



WOMEN
IN DATA SCIENCE
— SÃO PAULO —

Preparação para o Datathon

#WiDSDatathon

A iniciativa **Women in Data Science**

O **Women in Data Science (WiDS)** visa inspirar e educar os cientistas de dados em todo o mundo, independentemente do gênero, e apoiar as mulheres na área.

O WiDS começou como uma conferência em Stanford em novembro de 2015. Agora, o WiDS inclui:

- uma conferência global, com mais de 150 eventos regionais em todo o mundo;
- um **datathon**, encorajando os participantes a aprimorarem suas habilidades;
- e um podcast, apresentando líderes no campo falando sobre seu trabalho e suas jornadas.



Quem organiza o WiDS São Paulo



Bárbara Barbosa

Lead Data Scientist na Creditas



Priscilla Wagner

Data Scientist na NeuralMed



Jéssica Santos

Data Scientist na NeuralMed



Vivian Yamassaki

Data Scientist na Creditas



Datathon

2021

Histórico do Datathon

2018

Predição de gênero

2019

Predição de plantações de óleo de palma em imagens

2020

Predição de sobrevivência de paciente

2021

Identificação de diabetes



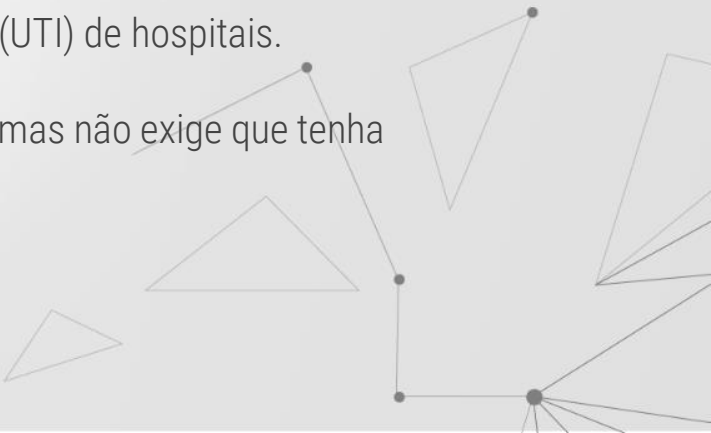


WiDS Datathon 2021

O desafio é criar um modelo com base em dados das primeiras 24 horas de tratamento intensivo para prever se um paciente tem diabetes.

A iniciativa da comunidade GOSSIS do MIT, com certificação de privacidade do Harvard Privacy Lab, forneceu um conjunto de dados com mais de 130.000 visitas de pacientes à Unidade de Terapia Intensiva (UTI) de hospitais.

O desafio é muito parecido com o do ano passado, mas não exige que tenha participado do anterior.



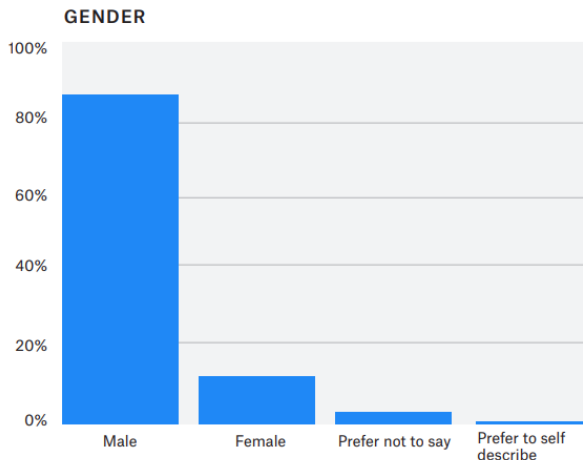
Quem pode participar do Datathon?

Data Scientist Profile

Gender

There's still a significant gender gap for data scientists, with **84% of users identifying as males**. The United States has a slightly smaller gender gap at 79%, while Japan has a slightly higher one at 90%. The results are relatively consistent regardless of region and has not changed since results of earlier Kaggle surveys.

* Com base nas respostas de 19.717 membros do Kaggle que responderam a pesquisa *Kaggle's State of Machine Learning and Data Science 2019*



Time

As inscrições podem ser individuais ou por times de até 4 pessoas.

Metade do time deve ser composto por mulheres!

Como participar do Datathon?



SE INSCREVA

Se registre em
http://bit.ly/wids_datathon2021_registration

ENTRE NO KAGGLE

Crie uma conta em
[kaggle.com](https://www.kaggle.com) e entre na
página do Datathon

FORME UM TIME

Se ainda não tem um time,
aproveite o form que
disponibilizamos nas redes
sociais para encontrar o seu
time 😊



SE CADASTRE NA MAILING LIST

Se cadastre na mailing list
do WiDS em bit.ly/mailings-wids
para receber notícias e
tutoriais sobre o Datathon



[Overview](#) [Data](#) [Notebooks](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

Manage Team

Team Name

Vivian Yamassaki [Save Team Name](#)

This name will appear on your team's leaderboard position.

Team Members

Vivian Yamassaki (you)

Leader

Invite Others

☒ Merge with other teams or invite users to your team by their team name

Team Name

[Request Merge](#)

Pending Merge Requests

You currently have no pending merge requests.

Teams Proposing a Merge

There are currently no teams proposing a merge with yours.

se
ferrame

249
Teams

422
Competitors

2,810
Entries

Points

This competition does not award standard ranking points

Tiers

This competition does not count towards tiers

Premiação

2º LUGAR

- ★ Uma inscrição gratuita por pessoa para o Stanford ICME Summer Data Science workshops
- ★ **1,5 mil dólares em dinheiro para o time vencedor**



1º LUGAR

- ★ Uma inscrição gratuita por pessoa para o Stanford ICME Summer Data Science workshops ou para o WiDS 2021
- ★ **2 mil dólares em dinheiro para o time vencedor**



3º LUGAR

- ★ Uma inscrição gratuita por pessoa para o Stanford ICME Summer Data Science workshops
- ★ **1 mil dólares em dinheiro para o time vencedor**










Submissão

Como deve ser feita a submissão

enconter_id	diabetes_mellitus
2	0.5
5	0.2
7	0.001

Como a submissão será avaliada

As submissões serão avaliadas com base na área abaixo da curva ROC (AUC) entre a mortalidade predita e o target observado (hospital_death).

Overview	Data	Notebooks	Discussion	Leaderboard	Rules	Team	My Submissions	Submit Predictions
Public Leaderboard Private Leaderboard								
This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.								
Raw Data Refresh								
#	Team Name	Notebook	Team Members	Score	Entries	Last		
1	Women Power		   	0.91439	73	1h		
2	Maria Wellen			0.91382	48	1d		
3	u++&mtmt		 	0.91374	59	2d		



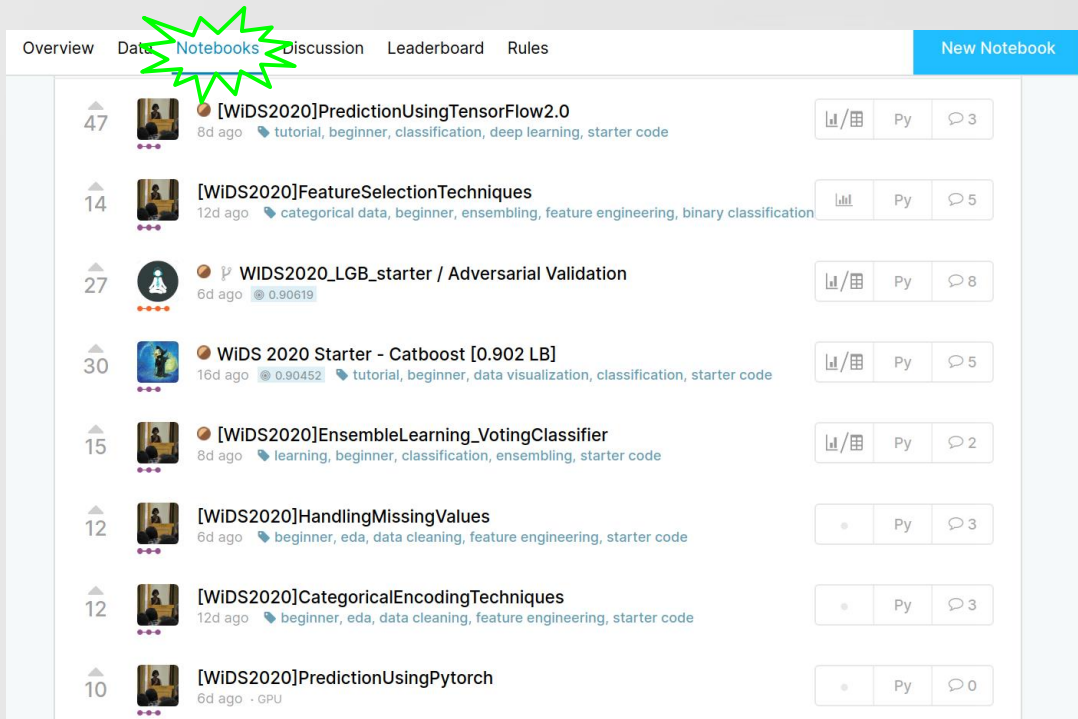
Dicas do Kaggle

1

Acompanhem os Notebooks

Nos notebooks é possível encontrar:

- ★ Análises exploratórias iniciais dos dados;
- ★ Como algumas técnicas estão se saindo;
- ★ Aprender técnicas novas;
- ★ Ver tratamentos diferentes para os dados



The screenshot shows the Kaggle interface with the 'Notebooks' tab selected. A green starburst highlights the 'Notebooks' tab in the top navigation bar. Below the navigation bar, a list of notebooks is displayed, each with a rank, a user profile picture, a title, a description, a date, a score, and a set of icons for visualization, language, and comments.

Rank	User	Title	Description	Date	Score	Visualization	Language	Comments
47	[User]	[WiDS2020]PredictionUsingTensorFlow2.0	tutorial, beginner, classification, deep learning, starter code	8d ago		Bar/Line	Py	3
14	[User]	[WiDS2020]FeatureSelectionTechniques	categorical data, beginner, ensembling, feature engineering, binary classification	12d ago		Bar/Line	Py	5
27	[User]	WiDS2020_LGB_starter / Adversarial Validation		6d ago	0.90619	Bar/Line	Py	8
30	[User]	WiDS 2020 Starter - Catboost [0.902 LB]	tutorial, beginner, data visualization, classification, starter code	16d ago	0.90452	Bar/Line	Py	5
15	[User]	[WiDS2020]EnsembleLearning_VotingClassifier	learning, beginner, classification, ensembling, starter code	8d ago		Bar/Line	Py	2
12	[User]	[WiDS2020]HandlingMissingValues	beginner, eda, data cleaning, feature engineering, starter code	6d ago			Py	3
12	[User]	[WiDS2020]CategoricalEncodingTechniques	beginner, eda, data cleaning, feature engineering, starter code	12d ago			Py	3
10	[User]	[WiDS2020]PredictionUsingPytorch		6d ago			Py	0



1

Acompanhem os Notebooks

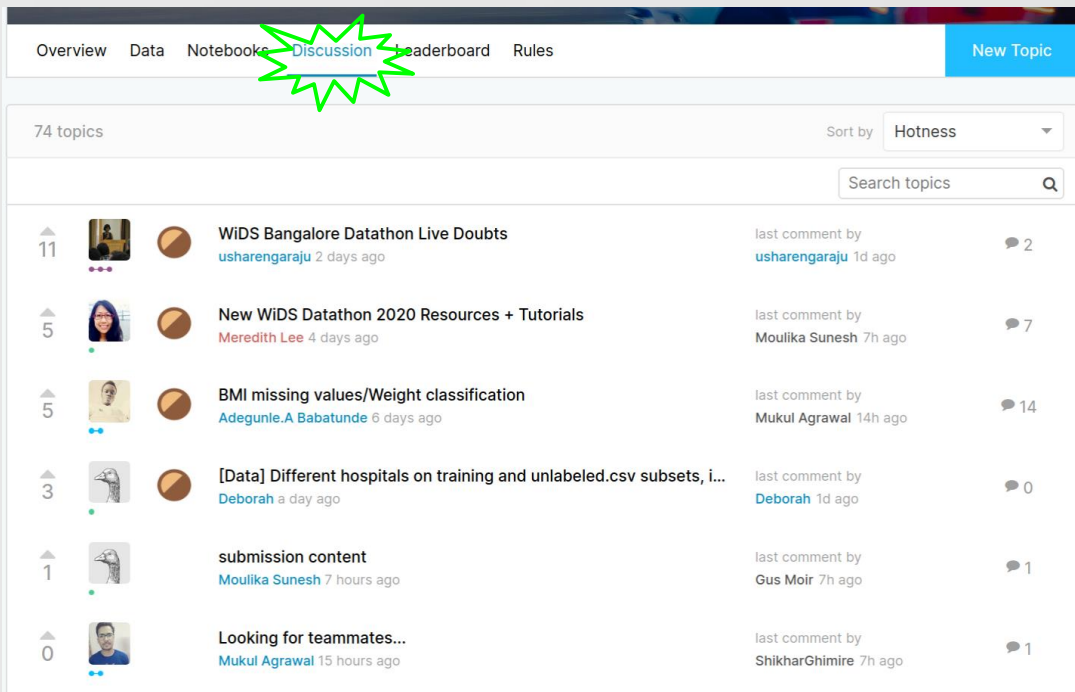
Vamos abrir [um notebook](#) da competição para ver como é!

2

Acompanhem as Discussões

Nas discussões os participantes podem falar sobre:

- ★ Problemas com o desafio;
- ★ Dificuldades com os dados;
- ★ Novas técnicas;
- ★ Dicas sobre a competição;
- ★ ...



Overview Data Notebooks **Discussion** Leaderboard Rules [New Topic](#)

74 topics Sort by Hotness

Upvotes	Profile	Topic Title	Author	Time	Last Comment	Time	Replies
11		WiDS Bangalore Datathon Live Doubts	usharengaraju	2 days ago	last comment by usharengaraju	1d ago	2
5		New WiDS Datathon 2020 Resources + Tutorials	Meredith Lee	4 days ago	last comment by Moulika Sunesh	7h ago	7
5		BMI missing values/Weight classification	Adegunle.A Babatunde	6 days ago	last comment by Mukul Agrawal	14h ago	14
3		[Data] Different hospitals on training and unlabeled.csv subsets, i...	Deborah	a day ago	last comment by Deborah	1d ago	0
1		submission content	Moulika Sunesh	7 hours ago	last comment by Gus Moir	7h ago	1
0		Looking for teammates...	Mukul Agrawal	15 hours ago	last comment by ShikharGhimire	7h ago	1



2

Acompanhem as Discussões

Vamos abrir [uma discussão](#) da competição para ver como é!

3

Caprichem na análise e no tratamento dos dados

Nesse desafio isso é ainda mais importante, já que a principal dificuldade é o tratamento de **Missing Values**. Mas o **tratamento** e **transformação** das variáveis fazem muita diferença na definição do vencedor.



4

Testem Ensembles

- Algoritmos como **XGBoost**, **CatBoost**, **Random Forest** ou diferentes combinações de algoritmos utilizando **pesos**, **médias** ou **stacking** costumam retornar os melhores resultados.

Use a criatividade nas combinações!



5

Prestem atenção na métrica de avaliação

Você pode ter desenvolvido um excelente modelo, mas ele não é bom para a métrica que o desafio está cobrando. Preste muita atenção e faça a validação do seu modelo com a métrica correta.

* A métrica do desafio atual é AUC.



6

Não deixem para testar apenas na submissão

- ★ Apesar de ter a validação do Kaggle, há um limite de submissões (não é recomendado fazer muitas) e só parte dos dados de teste vão para o leaderboard de início;
- ★ Faça sua própria validação de dados, se possível use cross-validation.



Separe um pedaço para validação



7

Usem a criatividade e compartilhem ideias

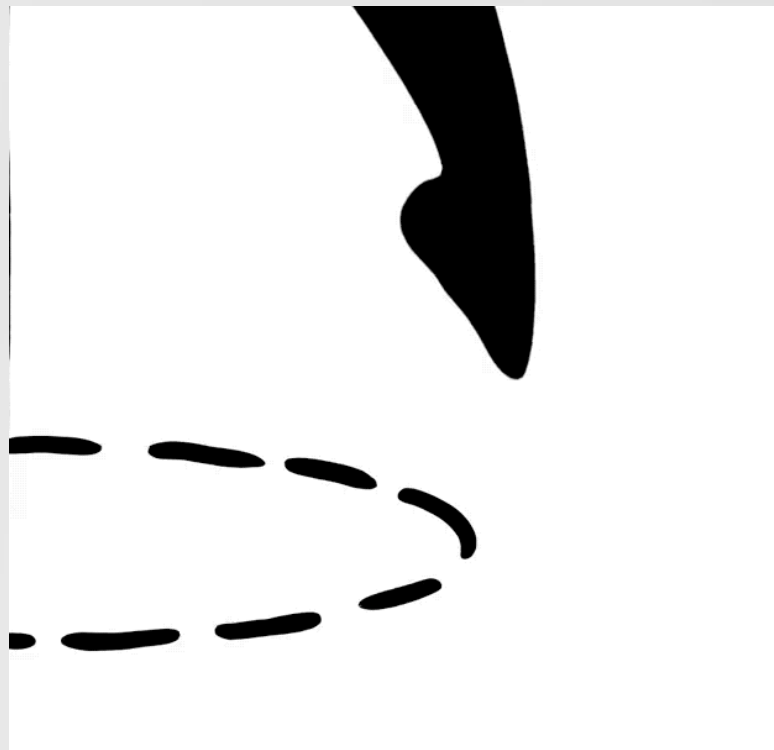
- ★ Faça combinações entre soluções diferentes entre o seu time;
- ★ Converse e discuta ideias entre vocês e nos fóruns.



8

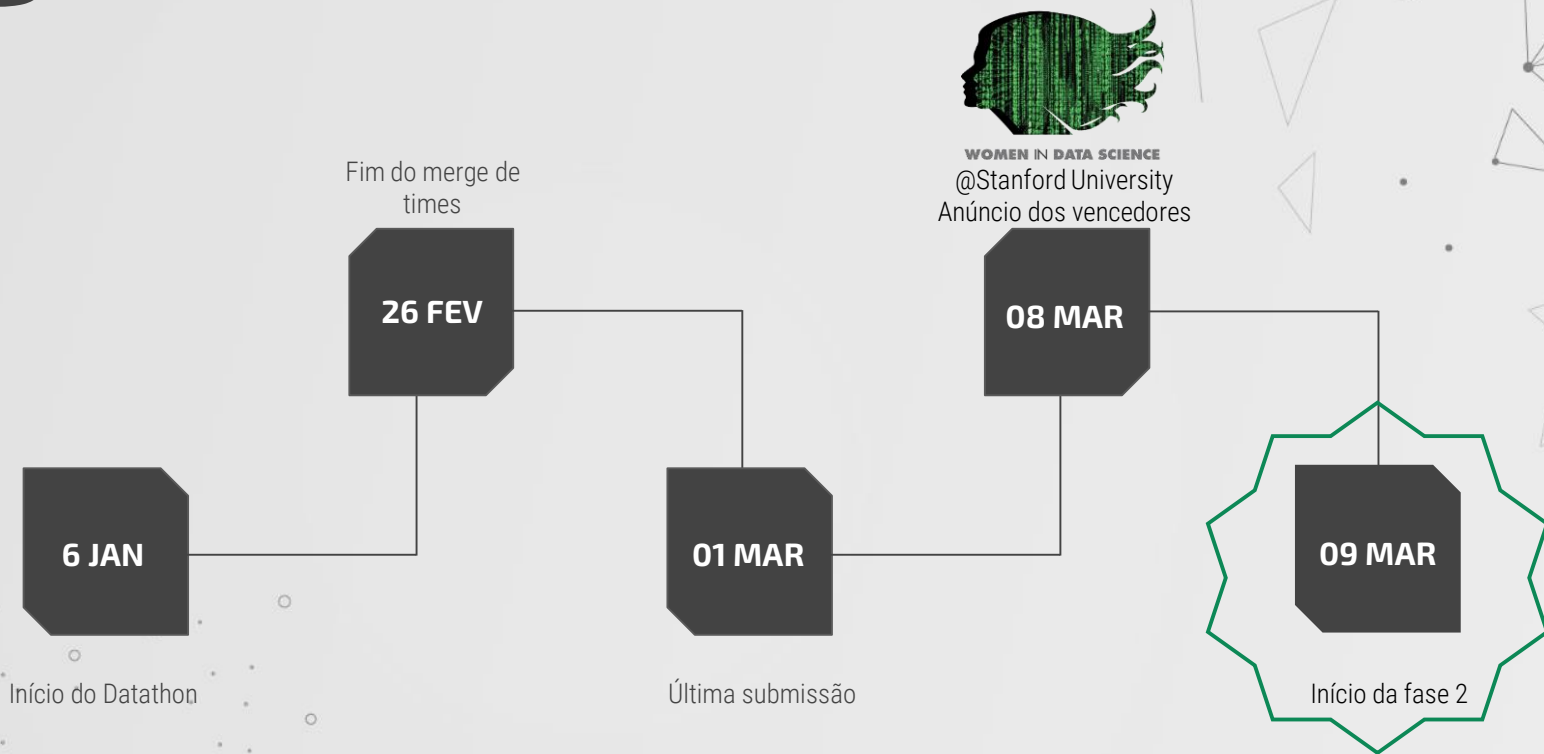
Faça melhorias aos poucos

- ★ Comece reproduzindo um notebook público;
- ★ Faça melhorias nas variáveis e veja o resultado;
- ★ Troque o método e análise;
- ★ Troque os hiperparâmetros e assim sucessivamente.



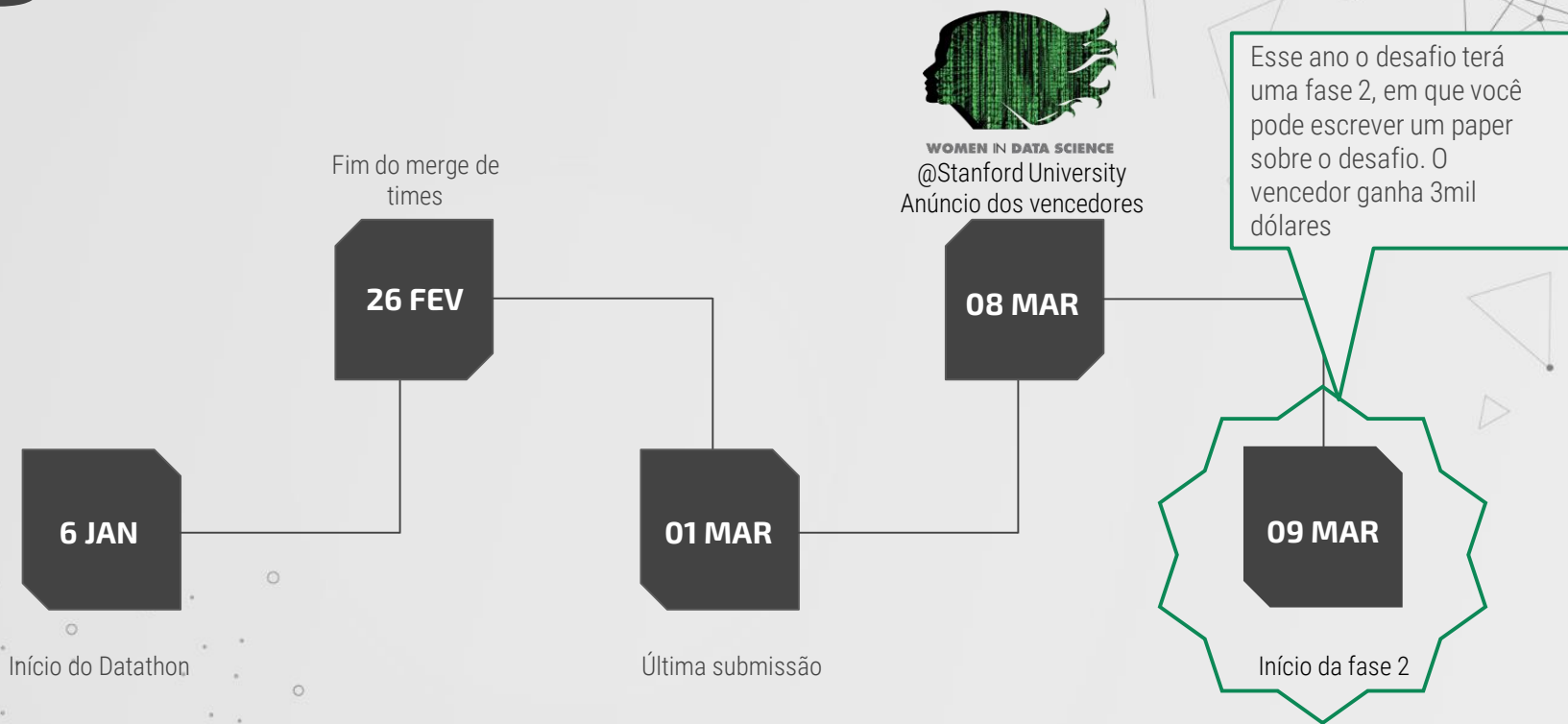
9

Prestem atenção nas datas



9

Prestem atenção nas datas




10

Use o Kaggle para Aprender

Mais do que ganhar a competição o importante é a quantidade de conhecimento que vocês vão compartilhar e aprender!

Abusem disso!





Esse ano, participantes do Datathon devem encontrar padrões e tirar insights de dados para auxiliar médicos a detectar pré-condições existentes, nesse caso a diabetes. Os pacientes com diabetes pré-existente que pegam a Covid-19 são mais propensos a apresentar agravamento da doença e portanto precisam de mais atenção.

> *Os **Datathons do WiDS** sempre focam em tópicos que proporcionem **benefícios sociais significativos**. Estamos otimistas de que Machine Learning pode ajudar a reduzir mortes em UTIs. Novos modelos voltados para essas previsões foram publicados recentemente, nos dando esperanças de que podemos capitalizar esses sucessos recentes*

Karen Matthys, Diretora executiva do ICME de Stanford e Co-diretora do WiDS



**Está sem grupo? Se inscreva no form
que te indicaremos outras pessoas que
estão procurando parceiros:**

bit.ly/datathon-wids_quero-grupo



Muito obrigada! :)

Se tiverem mais dúvidas, podem nos procurar:



wids_sp



wids_sp

Compartilhem e tirem dúvidas também no nosso grupo do Telegram, outras meninas estarão participando e poderão colaborar:

http://bit.ly/wids-sp_comunidade

