RMP
Revista dos Mestrados Profissionais
UFPE / CCSA – MGP
https://periodicos.ufpe.br/revistas/rmp
ISSN – 2317 - 0115
V. 13 - n. 2 (2024)
Recife – PE

NeoDataset: um conjunto de dados com *user stories e story points*

Giseldo da Silva Neo – Instituto Federal de Alagoas
giseldo.neo@ifal.edu.br
https://orcid.org/0000-0001-5574-9260

Alana Viana Borges da Silva Neo - Instituto Federal do Mato Grosso do Sul
alana.neo@ifms.edu.br
https://orcid.org/0009-0000-1910-1598

Kleber Jose Araújo Galvão Filho – Universidade Federal de Alagoas
kleber.filho@ifal.edu.br
https://orcid.org/0009-0001-1591-0272

José Antão Beltrão Moura– Universidade Federal de Campina Grande
antao@computacao.ufcg.edu.br
https://orcid.org/0000-0002-6393-5722

Olival de Gusmão Freitas Junior – Universidade Federal de Alagoas
olival@ic.ufal.br
https://orcid.org/0000-0003-4418-8386

Resumo – As equipes geralmente utilizam ferramentas de gerenciamento para acompanhar as *user stories*, controlar o seu código-fonte, registrar suas estimativas de esforço e os responsáveis. Essas ferramentas registram dados que podem ser utilizados em diversas pesquisas. Por outro lado, é desafiador encontrar dados para pesquisas, pois as empresas privadas são relutantes em compartilhá-los. O objetivo deste artigo é apresentar um conjunto de dados contendo dados brutos de 33 Projetos de *Software* Ágil de código aberto, minerados do GitLab, totalizando 122.627 *story points* e 20.474 *user stories*. Disponibilizamos esses dados publicamente para facilitar o seu uso pela comunidade científica. Acreditamos que esse conjunto pode ser utilizado em várias linhas de pesquisa de engenharia de *software*, incluindo classificação e vetorização de texto, aprendizagem de máquina, estimativa de esforço e priorização de tarefas.

Palavras-chave: *user story , story points*, conjunto de dados, ágil, linguagem natural.

NeoDataset - A dataset with user stories and story points

Abstract – Teams often use management tools to monitor outstanding User Stories, control their source code, record their effort estimates and those responsible for opening and closing tickets. These tools contain data that can be used in various software engineering research. It is necessary to find data for research as private companies are reluctant to share their data. This paper aims to present a dataset containing raw data from 33 open-source Agile Software Projects, mined from GitLab, totaling 122.627 Story Points and 20.474 User Stories. We make this data publicly available in CSV and JSON formats to facilitate its use by the interested scientific community. We believe this dataset can be used in multiple lines of software engineering research, including effort estimation.

Keywords: user story, story points, dataset, agile, natural language.

Data da Submissão: 24/08/2024 - **Data de aceitação:** 01/12/2024

DOI: https://doi.org.10.51359/2317-0115.2024.265431

RMP
Revista dos Mestrados Profissionais
UFPE / CCSA – MGP
https://periodicos.ufpe.br/revistas/rmp
ISSN – 2317 - 0115
V. 13 - n. 2 (2024)
Recife – PE

1. Introdução

Uma *user story* é uma técnica no desenvolvimento ágil de software, que visa capturar as necessidades dos usuários de maneira clara e acessível, facilitando a comunicação entre os membros da equipe de desenvolvimento e os stakeholders (COHN, 2005). Elas são escritas em uma forma narrativa que geralmente segue um formato simples: "Como [tipo de usuário], eu quero [funcionalidade], para que [benefício]". Este formato ajuda a garantir que o foco permaneça no valor que o usuário final obterá da funcionalidade solicitada, ao invés de nos detalhes técnicos de implementação. Ao representar requisitos de forma concisa e orientada ao usuário, as *user stories* permitem uma abordagem incremental e iterativa, adaptada às mudanças frequentes e melhorias contínuas.

Por outro lado, os *story points* são uma métrica usada para estimar o esforço necessário para implementar uma *user story*. Ao contrário das estimativas de tempo tradicionais, os *story points* se concentram na complexidade relativa de um conjunto de tarefas, incluindo o volume de trabalho, incertezas e riscos associados (COHN, 2005). Essa estimativa evita problemas relacionados à precisão inerente às estimativas de tempo e reconhece que diferentes equipes terão diferentes capacidades e velocidades de trabalho. Geralmente, a pontuação é atribuída usando uma escala numérica baseada na sequência de Fibonacci (1, 2, 3, 5, 8, e assim por diante), o que ajuda a evitar debates detalhados sobre pequenas diferenças de esforço e promove um consenso mais rápido dentro da equipe.

Uma *user story* bem elaborada fornece a base necessária para que a equipe de desenvolvimento possa discutir e compreender a tarefa, enquanto os *story points* registram a estimativa e são utilizadas para o planejamento do sprint. Esse processo, conhecido como planejamento de release ou de sprint, permite que a equipe visualize o trabalho a ser realizado em pequenas iterações, mantendo o alinhamento com as prioridades dos stakeholders e ajustando-se rapidamente às mudanças no escopo ou nos requisitos. Além disso, ao utilizar *story points*, equipes podem medir e melhorar sua velocidade ao longo do tempo, refinando suas estimativas e adaptando-se à realidade do projeto, resultando em entregas mais previsíveis e de maior qualidade (SCHWABER; SUTHERLAND, 2020).

A relevância de dados consolidados em pesquisas é amplamente reconhecida; no entanto, empresas privadas frequentemente demonstram relutância em divulgá-los. Este comportamento pode ser associado a vários motivos, incluindo preocupações com a privacidade, questões competitivas e o receio de que os dados possam ser mal interpretados ou utilizados fora do contexto (SMITH, 2017). Além disso, mesmo quando as empresas decidem compartilhar seus dados, nem sempre esses materiais são disponibilizados em um formato que facilite o acesso e a análise pelos pesquisadores (JOHNSON, 2019). Em diversos casos, os dados são desestruturados ou armazenados em sistemas proprietários que exigem ferramentas específicas e conhecimento especializado para a sua extração e manipulação (KIM; PARK, 2020).

O processo de consolidação e preparação dos dados, que é fundamental para transformar dados brutos em informações úteis e interpretáveis, ainda requer um esforço considerável. Esse desafio é amplificado pelo fato de que os dados frequentemente são armazenados em bancos de dados relacionais complexos. A extração de dados desses sistemas pode ser uma tarefa trabalhosa e demorada, envolvendo a escrita de consultas específicas em linguagens como SQL (*Structured Query Language*), a limpeza de dados para remover inconsistências e a integração de dados provenientes de múltiplas fontes. Esses passos são essenciais para garantir que os pesquisadores possam conduzir análises robustas e tirar conclusões precisas.

O acesso facilitado a dados de qualidade constitui uma componente frequente nas discussões sobre avanços na pesquisa científica, especialmente em domínios como a ciência de dados, economia e saúde (JOHNSON, 2019). No entanto, a mediação entre as necessidades dos pesquisadores e as condições impostas pelas empresas para o compartilhamento de dados continua a ser uma área que demanda atenção constante e soluções inovadoras para maximizar o aproveitamento do potencial informacional disponível (MAYER-SCHÖNBERGER, 2013; HARDT, 2019).

Para contribuir com pesquisas de engenharia de software e inspirados pela contribuição de outros conjuntos de dados (JUST, 2014; TAWOSI, 2022), disponibilizamos um novo conjunto de dados chamado NeoDataset. Ele é disponibilizado no GitHub para que toda a comunidade interessada possa contribuir, semelhante ao que acontece com outros conjuntos de dados (TOMASSI, 2019).

O objetivo deste artigo é apresentar um conjunto de dados com dados textuais do título e da descrição da *user story* e sua medição em *story points* realizada pelo time. Os dados coletados abrangem diversas equipes de desenvolvimento ao longo de vários projetos que podem revelar padrões e insights sobre práticas de estimativa e planejamento e esforço. Este novo conjunto de dados constitui uma fonte de informações para problemas de classificação de texto e para aprimorar a precisão das estimativas em projetos ágeis.

As seções seguintes apresentam a fundamentação teórica, a descrição do conjunto de dados, como ele foi extraído, suas características e estrutura, a sua originalidade e relevância e as considerações finais.

2. Referência Conceitual

Esta seção apresenta a *user story* e o *GitLab*, conceitos necessários para o entendimento de outras seções do artigo.

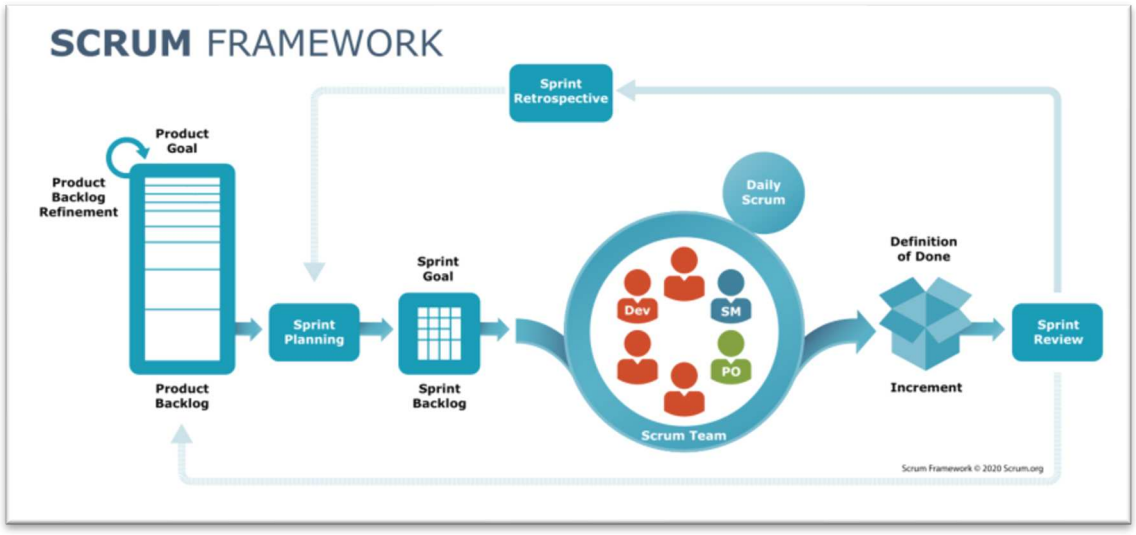
2.1 User story

Em 1993 o engenheiro de *software* Jeff Sutherland criou um processo de desenvolvimento chamado SCRUM (SUTHERLAND, 2014). Jeff trabalhava na empresa Easel Corporation e aplicou o SCRUM no projeto mais crítico dessa organização, obtendo bons resultados (SABBAGH, 2014). Em meados de 2001, ele, Ken Schwaber, Kent Beck e outros 16 pesquisadores discutiram as dificuldades no processo de desenvolvimento tradicional. No final desse encontro eles elaboram o manifesto ágil (BECK, 2001). Esse manifesto deu início ao movimento ágil, do qual o SCRUM faz parte.

O SCRUM é um método ágil baseado em ciclos de tempo fixo chamados *sprints*. Em uma *sprint* o time trabalha para atingir objetivos bem definidos. Esses objetivos são registrados em uma lista de coisas a fazer que é constantemente atualizada e re-priorizada, chamada *product backlog* (SUTHERLAND, 2014).

A característica do SCRUM é o foco em entregas rápidas com alto valor agregado em curtos períodos, lidando com as mudanças o mais rápido possível (DYBA, 2008). Esta abordagem tem mostrado resultados e tem sido utilizada em gerenciamento de projetos (PMI, 2017), tanto na indústria (TRIMBLE, 2016; RIGBY, 2018) quanto no governo (MERGEL, 2016) - Vide figura 1.

Figura 1 – Scrum Framework.



Fonte: Scrum Aliance (2024).

As *user stories* são peças centrais no registro de requisitos em times que usam o SCRUM. Essa narrativa deve ser concisa e descrever a funcionalidade desejada por um usuário em um sistema (COHN, 2009). Elas podem ser descritas em diferentes níveis de granularidade, desde funcionalidades abrangentes, até detalhes específicos da interface. Também podem ser estruturadas de diversas maneiras, porém, uma estrutura formal comumente utilizada é a proposta por (COHN, 2009) em três elementos principais:

- Quem: O tipo de usuário que necessita da funcionalidade;
- O que: A funcionalidade específica que o usuário precisa;
- Porque: A razão pela qual o usuário precisa da funcionalidade e os benefícios que ela trará.

A empresa *Mountain Goat Software* fornece 200 exemplos de *user stories* que podem ser utilizados para treinamentos ou para estudos. Essas *user stories* foram escritas durante a construção do *site* da Scrum Alliance entre 2003 e 2004. Elas estão disponíveis para *download* em formato PDF (MOUNTAIN, 2004). O quadro 1 apresenta alguns exemplos desse conjunto de dados.

Outros conjuntos de dados com *user stories* podem ser encontrados no *site* Mendeley Data. Esse *site* é um repositório que armazena vários conjuntos de dados

disponíveis para pesquisas. Um dos conjuntos de dados existentes nesse repositório consolida *user stories* de 22 projetos (DALPIAZ, 2018). O quadro 2 apresenta exemplos de *user stories* deste conjunto de dados.

Note que a descrição do texto da *user story* nos quadro 1 e quadro 2 segue a estrutura: "quem", geralmente utilizando a palavra "como"; também a estrutura "o que", geralmente utilizando as palavras: "quero", "posso", "desejo"; Por fim, a estrutura "Porque" com a palavra "para".

Quadro 1 - Exemplos de *user stories* do site Mountain Goat Software.

Descrição
Como membro do *site*, quero descrever-me na minha própria página de uma forma semiestruturada para que os outros possam aprender sobre mim. Ou seja, posso preencher campos predefinidos, mas também ter espaço para um ou dois campos de texto livre. (Seria bom deixar esse texto livre ser *Markdown* ou similar para permitir alguma formatação simples)
Como membro do site, posso preencher uma inscrição para me tornar um *Certified Scrum Practitioner* para que eu possa ganhar essa designação. *Certified Scrum Practitioner* foi o nome inicial do que ficou conhecido como *Certified Scrum Professional*
Como *Practitioner*, quero que minha página de perfil inclua detalhes adicionais sobre mim (ou seja, algumas das respostas à minha inscrição no *Practitioner*) para que eu possa mostrar minha experiência

Quadro 2 - Conjunto de dados com várias User Stories de Dalpiaz, disponível no Mendeley Data.

Descrição
Como designer de interface do usuário, desejo reprojetar a página Recursos para que ela corresponda aos novos estilos de design do *Broker*.
Como um designer de interface do usuário, quero relatar às agências sobre testes de usuários, para que eles estejam cientes de suas contribuições para tornar o *Broker* uma UX melhor.
Como designer de interface do usuário, quero passar para a rodada 2 de edições de página de destino DABS ou FABS, para que eu possa obter aprovações da liderança.

2.2 GitLab

O *GitLab* (https://gitlab.com/) é uma plataforma de desenvolvimento de *software* que combina diferentes fases do ciclo de vida de uma aplicação. Inicialmente concebido como um sistema de controle de versão utilizando *Git*, ele evoluiu para um ambiente completo que apoia desde a gestão de repositórios até a entrega contínua (KONONOV, 2018; MURPHY et al., 2020).

Entre as funcionalidades disponibilizadas pelo *GitLab*, destacam-se a gestão de projetos e repositórios, a integração contínua, a entrega contínua e a integração com outras ferramentas e serviços. A plataforma permite a colaboração entre equipes de desenvolvimento por meio de um sistema de *merge requests*, revisão de código e rastreamento de problemas (*issues*), tudo centralizado em uma interface única.

O *GitLab* pode ser utilizado tanto na nuvem, por meio de seu serviço hospedado, quanto em instalações *on-premises*, permitindo às organizações escolherem o modelo que melhor se adequa às suas necessidades de segurança e controle. Adicionalmente, o *GitLab* incorpora funcionalidades como monitoramento de desempenho, segurança de aplicações

e automação de *deploys*, oferecendo um ambiente coeso para todo o ciclo de vida do desenvolvimento. Essas características fazem do *GitLab* uma ferramenta abrangente para equipes de desenvolvimento que buscam uma integração eficiente de processos em um único arsenal de software.

A maioria das equipes que utilizam *user stories* também utiliza ferramentas de *software* para gerenciar o projeto e principalmente para manter um registro de suas *user stories* (JADHAV, 2023). Ao analisar os dados registrados por essas ferramentas podemos extrair informações para diversas pesquisas de engenharia de *software*, incluindo pesquisas sobre como melhorá-las (JIMÉNEZ et al., 2023).

O *GitLab* é uma das ferramentas de gerenciamento utilizadas por equipes ágeis para registrar *user stories* (CHOUDHURY et al., 2020). Ele permite que os engenheiros de *software* automatizem muitas ações durante o ciclo de desenvolvimento, incluindo registrar e alterar *user stories* (DIMITRIJEVI et al., 2015). No *GitLab*, a *user story* é registrada como um *issue*, e para cada *user story* são armazenadas diversas informações, como o título, a descrição e sua estimativa em *story points*. Outra ferramenta utilizada com funcionalidades semelhantes é o Jira® da empresa Atlassian.

Figura 2 - Quadrante mágico Gartner - *GitLab*

Fonte: Gatner (2024).

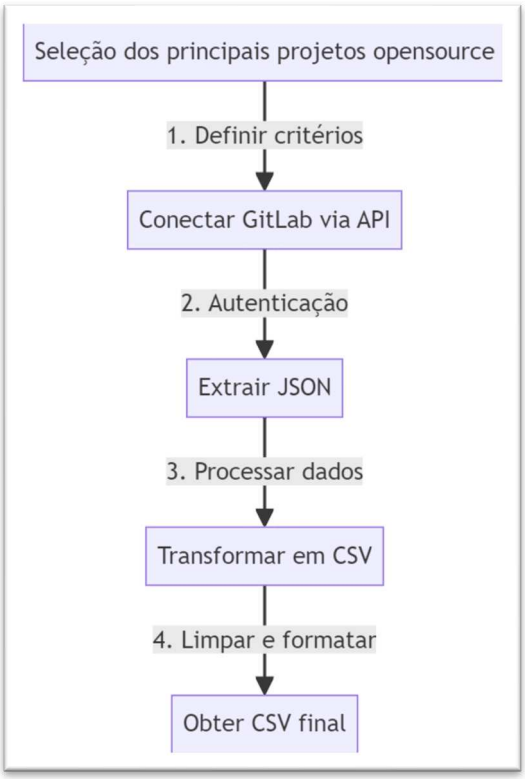
Esses dados vem sendo utilizados em diversas pesquisas relacionadas a problema em engenharia de software, tais como, atribuição de *user stories* (MANI et al., 2019), melhoria da descrição de *user stories* (CHAPARRO et al., 2017), planejamento da *sprint* (CHOETKIERTIKUL et al., 2018), análise de sentimento de desenvolvedores que escrevem as *user stories* (ORTU et al., 2015; ORTU et al., 2016; VALDEZ et al., 2020) estimativa de esforço de *user stories* (PORRU et al., 2016; SOARES et al., 2018, DRAGICEVIC et al., 2017; CHOETKIERTIKUL et al., 2019; TAWOSI et al., 2022) e

priorização de *user stories* (GAVIDIA-CALDERON et al., 2021; HUANG et al., 2021 e UMER et al., 2020). Além disso, o *GitLab* em 2023 está localizado no canto superior direito do quadrante Gartner como uma ferramenta líder em *DevOps* (figura 2). Os líderes do Quadrante Mágico das Plataformas de *DevOps* de 2023 influenciam a direção do espaço de *DevOps* com sua liderança de pensamento e serviços. Eles representam plataformas funcionalmente ricas com vários recursos e suportam o *software* durante todo o ciclo de vida de desenvolvimento.

3. Metodologia

O procedimento metodológico é composto por quatro etapas principais. Inicialmente, foram definidos os critérios de seleção dos projetos, eles foram filtrados pelo número de estrelas do repositório. Na segunda etapa, realizou-se a autenticação, e os dados foram extraídos no formato original (JSON). A terceira etapa envolveu o processamento desses dados. Finalmente, na quarta etapa, os dados foram agrupados resultando na obtenção do arquivo final no formato CSV, pronto para análise ou integração em outras plataformas, veja na figura 3. Cabe ressaltar que não foi realizado nenhum processamento de dados brutos antes da disponibilização do CSV.

Figura 3 – Procedimentos metodológicos.



Como resultado, foi produzido/consolidado um conjunto de dados que abrange informações de 33 projetos de desenvolvimento de *software*, com 20.474 *user stories* retiradas de repositórios do *GitLab*, totalizando 122.627 *story points*. Apenas *user stories*

com o atributo *state* igual a *closed* e que possuem o atributo *weight* preenchido foram coletadas. O campo *weight* é utilizado no *GitLab* para registrar o esforço em *story points*.

O conjunto de dados apresentado inclui projetos que não foram utilizados por estudos anteriores, de acordo com nossas buscas não sistemáticas em artigos. Já existem estudos prévios que extraíram dados da ferramenta de gestão *Jira* em diversas pesquisas de engenharia de *software* (TAWOSI et al., 2022; CHOETKIERTIKUL et al., 2019; PORRU et al., 2016 e SCOTT et al., 2018), contudo, projetos extraídos do *GitLab* são mais raros.

A reprodução de outros estudos, tal como o de Venkatraman *et al.* (2024) pode se beneficiar deste conjunto de dados, pois seus resultados podem ser reproduzidos para fortalecer a evidência sugerida, por exemplo de que determinado algoritmo é mais adequado para a finalidade de estimativa de esforço.

4. Resultados

4.1 Extração

Este conjunto de dados foi extraído durante o mês de janeiro de 2023 e abril de 2023. O processo de mineração teve como alvo os principais projetos de código aberto do *GitLab*. Os projetos selecionados empregam metodologias ágeis de desenvolvimento de *software* e tiveram registrado o tamanho em *story points* de suas tarefas pela equipe de desenvolvimento do projeto para minerar informações do *GitLab* foi criado uma ferramenta de extração implementada em *Python* que se conecta ao *GitLab* via API, o código dessa aplicação está disponível no *GitHub* em https://github.com/giseldo/neo-gitlab-extractor.

A figura 4 apresenta um exemplo no conjunto de dados. Ela contém as informações do "Título", "Descrição" e "SP" de três *user stories*. Na coluna "Descrição", são apresentados detalhes sobre cada problema ou tarefa. A primeira linha é a *user story* "*Posts repeat in Groups feed*" ela descreve um problema em que posts se repetem em um feed de grupos, o que parece ser uma questão em uma instância de teste. Detalhes técnicos sobre o comportamento e possíveis soluções são discutidos, como criar um novo grupo e postar uma sequência de *posts* para replicar o problema.

Ainda na figura 4, a segunda linha é a *user story* "*Expose bot accuracy scores on front end for admins*" que fala sobre a necessidade de ajudar os administradores a tomar decisões informadas ao exibir as pontuações de precisão de um *bot* na interface de administrador, juntamente com uma descrição de como essas informações devem ser apresentadas, incluindo números ao lado do avatar do usuário. A terceira linha é a *user story* "*Refactor experiments to use events instead of contexts*" ela discute a necessidade de reformular como os experimentos são realizados, sugerindo que, em vez de basear experimentos em contextos como o carregamento de uma página, eles sejam baseados em eventos específicos que os usuários desencadeiam, com propostas para melhorias no processo. Por fim, na coluna "SP", são atribuídos números de prioridade (5, 3, 8) a cada uma das tarefas descritas, indicando o esforço de cada uma delas.

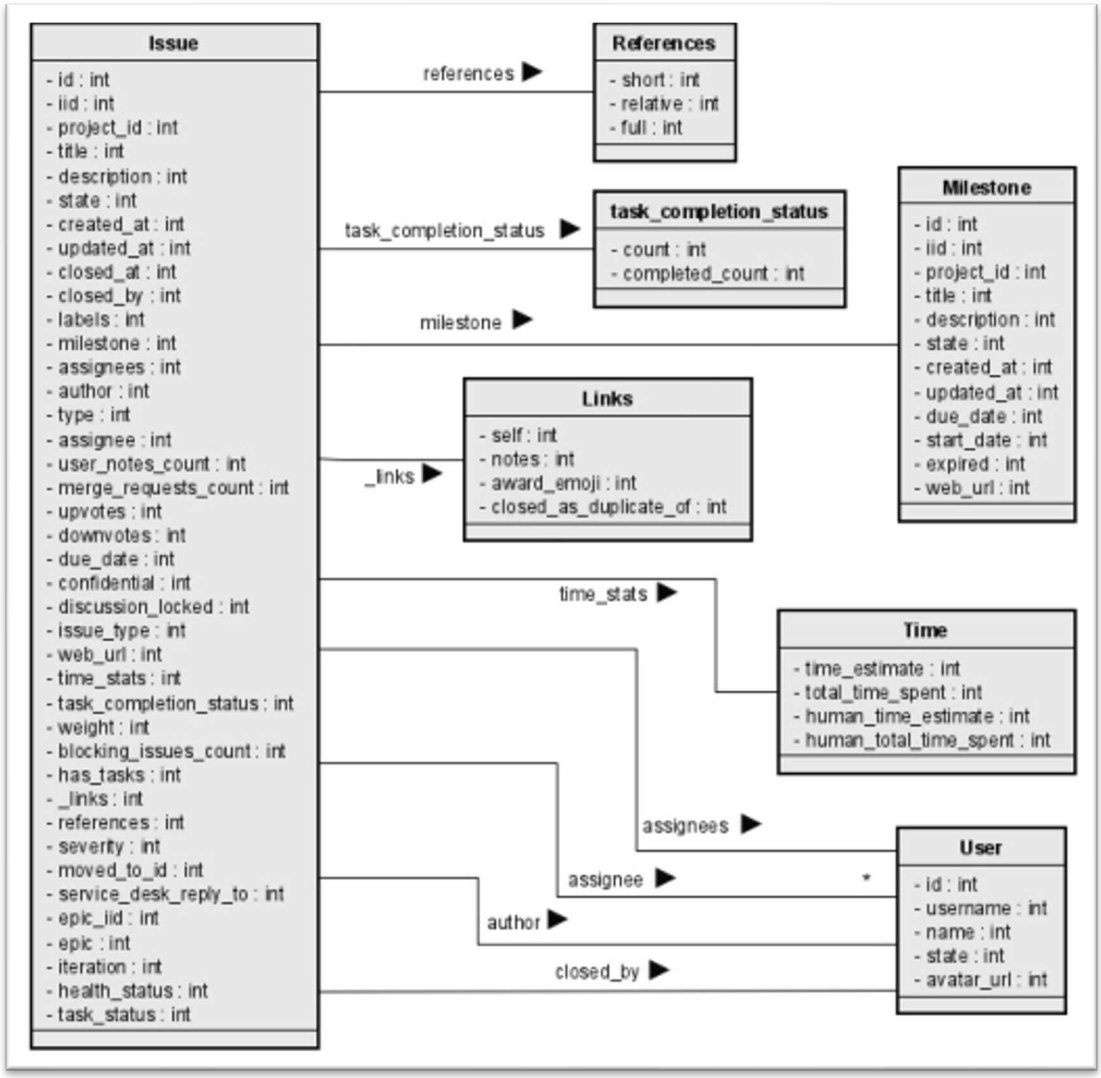
Figura 4 – Exemplo do conjunto de dados.

Título	Descrição	SP
Posts repeat in Groups feed	We have a few reports of posts repeating in groups. [This Gitlab issue](url...) makes reference to the issue in a test Group. Additionally, a user alerted me that [this Group](url...) exhibits the issue, though notably the posts repeat only after scrolling VERY deep into the timeline (~100+ posts). The pinned comment re-appears in the feed, after which the feed appears to loop. Note timestamps on the posts in this sequence: ![image](url...) ![image](url...) ![image](url...) ### Replication steps Effect can be seen here: (url...) To replicate a base-case: 1. Make a new group 2. Make a post titled 1, and a post titled 2 3. Refresh	5
Expose bot accuracy scores on front end for admins	In order to (a) help with making admin decisions, and (b) QA check the scores so that admins can see what Tasman is spitting out and provide feedback on the effectiveness of the current scoring, we'd like to make Tasman-derived trust scores visible on the front end of Minds website for users with Admin access. (url...) >Designs are in progress - Given Tasman trust scores for users are available, - and UserA is logged into Minds, - and UserA's account has Admin access, - when any user avatar + name component is displayed, - then a number is displayed alongside the avatar + name that exposes the Tasman trust score for that user. Proposed approach for displaying the correct colour based on a 0-100 value ->(url...) ### Notes * Uniquely an admin feature for the first iteration * In the future this will be visible to all users, including an explainer (via a modal) about the score, why scores are valuable and how to improve scores.	3
Refactor experiments to use events instead of contexts	## The Problem Currently all users are bucketed into experiments on app initialisation and not when they actually see the experiment. This was done so that pageviews could register the experiments as contexts when someone lands on the site. Issues arise when experiment reports rely on only including users who have seen the experiment to be included. For example, an experiment that forwards users to discovery instead of the newsfeed after registration should not just only include users that have just signed up, it should also only apply to users that do not have any referral logic (ie. take newly loggedin/registered users back to previous page). ## Proposed change - [x] Create a new iglu schema `growthbook_event` which includes the experiment_id that the user is in - [x] Remove backend registration and client on init logic - [x] Record a `growthbook_event` via snowplow when an experiment is run - [x] To avoid unnecessary events, cache a reference to the events so that they only get sent every 24 hours. - [x] Now that we support psuedo_id, attach both the anonymous_id and user_id to growthbook events	8

A figura 5 apresenta os dados *JSON* originais que foram coletados com a *API*, porém somente o título, descrição, id do projeto, id da *user story* e data da criação foram consolidados na versão final do conjunto de dados. O diagrama de entidade-relacionamento ilustra a estrutura de dados associada ao gerenciamento da *user story*.

Ainda na figura 5, a entidade central, denominada "*Issue*", engloba atributos essenciais como identificadores, título, descrição, estado, datas de criação e atualização, além de informações sobre o responsável pela tarefa e suas referências. Conectadas a essa entidade principal, estão outras entidades, como "*Task Completion Status*", que monitora o progresso da tarefa, registrando o total de subtarefas e o número concluído; e "*Milestone*", que marca etapas significativas do projeto, oferecendo informações como datas de início e término, além de descrições. Outras entidades incluem "*Links*", que armazena links associados e outras informações complementares, "Time", que acompanha o tempo estimado e o efetivamente gasto na tarefa, e "*User*", que representa os usuários do sistema com seus respectivos identificadores e informações de perfil. As relações entre essas entidades demonstram a interconexão dos dados de uma "*Issue*" com diversos componentes do sistema, refletindo o modelo utilizado no *GitLab*.

Figura 5 – *JSON* Original coletado via *API*, antes da conversão para *CSV*.



4.2 Armazenamento

O conjunto de dados em sua versão final está consolidado em um único arquivo no formato *CSV*. O formato *CSV* é amplamente utilizado para armazenar e trocar dados tabulares de maneira eficiente e simples. Ele é um dos formatos mais comuns e versáteis para armazenar dados tabulares. Devido à sua simplicidade, é adotado em diversas áreas acadêmicas, comerciais e tecnológicas. A principal característica que define um arquivo *CSV* é que ele representa dados em forma de texto, onde cada linha corresponde a um registro ou linha da tabela, e os valores dentro de uma linha são separados por vírgulas.

4.3 Características

Os projetos do conjunto de dados possuem diferentes características, abrangem diferentes linguagens de programação, diferentes domínios de negócio e diferentes localizações geográficas da equipe. Não foi realizado nenhum processo de limpeza ou transformação nas colunas.

O quadro 3 apresenta uma descrição dos atributos (também chamados de variáveis ou colunas) do conjunto de dados além de sua categorização, se quantitativo ou qualitativo. O *idproject* é um número que identifica aquele projeto. A *issuekey* é um número chave que identifica aquela *user story* , o campo *created* é a data/hora de criação da *user story*, o campo *title*, é o título daquela *user story* em linguagem natural, o campo *description* é a descrição da *user story* escrito em linguagem natural, o campo *story point* é a quantidade de *story points* registrada pela equipe de projeto para aquela *user story*.

Quadro 3 – Descrição dos atributos do conjunto de dados.

Nome	Descrição	Categorização
idproject	Chave de identificação do projeto	Qualitativa (campo chave)
issuekey	Chave de identificação da u*ser story*	Qualitativa (campo chave)
created	Data de criação da *User story*	Qualitativa
title	Título da u*ser story*	Texto em linguagem natural
description	Descrição completa da u*ser story*	Texto em linguagem natural
storypoint	Quantidade de *story point* daquela u*ser story*	Quantitativa Ordinal

A tabela 1 apresenta a estatística descritiva do conjunto de dados, incluindo quantidade, média, desvio padrão e percentis, agrupados pelo identificador do projeto. Os projetos utilizam escalas de *story points* distintas; por exemplo, o projeto com o *idproject* 28419588 tem um valor máximo de *story points* de 300, enquanto o projeto 7764 apresenta um valor máximo de 128. Além disso, em alguns projetos, observa-se *story points* com valor zero, o que pode indicar a ausência de esforço associado.

4.4 Disponibilidade

Existem diversas plataformas disponíveis para a disponibilização de conjuntos de dados, entre as quais destacam-se o *HuggingFace* (https://huggingface.co/) e o *Mendeley Data* (https://data.mendeley.com/). O *NeoDataset* está acessível em ambas as plataformas.

O conjunto de dados disponível no *HuggingFace* pode ser acessado no endereço: https://huggingface.co/datasets/giseldo/neodataset. O *HuggingFace* é uma plataforma reconhecida por oferecer uma vasta gama de recursos para a manipulação e compartilhamento de aplicativos e conjuntos de dados relacionados a aprendizagem de máquina. Grandes modelos de linguagem e conjuntos de dados utilizados são acessíveis a partir desta ferramenta.

Os conjuntos de dados hospedados no *HuggingFace* são facilmente acessíveis por pesquisadores e desenvolvedores ao redor do mundo, aumentando a visibilidade e uso de seus dados. Disponibilizar um conjunto de dados no *HuggingFace* pode ser benéfico para a comunidade científica de aprendizagem de máquina, uma vez que pode promover a transparência, a replicabilidade e a inovação.

Tabela 1 – Estatística descritiva.

idproject	count	mean	std	min	25%	50%	75%	max
7764	355.0	2.732394	7.186298	0.0	1.00	2.0	3.0	128.0
250833	13.0	2.692308	2.213015	1.0	1.00	2.0	3.0	9.0
734943	121.0	2.190083	2.608684	1.0	1.00	1.0	3.0	20.0
1304532	1724.0	2.621810	2.487977	0.0	1.00	2.0	3.0	21.0
1714548	154.0	2.597403	1.510397	1.0	1.00	2.0	3.0	8.0
2009901	171.0	2.345029	1.460546	1.0	1.00	2.0	3.0	13.0
2670515	1574.0	1.993011	1.226670	0.0	1.00	2.0	2.0	15.0
3828396	15.0	2.066667	1.162919	1.0	1.00	2.0	2.0	5.0
3836952	103.0	26.902913	38.626293	0.0	2.50	5.0	32.0	100.0
4456656	344.0	1.988372	1.551782	0.0	1.00	1.0	3.0	13.0
5261717	106.0	1.745283	0.744009	1.0	1.00	2.0	2.0	4.0
6206924	42.0	2.166667	0.960606	1.0	1.25	2.0	3.0	4.0
7071551	310.0	1.935484	2.005421	0.0	1.00	2.0	2.0	32.0
7128869	327.0	4.143731	3.562003	0.0	2.00	3.0	5.0	21.0
7603319	237.0	4.670886	5.340257	0.0	1.00	4.0	8.0	40.0
7776928	216.0	2.046296	1.156452	1.0	1.00	2.0	2.0	10.0
10152778	521.0	3.059501	4.173901	0.0	1.00	2.0	3.0	80.0
10171263	982.0	3.810591	3.652708	0.0	2.00	3.0	5.0	32.0
10171270	1845.0	3.086721	3.244228	0.0	1.00	2.0	4.0	40.0
10171280	2796.0	2.579757	2.710954	0.0	1.00	2.0	3.0	50.0
10174980	178.0	2.820225	2.759930	1.0	1.00	2.0	3.0	15.0
12450835	424.0	13.674528	22.051256	4.0	6.00	6.0	12.0	260.0
12584701	171.0	6.467836	7.389409	1.0	2.00	4.0	8.0	48.0
12894267	285.0	4.978947	5.026645	0.0	2.00	4.0	6.0	40.0
14052249	167.0	3.047904	2.895381	1.0	2.00	2.0	3.0	20.0
14976868	113.0	6.380531	7.624348	1.0	2.00	4.0	8.0	42.0
15502567	3284.0	10.563946	14.193085	4.0	6.00	6.0	10.0	268.0
19921167	420.0	4.290476	4.026927	1.0	2.00	3.0	5.0	40.0
21149814	1942.0	2.918641	1.507178	1.0	2.00	3.0	3.0	20.0
23285197	284.0	1.785211	1.455947	0.0	1.00	1.0	2.0	13.0
28419588	144.0	123.611111	48.740947	100.0	100.00	100.0	100.0	300.0
28644964	102.0	139.215686	63.177986	100.0	100.00	100.0	200.0	300.0
28847821	1004.0	2.654382	2.017143	0.0	1.00	2.0	4.0	14.0

Fonte: dados da pesquisa (2024).

Além disso, o conjunto de dados disponível no *Mendeley Data* pode ser acessado em https://huggingface.co/datasets/giseldo/neodataset. O *Mendeley Data* é uma plataforma projetada para a preservação e compartilhamento de dados científicos, promovendo a transparência, a reprodutibilidade e a colaboração interdisciplinar. A plataforma permite que pesquisadores carreguem conjuntos de dados de qualquer

tamanho e formato, acompanhados de metadados detalhados que facilitam a descoberta e reutilização.

Ao disponibilizar dados de forma aberta e acessível, os pesquisadores permitem que outros estudiosos possam validar resultados, realizar meta-análises ou até mesmo identificar novas linhas de investigação. Assim, a publicação de dados no ecossistema *Mendeley* e *HuggingFace* contribui para o avanço da ciência aberta e colaborativa, beneficiando tanto os autores originais quanto a comunidade científica ao redor do mundo.

4.5 4.5 Exemplo em *Python*

Um exemplo de como utilizar baixar o repositório do *HuggingFace* com código em linguagem de programação *Python*, com o suporte da biblioteca *Pandas* (PANDAS, 2023) é apresentado na figura 6. O código faz o *download* do conjunto de dados para o computador onde está sendo executado e exibe suas primeiras linhas. A saída do comando é a tabela 2, nesta figura são exibidas as 5 primeiras linhas do conjunto de dados.

Figura 6 – Primeiras colunas do conjunto de dados.

```
1. import pandas as pd
2. df = pd.read_csv("hf://datasets/giseldo/neodataset/issues.csv")
3. df.head()
```

Tabela 2 – Saída das primeiras colunas do conjunto de dados de um do projeto do conjunto de dados.

	idproject	issuekey	created	title	description	storypoints
0	10152778	19541701	2019-03-28 15:04:16.070	(feat): change 'from' email address to 'no-rep...	Emails should not point to info@minds.com	1.0
1	10152778	19273465	2019-03-20 02:11:19.496	(bug): Link previews disappear when editing a ...	### Summary\r\n\r\nWhen editing a status post ...	3.0
2	10152778	18774779	2019-03-04 11:21:14.993	(bug): subscribed blog post duplicates filling...	NaN	2.0
3	10152778	18438523	2019-02-20 16:14:17.411	(bug): Cant delete a reply to a comment on a b...	### Summary\r\n\r\nOn ones own blog post, whil...	5.0
4	10152778	18177704	2019-02-12 16:55:06.846	Counter in group convo-feeds counts 2 seconds ...	### Summary\r\n\r\nWhen a comment is is left i...	5.0

Fonte: Dados da pesquisa (2024).

4.6 Possíveis aplicações na aprendizagem de máquina

Os dados brutos de *user stories* e *story points*, extraídos de repositórios do *GitLab*, podem alimentar algoritmos de aprendizado de máquina. Esse conjunto de dados permite o desenvolvimento de modelos preditivos para estimar o esforço necessário em tarefas ágeis, além de suportar o aprimoramento de técnicas de processamento de linguagem natural para a classificação e priorização de tarefas de software.

O impacto desse tipo de exploração no futuro do trabalho em engenharia de *software* pode ser significativo, já que automatizar previsões de esforço e priorização pode aumentar a eficiência das equipes, permitindo uma melhor alocação de recursos e adaptação às mudanças em projetos complexos.

Um cenário de uso dos dados disponibilizados pode ser a criação de um modelo de aprendizado supervisionado que, ao ser treinado com o NeoDataset, seja capaz de prever *story points* para novas *user stories* com base em sua descrição textual. Essa ferramenta pode ser integrada diretamente em plataformas de gerenciamento de projetos como *GitLab* ou *Jira*, oferecendo sugestões automáticas para equipes de desenvolvimento

ao planejar sprints, facilitando assim o trabalho colaborativo em grandes equipes distribuídas. Como exemplo o *user story tutor* (NEO et al., 2024) que propôs um modelo preditivo utilizando o NeoDataset disponível no *HuggingFace*.

5. Ameaças a validade

As ameaças à validade deste estudo podem incluir:

- Viés de seleção dos projetos: Os projetos selecionados para a extração dos dados não são representativos de todos os projetos de desenvolvimento de software. Isso pode limitar a generalização dos resultados obtidos;
- Viés de seleção das *user stories*: A coleta de *user stories* que estão fechadas (situação *closed*) e possuem atributo de *story Points* preenchido pode introduzir um viés, pois pode excluir uma parte significativa das *user stories* que ainda estão em andamento ou não possuem estimativas feitas pela equipe;
- Precisão dos *story points*: Os dados de *story points* são baseados em estimativas feitas pelas equipes de desenvolvimento. Essas estimativas podem conter erros ou serem influenciadas por fatores subjetivos, o que pode impactar a precisão das análises feitas com esses dados;
- Dependência da ferramenta *GitLab*: A extração dos dados é baseada na integração com a *API* do *GitLab*. Qualquer alteração na *API* ou na estrutura dos dados pode impactar a coleta e a disponibilidade dos dados futuramente;
- Relevância das informações: Embora os dados coletados possam fornecer insights sobre práticas de estimativa e planejamento de esforço, pode haver outros fatores além das *user stories* e dos *story points* que influenciam essas práticas. Portanto, é importante considerar outras fontes de informação para obter uma visão completa do processo de desenvolvimento ágil de *software*;
- Generalização dos resultados: Os resultados obtidos com este conjunto de dados específico podem não se aplicar a todos os contextos de desenvolvimento ágil de *software*. É importante considerar as características específicas dos projetos e das equipes ao interpretar e generalizar os resultados encontrados.

6. Trabalhos relacionados

Diversos trabalhos na área de engenharia de *software* têm se dedicado à criação e disponibilização de conjuntos de dados que apoiam a pesquisa em metodologias ágeis, principalmente no que tange à análise de *user stories* e estimativas de esforço. Um dos exemplos mais notáveis é o trabalho de Just (2014), que disponibilizou um conjunto de dados focado em métricas de código-fonte e bugs, extraídos de repositórios GitHub. Embora relevante, esse conjunto de dados não explora diretamente as *user stories* ou *story points*, deixando uma lacuna para análises mais focadas nas práticas ágeis de desenvolvimento de *software*.

Tawosi *et al*., (2022) apresentou um estudo que extraiu dados do *Jira*, uma das ferramentas mais utilizadas para gestão de projetos ágeis, oferecendo *insights* sobre o

planejamento de sprints e a estimativa de esforço em ambientes corporativos. No entanto, a maioria dos dados coletados por aqueles autores, se restringe a contextos corporativos e privados, onde o acesso aos dados é limitado Em contrapartida, o NeoDataset contribui significativamente ao disponibilizar um conjunto de dados oriundo de projetos de código aberto no *GitLab*, ampliando a possibilidade de pesquisa em ambientes não corporativos e em projetos onde os dados são mais acessíveis e transparentes.

Outro trabalho relevante é o de Porru *et al*. (2016), que investigou a eficácia das estimativas de *story points* utilizando dados extraídos do *Jira*. Similarmente, Scott *et al*., (2018) também exploraram estimativas de esforço. Por fim, vale mencionar que o uso de dados extraídos do GitLab, como no NeoDataset, é menos explorado na literatura em comparação com dados provenientes de ferramentas como o *Jira*. Isso permite novas investigações sobre práticas ágeis em diferentes contextos e contribuindo para a diversificação das fontes de dados disponíveis para a comunidade de pesquisa.

7. Considerações finais

Este conjunto de dados minerado do *GitLab* e disponibilizado em formato CSV, tem como objetivo auxiliar na educação e pesquisa sobre desenvolvimento ágil de software. Foi criado com a finalidade de treinamento e pesquisa em Estimativa de *user stories* em *story points*, mas também contém informações relevantes para outros aspectos da engenharia de *software*. Além disso, permite a replicação de descobertas de estudos anteriores.

Como trabalho futuro, sugerimos a disponibilização do conjunto de dados em diferentes formatos, a fim de aumentar sua acessibilidade e adaptabilidade a diferentes contextos de pesquisa e desenvolvimento ágil de software. Outra sugestão é disponibilizar o conjunto de dados no formato NoSQL.

Outra sugestão de trabalho futuro é converter os dados em *scripts* SQL, permitindo a sua utilização em modelos relacionais de bancos de dados. Adicionalmente, outra sugestão é oferecer um contêiner *Docker* com um Sistemas de Gerenciamento de Banco de Dados configurado, visando simplificar o uso e a adoção do conjunto de dados por parte de outros pesquisadores.

Referências

BECK, K. Extreme Programming Explained: Embrace Change. 1ª ed. Boston: Addison-Wesley, 2001.

CHAPARRO, O., LU, J., ZAMPETTI, F., MORENO, L., DI PENTA, M., MARCUS, A., BAVOTA, G., AND NG, V. Detecting missing information in bug descriptions. **Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering** Part F130154 (2017), 396–407.

CHOETKIERTIKUL, M., DAM, H. K., TRAN, T., GHOSE, A., AND GRUNDY, J. Predicting Delivery Capability in Iterative Software Development. **IEEE Transactions on Software Engineering** 44, 6 (2018), 551–573.

CHOU, P., CROWSTON, K., DAHLANDER, L., MINERVINI, M. S., AND RAGHURAM, S. GitLab: work where you want, when you want. Journal of Organization Design 9, 1 (2020).

COHN, M. **User Stories Applied: For Agile Software Development**. 1ª ed. Boston: Addison-Wesley, 2005.

DALPIAZ, F. 2018. Disponível em: https://data.mendeley.com/datasets/7zbk8zsd8y.

DIMÍTRIJEVIC, S., JOVANOVIC, J., AND DEVEDZIC, V. A comparative study of software tools for user story management. **Information and Software Technology** 57 (2015), 352–368.

DRAGICEVIC, S., CELAR, S., AND TURIC, M. Bayesian network model for task effort estimation in agile software development. **Journal of Systems and Software** 127 (2017), 109–119.

DYBÅ, T., AND DINGSØYR, T. Empirical studies of agile software development: A systematic review. **Information and Software Technology** 50, 9-10 (2008), 833–859.

GAVIDIA-CALDERON, C., SARRO, F., HARMAN, M., AND BARR, E. T. The Assessor's Dilemma: Improving Bug Repair via Empirical Game Theory. **IEEE Transactions on Software Engineering** 47, 10 (2021), 2143–2161.

HARDT, M.; NARAYANAN, A. **Data and Society: A Critical Introduction**. Cambridge: Cambridge University Press, 2019.

HUANG, Y., WANG, J., WANG, S., LIU, Z., WANG, D., AND WANG, Q. Characterizing and predicting good first issues. International Symposium on Empirical Software Engineering and Measurement (2021).

JADHAV, D., KUNDALE, J., BHAGWAT, S., AND JOSHI, J. A Systematic Review of the Tools and Techniques in Distributed Agile Software Development. Agile Software Development: Trends, Challenges and Applications (2023), 161–186.

JIMÉNEZ, S., ALANIS, A., BELTRÁN, C., JUÁREZ-RAMÍREZ, R., RAMÍREZ-NORIEGA, A., AND TONA, C. Usqa: A user story quality analyzer prototype for supporting software engineering students. **Computer Applications in Engineering Education** (2023).

JUST, R., JALALI, D., AND ERNST, M. D. Defects4J: A database of existing faults to enable controlled testing studies for Java programs. **2014 International Symposium on Software Testing and Analysis**, ISSTA 2014 - Proceedings (2014), 437–440.

KIM, D.; PARK, S. Structured versus unstructured data: Implications for data sharing among corporations. **Data Science Review**, v. 5, n. 2, p. 125-138, 2020.

KONONOV, Dmitriy. ICT for my work: **The GitLab – is it a more powerful alternative to GitHub**?. 2018. Disponível em: https://kononovdm.github.io/2018/11/GitLab-review/.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data: A Revolution That Will Transform How We Live, Work, and Think.** New York: Houghton Mifflin Harcourt, 2013.

MANI, S., SANKARAN, A., AND ARALIKATTE, R. Deeptriage: Exploring the effectiveness of deep learning for bug triaging. **ACM International Conference Proceeding Series** (2019), 171–179.

MERGEL, I. **Agile innovation management in government: A research agenda**. Government Information Quarterly 33, 3 (2016), 516–523.

MOUNTAIN SOFTWARE, G., 2004. Disponível em https://www.mountaingoatsoftware.com/uploads/documents/example-user-stories.pdf.

MURPHY, Chris et al. GitLab **CI/CD and DevOps: Reference Design and Implementation Guide for GitLab Pipelines**. 2020. Disponível em: https://docs.gitlab.com/ee/ci/examples/..

NEO, Giseldo et al. User Story Tutor (UST) to Support Agile Software Developers. In: **CSEDU (2)**. 2024. p. 51-62.

ORTU, M., DESTEFANIS, G., ADAMS, B., MURGIA, A., MARCHESI, M., AND TONELLI, R. The JIRA repository dataset: Understanding social aspects of software development. **ACM International Conference Proceeding Series** 2015-October (2015).

ORTU, M., MURGIA, A., DESTEFANIS, G., TOURANI, P., TONELLI, R., MARCHESI, M., AND ADAMS, B. The emotional side of software developers in JIRA. **Proceedings - 13th Working Conference on Mining Software Repositories**, MSR 2016 (2016), 480–483.

PANDAS Documentation. **IO Tools** (Text, CSV, HDF5, …). Disponível em: https://pandas.pydata.org/pandas-docs/stable/reference/io.html.

PMI. Success Rates Rise - **2017 9th Global Project Management Survey**. Tech. rep., PMI, 2017.

PORRU, S., MURGIA, A., DEMEYER, S., MARCHESI, M., AND TONELLI, R. Estimating story points from issue reports. **ACM International Conference Proceeding Series** (2016).

RIGBY, D. K., SUTHERLAND, J., AND NOBLE, A. Agile Scale: How to go from teams to hundreds. **Havard Business Review** May-June, June (2018), 1–3.

SABBAGH, R. Scrum: Gestão ágil para projetos de sucesso. Editora Casa do Código, 2014.

SCHWABER, K.; SUTHERLAND, J. **The Scrum Guide**. 2020. Disponível em: https://scrumguides.org.

SMITH, J. Privacy concerns in the sharing of corporate data for research purposes. **Corporate Data Journal**, v. 8, n. 1, p. 112-126, 2017.

SOARES, R. G. Effort Estimation via Text Classification And Autoencoders. In 2018 International Joint Conference on Neural Networks (IJCNN) (2018), vol. July, **IEEE**, pp. 1–8.

SUTHERLAND, J. **SCRUM: A arte de fazer o dobro de trabalho na metade do tempo**. Leya, 2014.

TAWOSI, V., AL-SUBAIHIN, A., MOUSSA, R., AND SARRO, F. Agile effort estimation: Have we solved the problem yet? insights from a replication study. **IEEE Transactions on Software Engineering** 49, 4 (2022), 2677–2697.

TAWOSI, V., AL-SUBAIHIN, A., MOUSSA, R., AND SARRO, F. A Versatile Dataset of Agile Open Source Software Projects. In **Proceedings - 2022 Mining Software Repositories Conference**, MSR 2022 (2022), pp. 707–711.

TAWOSI, V., SARRO, F., PETROZZIELLO, A., AND HARMAN, M. Multi-Objective Software Effort Estimation: A Replication Study**. IEEE Transactions on Software Engineering** 48, 8 (2022), 3185–3205.

TAWOSI, V., MOUSSA, R., AND SARRO, F. Investigating the Effectiveness of Clustering for Story Point Estimation. In Proceedings - 2022 **IEEE International Conference on Software Analysis, Evolution and Reengineering**, SANER 2022 (2022), pp. 827–838.

TOMASSI, D. A., DMEIRI, N., WANG, Y., BHOWMICK, A., LIU, Y. C., DEVANBU,P. T., VASILESCU, B., AND RUBIO-GONZALEZ, C. BugSwarm: Mining and Continuously Growing a Dataset of Reproducible Failures and Fixes. Proceedings - **International Conference on Software Engineering** 2019-May (2019), 339–349.

TRIMBLE, J., SHIRLEY, M. H., AND HOBART, S. G. Agile: From software to mission system. **14th International Conference on Space Operations**, 2016 (2016), 1–8.

UMER, Q., LIU, H., AND ILLAHI, I. CNN-Based Automatic Prioritization of Bug Reports. **IEEE Transactions on Reliability** 69, 4 (2020), 1341–1354.

VALDEZ, A., OKTABA, H., GOMEZ, H., AND VIZCAINO, A. Sentiment analysis in jira software repositories. **Proceedings - 2020 8th Edition of the International Conference in Software Engineering Research and Innovation**, CONISOFT 2020 (2020), 254–259.

VENKATRAMAN, K.; AKASHVARMA, M.; SIDDHARTH, S. Enhancing Software Test Effort Estimation using Ensemble Learning Algorithms. In: **2023 4th International Conference on Intelligent Technologies (CONIT)**. IEEE, 2024. p. 1-5.

WIKIPEDIA. CSV. Disponível em: https://pt.wikipedia.org/wiki/Comma-separated_values.