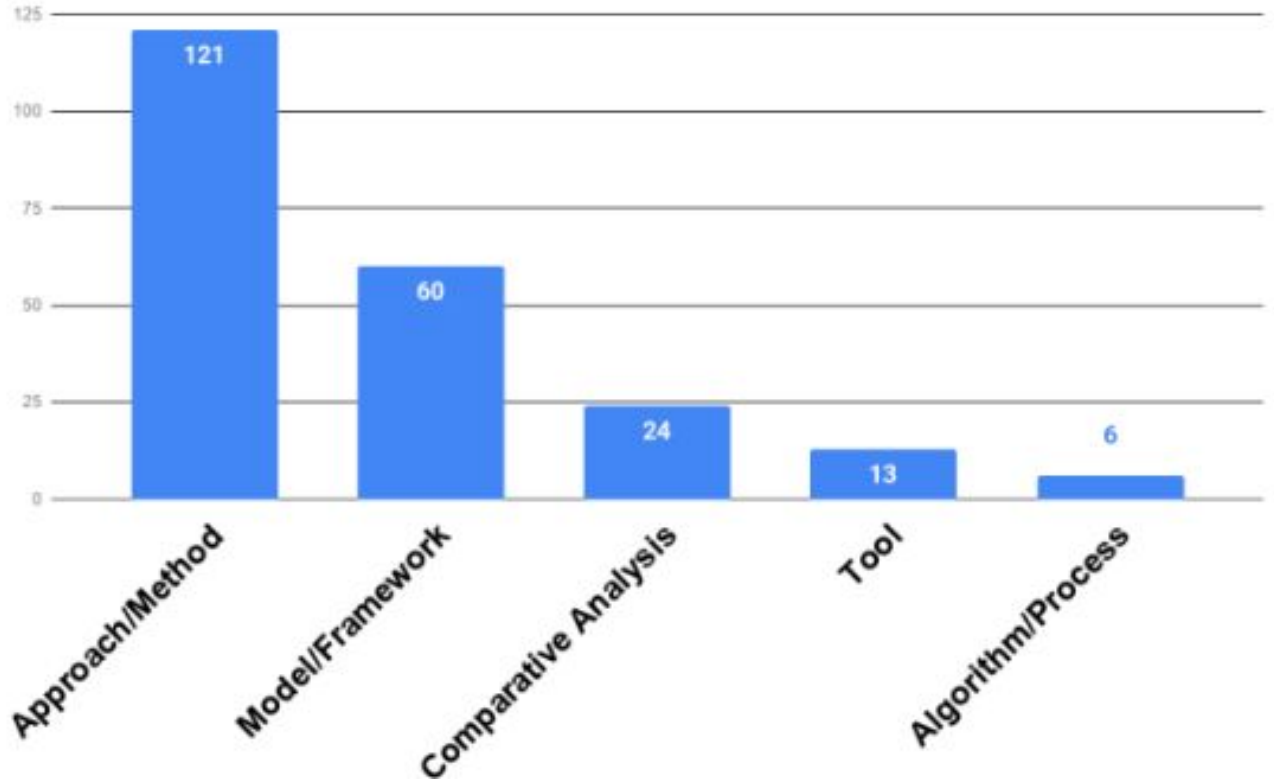
Tipo de ML Utilizado nos Artigos



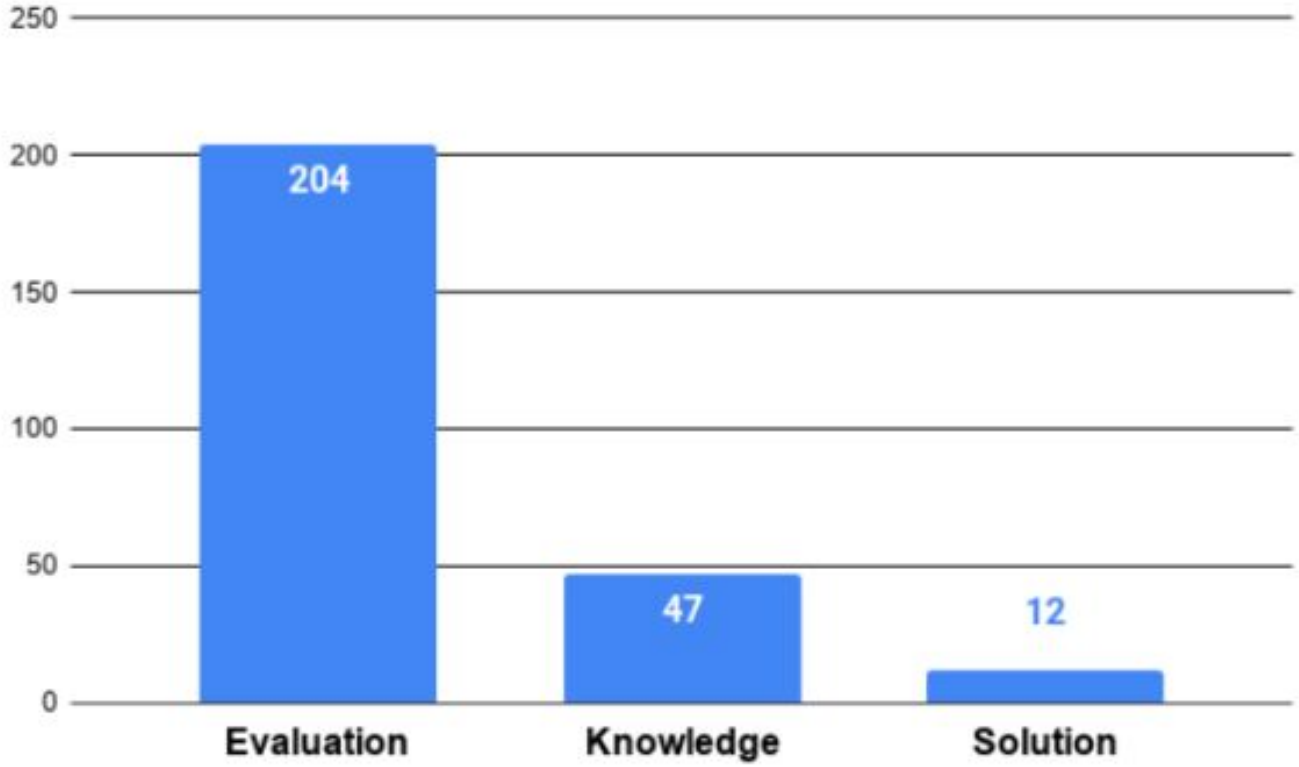
Técnica de ML Utilizada nos Artigos



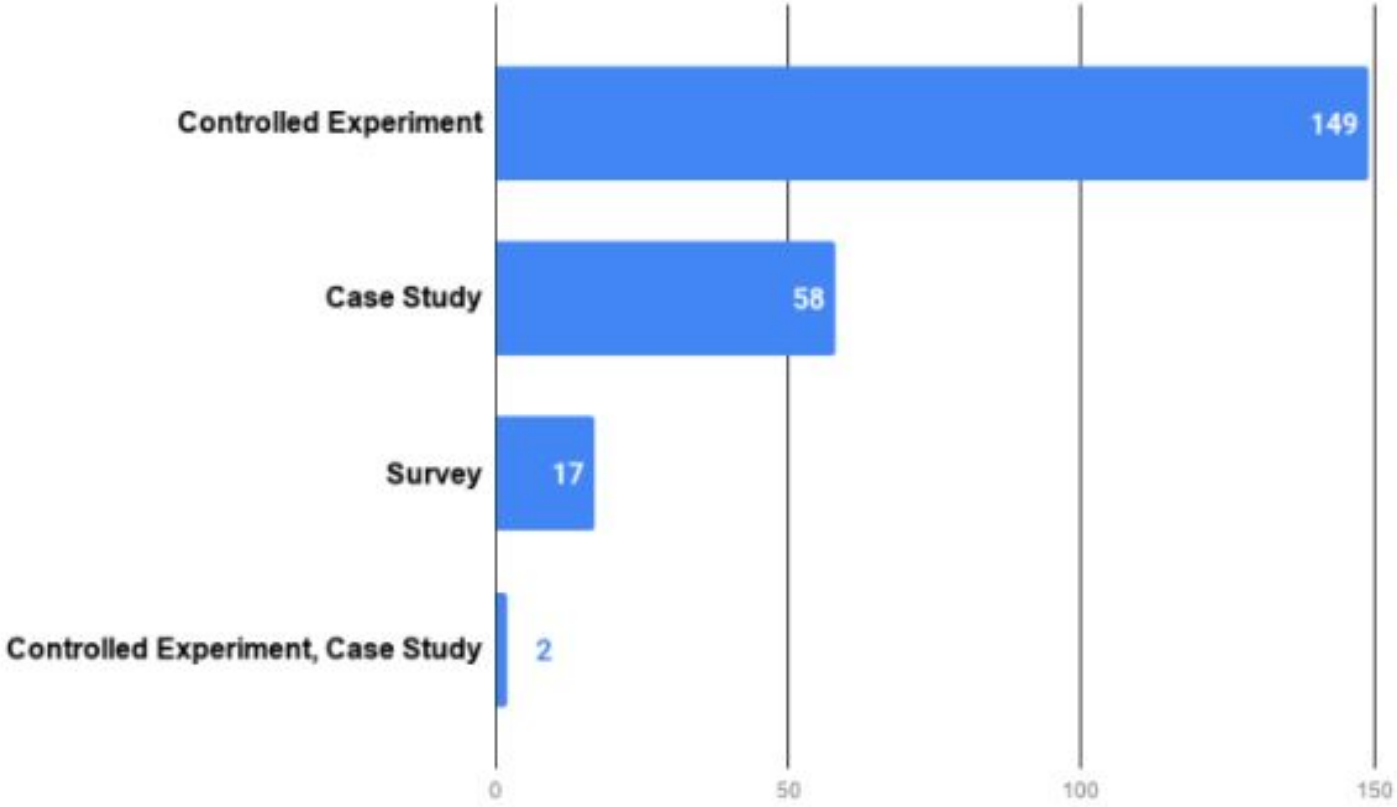
Tipo de Contribuição dos Artigos



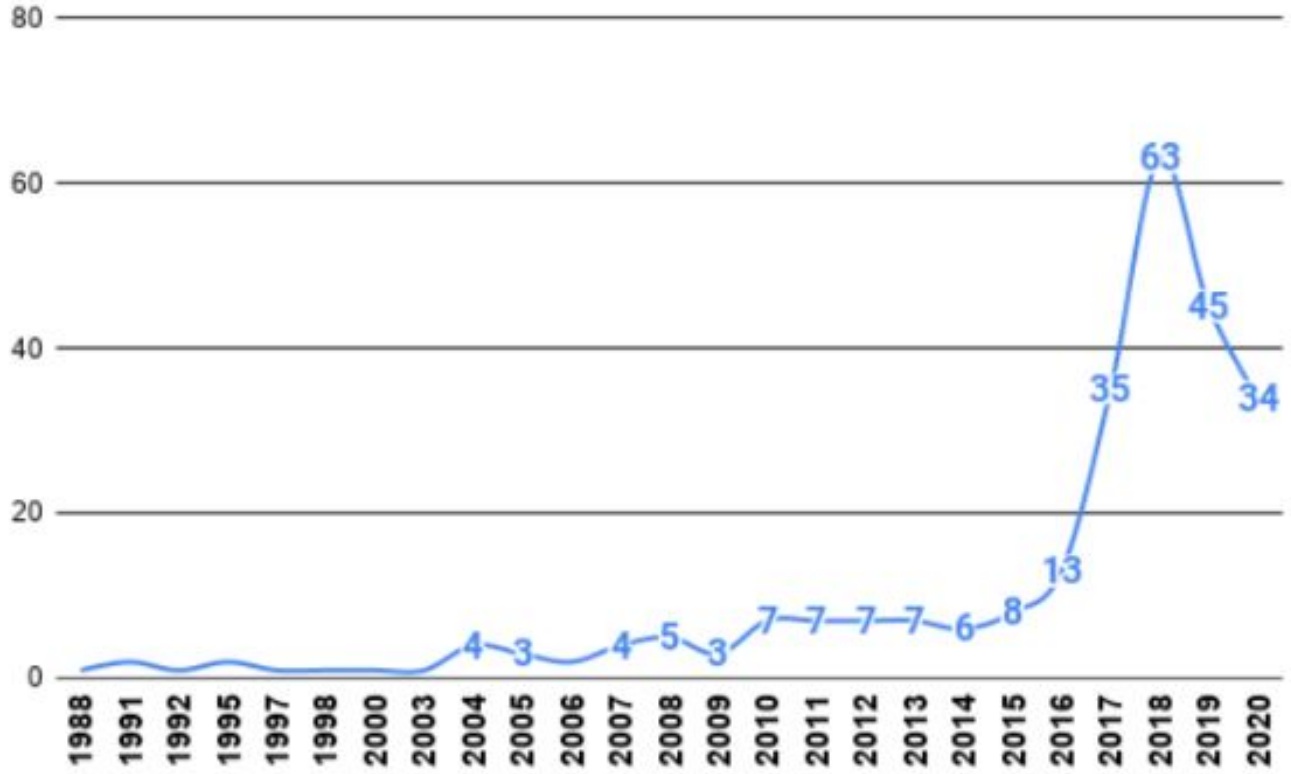
Tipo de Contribuição dos Artigos



Tipo de Pesquisa dos Artigos

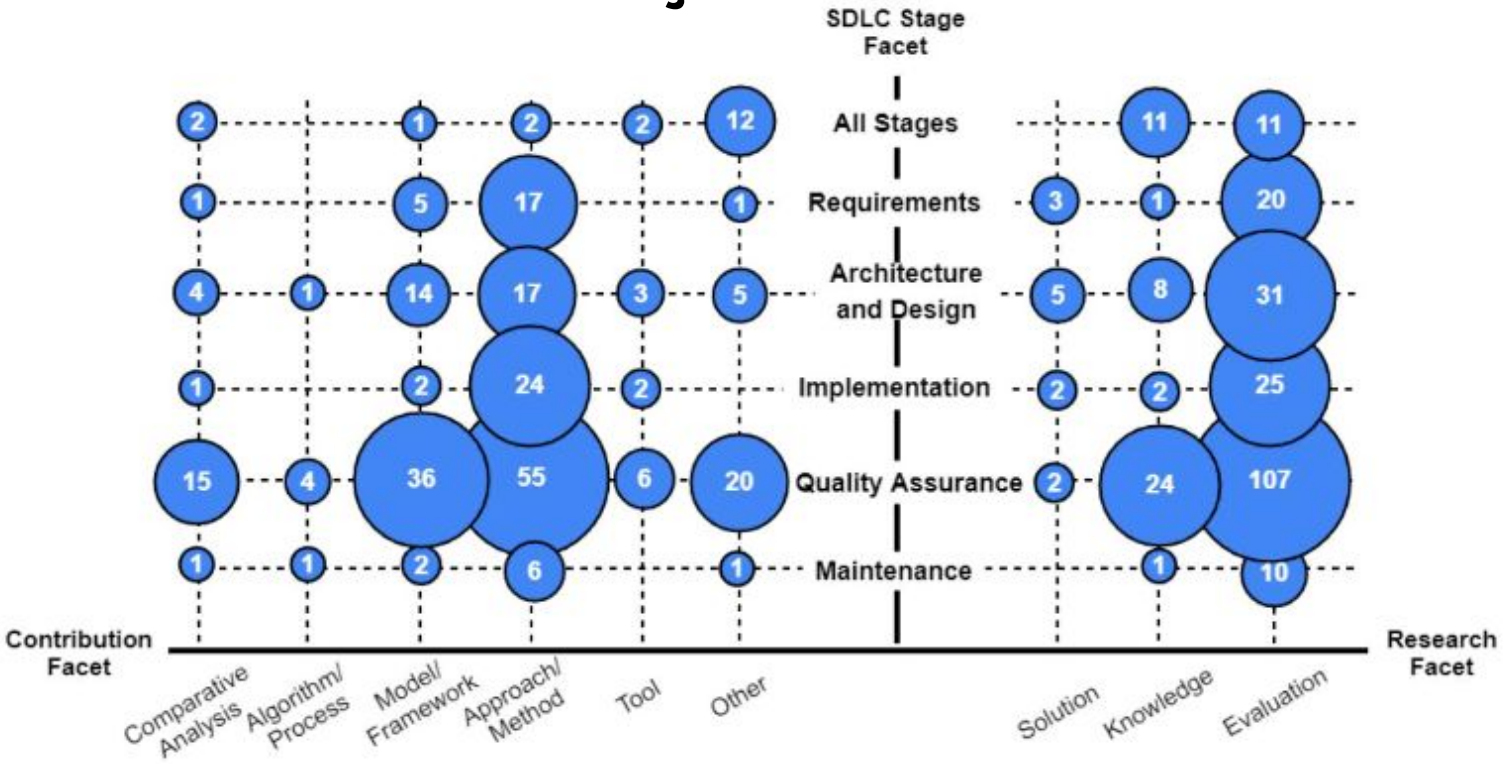


Quantidade de Artigos por Ano Nesse Tema

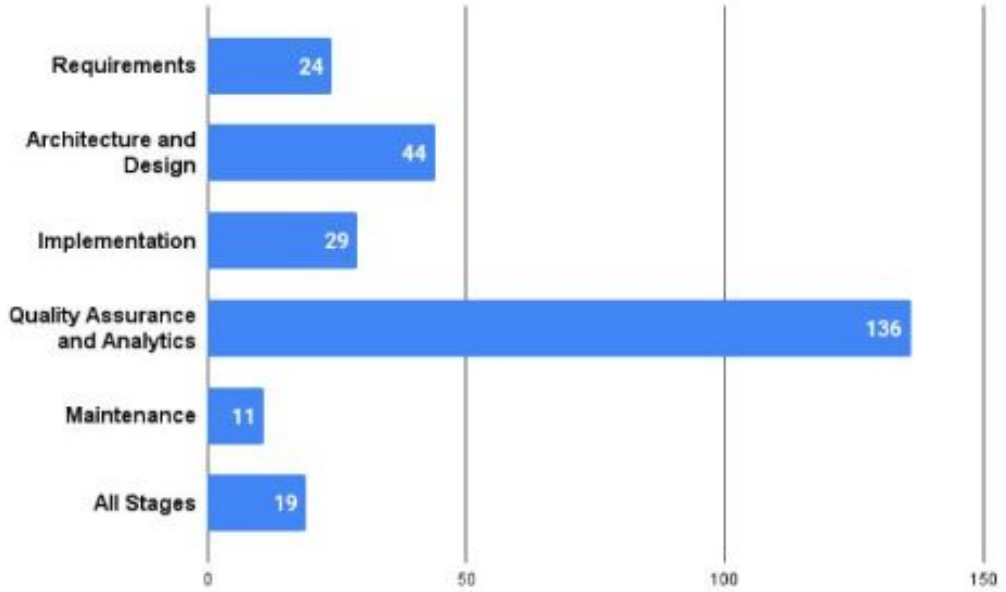


Conferências com Mais Publicações Nesse Tema

Relação da estapas de SDLC com pesquisa e formas de contribuição



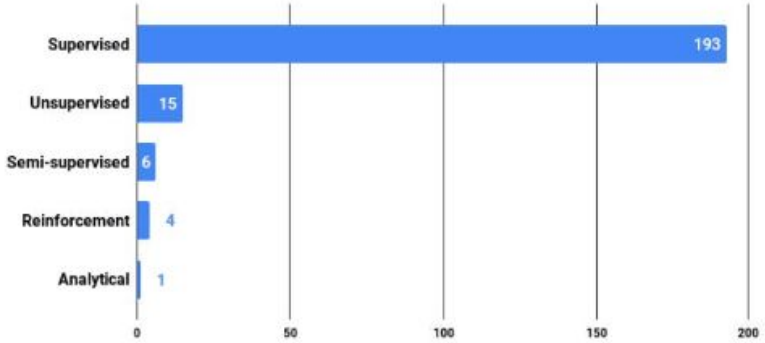
Relação da estapas de SDLC com ML

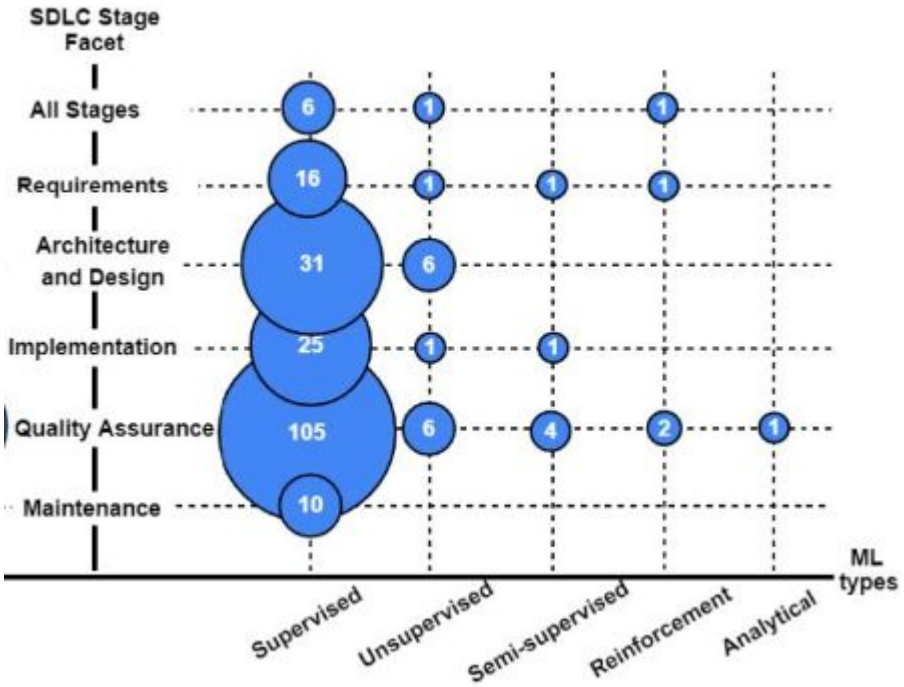


Relação da estapas de SDLC com ML

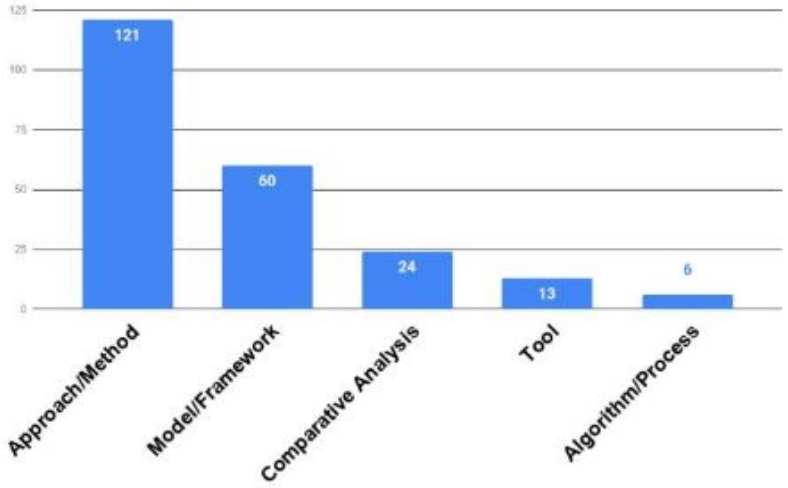
SDLC Stages	Applications of ML for SE	Articles
All Stages	N/A	[1, 11, 14, 48, 49, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69]
Requirements	Requirements Modeling and Analysis	[21, 22, 51, 70, 71, 72, 73, 74, 75]
	Requirements Selection/Prioritization/Classification	[23, 24, 76, 77, 78, 79, 80, 81]
	Requirements Traceability	[3, 25, 82, 83, 84, 85, 86]
Architecture and Design	Design Modeling	[9, 26, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102]
	Design Pattern Prediction	[27, 53, 103, 104, 105, 106]
	Development Effort Estimation	[4, 5, 28, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123]
Implementation	Code Clone/Localization/Refactoring/Labeling	[29, 124, 125, 126, 127, 128, 129, 130, 131, 132]
	Code/Bad smell detection	[30, 31, 133, 134]
	Code Inspection/Analysis	[32, 135, 136, 137, 138, 139, 140, 141, 142]
	Code/Program Similarity	[33, 143, 144, 145, 146, 147]
Quality Assurance and Analytic	Fault/Bug/Defect Prediction	[7, 34, 35, 36, 37, 38, 39, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203]
	Test Case/Data/Oracle Generation	[40, 54, 204, 205, 206, 207, 208, 209]
	Test Case Selection/Prioritization/Classification	[41, 210, 211, 212, 213]
	Vulnerability/Anomaly/Malware Discovery/Analysis	[42, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232]
	Software Analysis	[43, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243]
	Technique Assessment	[244, 245, 246, 247, 248]
	Software Process Assessment	[249, 250, 251]
	Verification and Validation	[44, 246, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265]
	Testing Effort Estimation	[6, 266, 267, 268]
Maintenance	Software Maintainability Prediction	[45, 269, 270, 271]
	Software Aging Detection	[46, 272, 273, 274, 275]
	Maintenance Effort Estimation	[47, 276]

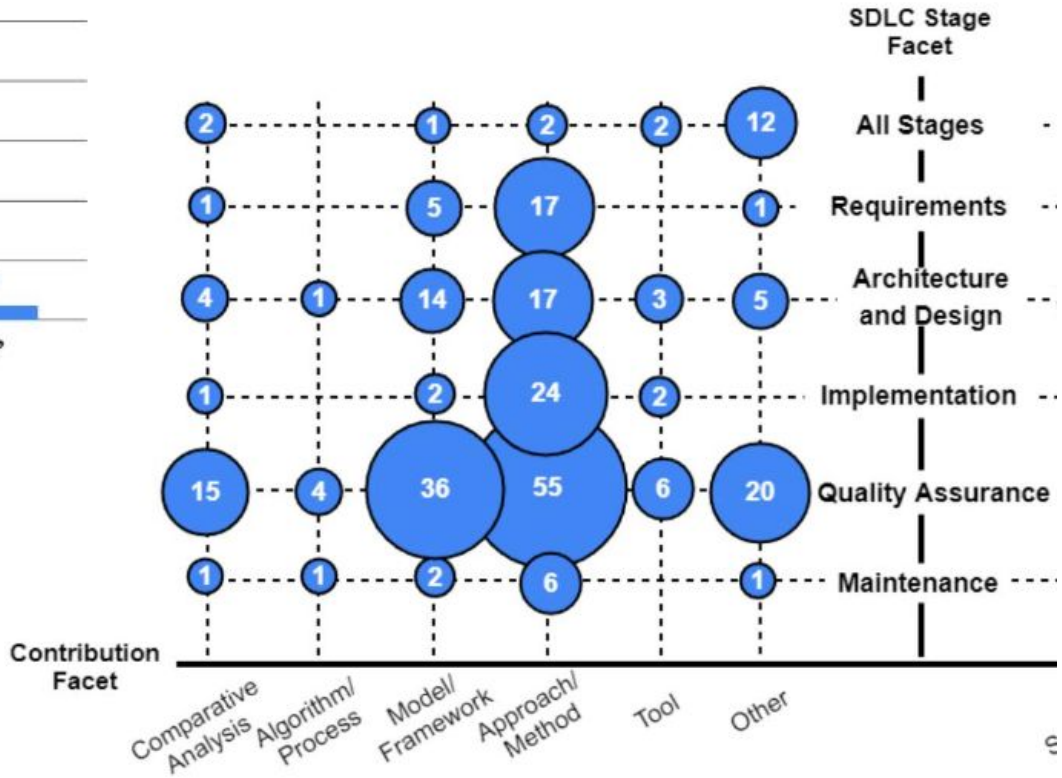
Relação da estapas de SDLC com tipos de ML



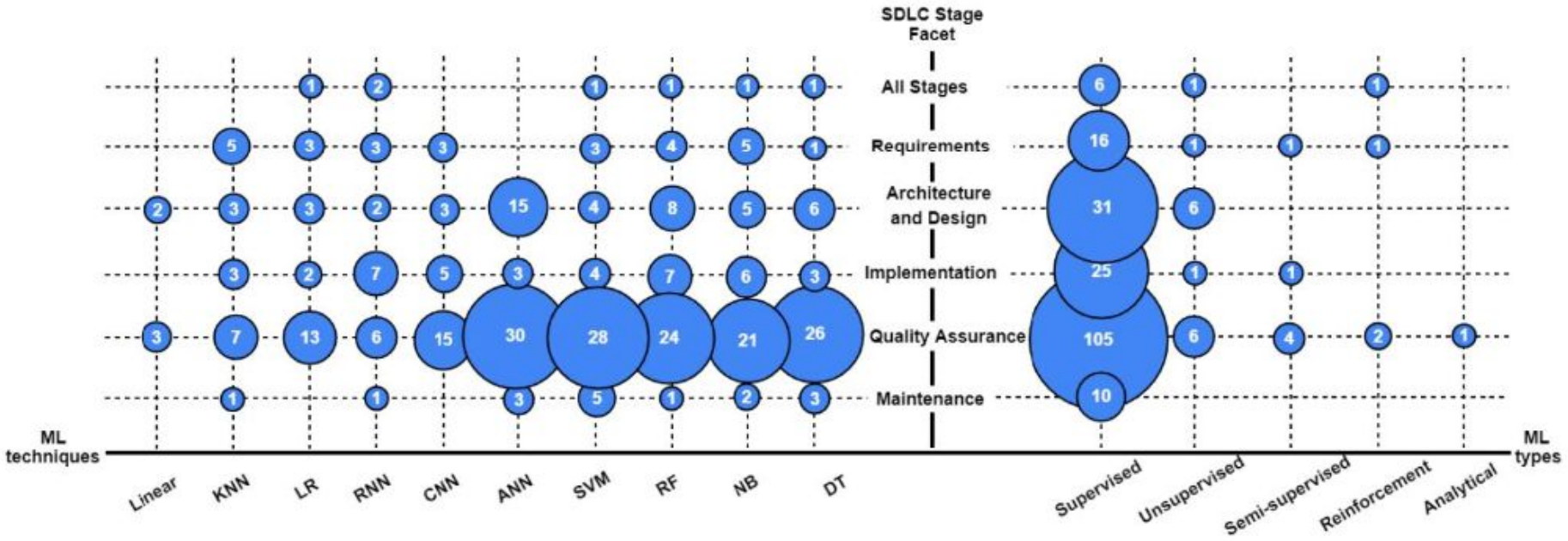


Relação da estapas de SDLC com ferramentas de ML





Relação da estapas de SDLC com técnicas de ML



Desafios, Limitações e Pesquisas Futuras

- Um dos maiores **desafios** é a natureza incerta e estocástica das técnicas de ML utilizadas e a diferença dos dados utilizados e das aplicações necessárias.
- Há também dificuldade em **obter dados em grande quantidade, estruturados e categorizados**.
- **Poucos dados** levam ao problema de *overfitting* das soluções de ML.
- Pesquisas futuras devem focar em **expandir os conjuntos de dados** existentes e em buscar **métricas mais condizentes** com o tema em questão.
- A busca dos artigos também sofre com a **falta de generalização dos resultados**, e essa ameaça também deve ser levada em consideração.