



(<https://programs.upgrad.com/data-science-advanced-certificate-bdm-iimk/>)

(https://analyttica.com/innovative-datasience-learning-platform?utm_source=BRAND&utm_medium=AIM&utm_campaign=DV_DA_IIMK_BRAND_AIM_MARCH5-ROADBLOCK_METRO_WEBSITE)



Innovative Platform to learn Data Science
150,000+ Data Enthusiasts Globally

[Subscribe Now](#)

(https://leaps.analyttica.com/innovative-datasience-learning-platform?utm_source=analyticsindiamagazine&utm_medium=partner_website&utm_campaign=aim_learndatascience_partnership)

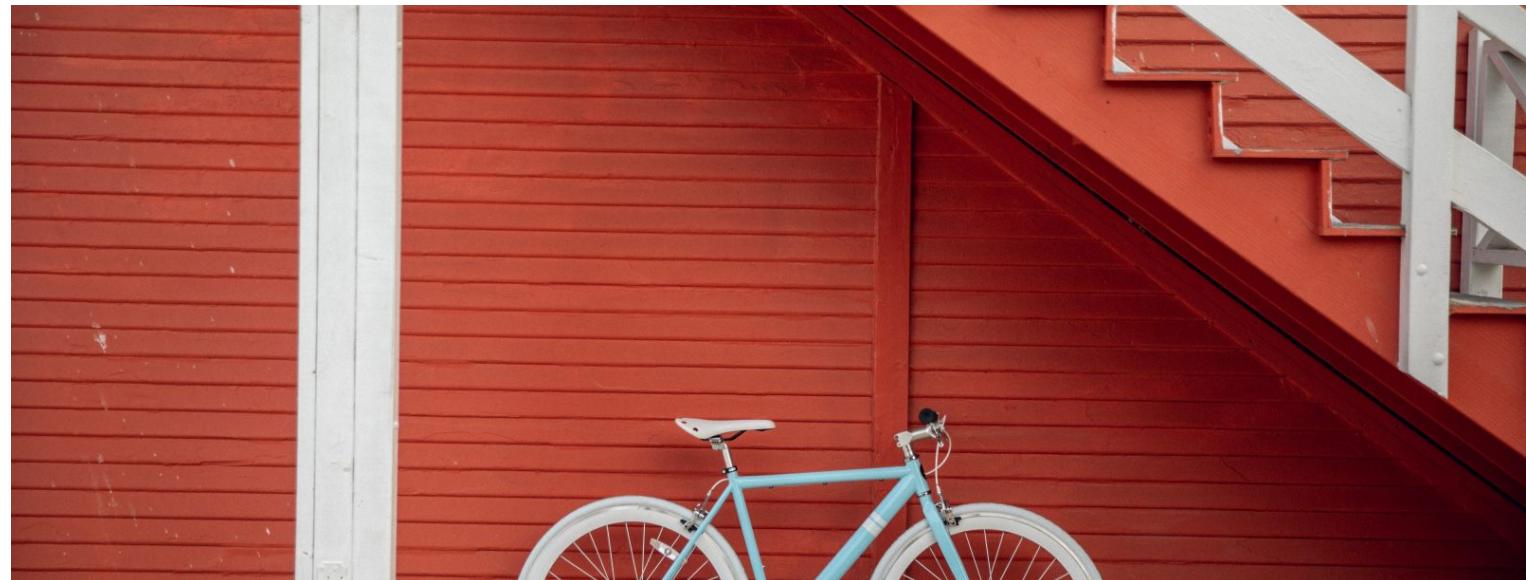
[OPINIONS \(HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/\)](#)

8 Explainable AI Frameworks Driving A New Paradigm For Transparency In AI



BY RAM SAGAR (<https://analyttica.com/author/ram-sagar/>)

18/10/2019



(https://www.qpiae-explorer.tech/certification/?utm_source=aimagazine&utm_medium=banner&utm_campaign=preregistration)

Due to the ambiguity in [Deep Learning solutions](https://analyttica.com/8-platforms-you-can-use-to-build-mobile-deep-learning-solutions/) (<https://analyttica.com/8-platforms-you-can-use-to-build-mobile-deep-learning-solutions/>), there has been a lot of talk about how to make explainability inclusive of an ML pipeline. Explainable AI refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. It contrasts with the concept of the “black box” in machine learning and enables transparency.

For example, consider the following two images:

$f(\mathbf{x}) = 0.293 \tanh(0.337 x_1 - 0.329 x_2 + 0.251 x_3 - 0.288 x_4 - 0.297 x_5 + 0.436 x_6 +$
 + 0.166 x₇ - 0.184 x₈ + 0.219 x₉ + 0.483 x₁₀ - 0.22 [pm](https://analyticsindiamag.com/) (<https://analyticsindiamag.com/>)
 + 0.352 x₁₄ + 0.259 x₁₅ + 0.176 x₁₆ + 0.345 x₁₇ + 0.314 x₁₈ + 0.177 x₁₉ - 0.329 x₂₀ +
 - 0.363 x₂₁ + 0.216 x₂₂ - 0.148 x₂₃ - 0.043 x₂₄ + 0.316 x₂₅ - 0.068 x₂₆ - 0.421 x₂₇₍₀₎ +
 + 0.15 x₂₇₍₁₎ - 0.289 x₂₇₍₂₎ - 0.241 x₂₈ + 0.16 x₂₉ + 0.199 x₃₀ - 0.111 x₃₁ - 0.164 x₃₂ +
 + 0.117 x₃₃ + 0.466 x₃₄ + 0.457 x₃₅ + 0.133 x₃₆ + 0.331 x₃₇ - 0.362 x₃₈ - 0.43 x₃₉ +
 - 0.491 x₄₀ - 0.155 x₄₁ + 0.371 x₄₂ - 0.05 x₄₃ - 0.177 x₄₄ - 0.044 x₄₅ + 0.225 x₄₆ +
 + 0.328 x₄₇ - 0.118 x₄₈ - 0.3) +
 - 1.934 \tanh(-0.233 x₀ + 0.174 x₁ - 0.252 x₂ - 0.501 x₃ - 0.125 x₄ + 0.311 x₅ - 0.573 x₆ +
 - 0.299 x₇ + 1.123 x₈ + 0.318 x₉ - 1.169 x₁₀ + 0.105 x₁₁ - 0.429 x₁₂ - 0.075 x₁₃ +
 - 0.143 x₁₄ + 0.146 x₁₅ - 0.531 x₁₆ + 0.077 x₁₇ - 0.133 x₁₈ - 0.122 x₁₉ + 0.162 x₂₀ +
 - 0.08 x₂₁ - 0.496 x₂₂ - 0.21 x₂₃ - 0.113 x₂₄ + 0.485 x₂₅ + 0.575 x₂₆ - 0.126 x₂₇₍₀₎ +
 + 0.135 x₂₇₍₁₎ + 0.022 x₂₇₍₂₎ - 0.352 x₂₈ - 0.693 x₂₉ + 0.379 x₃₀ + 0.409 x₃₁ - 0.109 x₃₂ +
 + 0.228 x₃₃ + 0.292 x₃₄ + 0.161 x₃₅ - 0.086 x₃₆ - 0.3 x₃₇ - 0.089 x₃₈ + 0.163 x₃₉ +
 - 0.074 x₄₀ + 0.31 x₄₁ - 0.849 x₄₂ + 0.14 x₄₃ + 0.754 x₄₄ + 0.291 x₄₅ - 0.533 x₄₆ + 0.273 x₄₇ +
 - 0.285 x₄₈ - 0.286) + 0.252

(<https://analyticsindiamag.com/wp-content/uploads/2019/10/Equation-705x382.png>)

via Rulex XAI

```

IF (customer_province in {A, B, C, D} AND damage_class in {1} AND Number of days between policy start and date of accident <= 371
THEN Fraud = Yes
IF (customer_province in {E, B, C, F} AND Customer age > 48 AND Number of days between date of accident and complaint > 1)
THEN Fraud = Yes
IF (customer_province in {G, H, I, J, K, L, M, N, B, O, P, Q, R, S})
THEN Fraud = No
IF (Number of days between date of accident and policy end <= 2)
THEN Fraud = No
  
```

(<https://analyticsindiamag.com/wp-content/uploads/2019/10/ruledx.png>)

via Rulex XAI

The first picture consists of a bunch of mathematical expressions chained together that represent the way inner layers of an algorithm or a neural network function. Whereas the second picture also contains the working of an algorithm but the message is more lucid.

Given an opportunity to choose, any client would prefer the second example. The kind of reputation that ML models have gained over the years has made users skeptical. This problem is more prevalent in use cases where the results are of critical nature. These can be an image recognition algorithm identifying criminals or a model deployed for cancer diagnosis or it can be a recommendation model that pushes certain news and products forward.



(<https://www.analytixlabs.co.in/>)

The non-transparency of the algorithms can lead to the exploitation of certain groups.

To promote explainable AI, researchers have been developing tools and techniques and here we look at a few which have shown promising results: over the past couple of years:

What-if Tool

TensorFlow team announced the [What-If Tool](https://pair-code.github.io/what-if-tool/) (<https://pair-code.github.io/what-if-tool/>), an interactive visual interface designed to help visualize datasets and better understand the output of TensorFlow models. To analyze the models deployed. In addition to TensorFlow models, one can also use the What-If Tool for XGBoost and Scikit Learn models.

Once a model has been deployed, its performance can be viewed on a dataset in the What-If tool.

Additionally, one can slice the dataset by features and compare performance across those slices, identifying subsets of data on which the model performs best or worst, which can be very helpful for ML fairness investigations.

Local Interpretable Model-Agnostic Explanations LIME is an actual method developed by researchers at the University Of Washington to gain greater transparency on what's happening inside an algorithm.

When the number of dimensions is high, maintaining local fidelity for such models becomes increasingly hard. In contrast, LIME solves the much more feasible task of finding a model that approximates the original model locally

[LIME \(https://arxiv.org/pdf/1602.04938v1.pdf\)](https://arxiv.org/pdf/1602.04938v1.pdf) incorporates interpretability both in the optimization and the notion of interpretable representation, such that domain and task specific interpretability criteria can be accommodated.

LIME, a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner. The team also introduced SP-LIME, a method to select representative and non-redundant predictions, providing a global view of the model to users

DeepLIFT

DeepLIFT is a method that compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference. It gives separate consideration to positive and negative contributions, [DeepLIFT \(https://github.com/kundajelab/deeplift\)](https://github.com/kundajelab/deeplift) can also reveal dependencies which are missed by other approaches. Scores can be computed efficiently in a single backward pass.

DeepLIFT is on pypi, so it can be installed using pip:

```
pip install deeplift
```

Skater

Skater is a unified framework to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-case. Skater is an open source python library designed to demystify the learned structures of a black box model both globally and locally.

Shapley

The Shapley Value SHAP (SHapley Additive exPlanations) is the average marginal contribution of a feature value over all possible coalitions.

Coalitions are basically combinations of features which are used to estimate the shapley value of a specific feature. It is a unified approach to explain the output of any machine learning model.

SHAP connects game theory with local explanations, uniting several previous methods and representing the only possible consistent and locally accurate additive feature attribution method based on expectations.

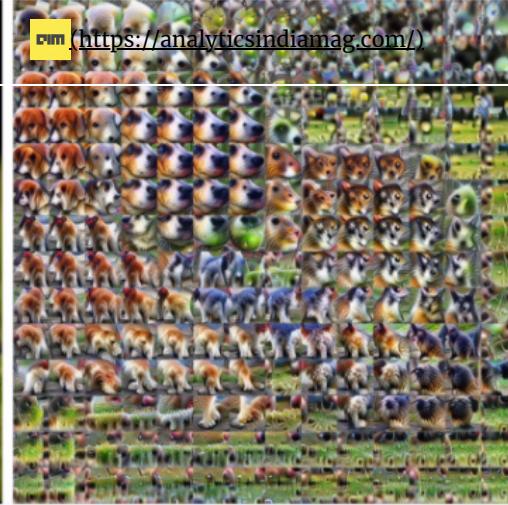
AIX360

The [AI Explainability 360 \(https://github.com/IBM/AIX360\)](https://github.com/IBM/AIX360) toolkit is an open-source library developed by IBM in support of interpretability and explainability of datasets and machine learning models. The AI Explainability 360 is released as a Python package that includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.

Activation Atlases



Input image from ImageNet.



Activation grid from InceptionV1, layer mixed4d.

(<https://analyticsindiamag.com/wp-content/uploads/2019/10/aas.png>)

Google (<https://ai.googleblog.com/2019/03/exploring-neural-networks.html>) in collaboration with OpenAI (<https://blog.openai.com/introducing-activation-atlases/>), came up with Activation Atlases, which was a novel technique aimed at visualising how neural networks interact with each other and how they mature with information along with the depth of layers.

This approach was developed to have a look at the inner workings of convolutional vision networks and derive human-interpretable overview of concepts within the hidden layers of a network.

SEE ALSO



(<https://analyticsindiamag.com/what-is-the-hiring-process-of-data-scientists-at-cisco/>)

OPINIONS ([HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/](https://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/))

What Is The Hiring Process Of Data Scientists At Cisco?
[\(https://analyticsindiamag.com/what-is-the-hiring-process-of-data-scientists-at-cisco/\)](https://analyticsindiamag.com/what-is-the-hiring-process-of-data-scientists-at-cisco/)

Rulex Explainable AI

Rulex (<https://www.rulex.ai/rulex-explainable-ai-xai/>) is a company that creates predictive models in the form of first-order conditional logic rules that can be immediately understood and used by everybody.

Rulex's core machine learning algorithm, the Logic Learning Machine (LLM), works in an entirely different way from conventional AI. The product is designed so that it produces conditional logic rules that predict the best decision choice, in plain language that is immediately clear to process professionals. Rulex rules make every prediction fully self-explanatory.

And unlike decision trees and other algorithms that produce rules, Rulex rules are stateless and overlapping.

The Need For Transparency In Models



(<https://analyticsindiamag.com/wp-content/uploads/2019/10/fouad-ghazizadeh-jRHU8rNXrlA-unplash.jpg>).

Data quality and its accessibility are two main challenges one will come across in the initial stages of building a machine learning pipeline.

But there can be problems associated with the information that is deployed into the model such as:

- an incorrect model gets pushed
- incoming data is corrupted
- incoming data changes and no longer resembles datasets used during training.

Today, developers are even using deep learning to analyze another deep learning network, allowing them to understand the inner working of a model and the kind of influence as it flows to layers above or below and ultimately back down to the underlying data source.

Operating in the dark has made AI less trustworthy as a solution. Providing a solution to any problem is the end of the story to any ML model. However, for the practitioners it is crucial to explain their results in the most intuitive to their clients.

The European Union's General Data Protection Regulation (GDPR), which went into effect in 2018, insists on having high level data protection for consumers and harmonizes data security regulations within the European Union (EU).

Today, the companies have to inform subjects about any personal data collection and processing and obtain their consent before collecting such data as the GDPR threatens the use of traditional machine learning AI technology for automated decisions.

According to article 14 of GDPR, when a company uses automated decision-making tools, it must provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

For a safer, more reliable inclusion of AI, a seamless blend of human and artificial intelligence is needed. Human intervention should also be considered in developing techniques that allow practitioners to easily evaluate the quality of the decision rules in use and reduce false positives.

What Do You Think?



Add a comment...

Facebook Comments Plugin

Subscribe to our Newsletter

Get the latest updates and relevant offers by sharing your email.

ENTER YOUR EMAIL

SUBSCRIBE NOW

Join Our Telegram Group. Be part of an engaging online community. [Join Here](https://t.me/joinchat/NJLxnhZB7GkX3CPvjs9QGQ)



RAM SAGAR ([HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/RAM-SAGAR/](https://analyticsindiamag.com/author/ram-sagar/))

I have a master's degree in Robotics and I write about machine learning advancements.
email:ram.sagar@analyticsindiamag.com

f SHARE

 [TWEET](https://twitter.com/intent/tweet?text=http://8%20Explainable%20AI%20Frameworks%20Driving%20A%20New%20Paradigm%20in-ai/)

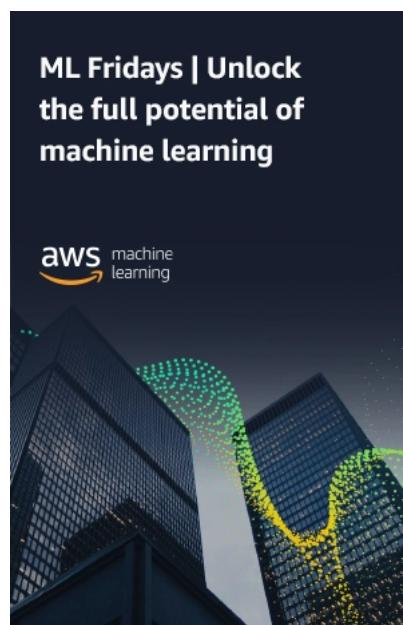
[in](https://www.linkedin.com/cws/share?url=https://analyticsindiamag.com/8-explainable-ai-frameworks-driving-a-new-paradigm-for-data-science/)(<https://www.linkedin.com/cws/share?url=https://analyticsindiamag.com/8-explainable-ai-frameworks-driving-a-new-paradigm-for-data-science/>)

✉(https://wa.me/?text=http://8%20Explainable%20AI%20Frameworks%20Driving%20A%20New%20Paradigm%20For%20Transparer
 (mailto:?)

subject=http://%20Explainable%20AI%20Frameworks%20Driving%20A%20New%20Paradigm%20For%20Transparency%20In%20explainable-ai-frameworks-driving-a-new-paradigm-for-transparency-in-ai/)

1(<https://t.me/share/url?&text=http://%E2%9C%93Explainable%20AI%20Frameworks%20Driving%20A%20New%20Paradigm%20For%20Ti>)

<https://share.flipboard.com/bookmarklet/popout?v=2&title=http://8%20Explainable%20AI%20Frameworks%20Driving%20A%20New%20Age%20of%20Machine%20Learning>





(<https://ad.doubleclick.net/ddm/clk/4.91374.101;29816944.8;q>).



(https://business.louisville.edu/learnmore/UofLMSBA/?utm_campaign=MSBA&utm_source=analyticsindia&utm_medium=display&utm_keyword=analyticsindia&utm_content=GetPaid)

OUR UPCOMING EVENTS

SKILLUP 2021 | Data Science Education Fair | 22–23rd April | [Register here>>](https://skillup.analyticsindiasummit.com/) (<https://skillup.analyticsindiasummit.com/>)

Rising 2021 (<https://rising.analyticsindiasummit.com/>) | Women in AI Conference | May 21 & 22 | Virtual

RELATED POSTS

[OPINIONS \(HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/\)](#)

The Struggle For Data Privacy
[\(https://analyticsindiamag.com/the-struggle-for-data-privacy/\)](https://analyticsindiamag.com/the-struggle-for-data-privacy/)

11/03/2021 · 3 MINS READ



(<https://analyticsindiamag.com/the-struggle-for-data-privacy/>)

[DEVELOPERS CORNER \(HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/DEVELOPERS_CORNER/\)](#)

Hands-on Guide to Interpret Machine Learning with SHAP
[\(https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/\)](https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/)

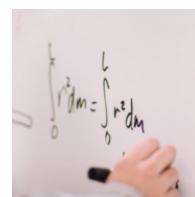
06/03/2021 · 6 MINS READ



(<https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/>)

[DEVELOPERS CORNER \(HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/DEVELOPERS_CORNER/\)](#)

Guide To AI Explainability 360: An Open Source Toolkit By IBM
[\(https://analyticsindiamag.com/guide-to-ai-explainability-360-an-open-source-toolkit-by-ibm/\)](https://analyticsindiamag.com/guide-to-ai-explainability-360-an-open-source-toolkit-by-ibm/)



(<https://analyticsindiamag.com/guide-to-ai-explainability-360-an-open-source-toolkit-by-ibm/>)

[to-ai-explainability-360-an-open-source-toolkit-by-ibm/](https://analyticsindiamag.com/explainability-360-an-open-source-toolkit-by-ibm/)

29/01/2021 · 16 MINS READ

[explainability-360-an-open-source-toolkit-by-ibm/](https://analyticsindiamag.com/explainability-360-an-open-source-toolkit-by-ibm/)

DEVELOPERS CORNER (<https://analyticsindiamag.com/category/developers-corner/>)



Hands-On Guide To Adversarial Robustness Toolbox (ART): Protect Your Neural Networks Against Hacking

(<https://analyticsindiamag.com/adversarial-robustness-toolbox-art/>)

07/01/2021 · 7 MINS READ

(<https://analyticsindiamag.com/robustness-toolbox-art/>)

PEOPLE (<https://analyticsindiamag.com/category/interviews/>)



Model Explainability Validates AI For Safety-Critical Systems, Says Prashant Rao, MathWorks

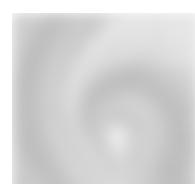
(<https://analyticsindiamag.com/explainability-validates-ai-for-safety-critical-systems-says-prashant-rao-mathworks-india/>)

India

(<https://analyticsindiamag.com/model-explainability-validates-ai-for-safety-critical-systems-says-prashant-rao-mathworks-india/>)

29/12/2020 · 5 MINS READ

OPINIONS (<https://analyticsindiamag.com/category/articles/>)



A Greater Foundation For The Triumph Of Deep Learning With XAI

(<https://analyticsindiamag.com/a-greater-foundation-for-the-triumph-of-deep-learning-with-xai/>)

(<https://analyticsindiamag.com/a-greater-foundation-for-the-triumph-of-deep-learning-with-xai/>)

21/12/2020 · 5 MINS READ

CONNECT WITH US	OUR BRANDS	OUR CONFERENCES	OUR VIDEOS	BRAND PAGES	LISTS
About Us (https://analyticsindiamag.com/)	MachineHack – ML Hackathons (https://www.machinehack.com/)	Cypher (https://www.analyticsindia.com/cypher/)	Documentary – The Transition Cost (https://www.youtube.com/on-intel/)	Intel AI Hub (https://analyticsindiamag.com/intel-ai-hub/)	Academic Rankings Best Firms To Work For (https://analyticsindiamag.com/best-firms-to-work-for/)
Advertise with-us/ (https://analyticsindiamag.com/)	AIM Research (https://aimresearch.ai/)	The MachineCon (https://themachinecon.com/)	Web Series – The Web Series (https://www.youtube.com/watch?v=7pvGjbzTTWk&list=PL9Kc1zSa4_6OzMfxoI1SJZOGpNp371vbd)	ASSOCIATION OF DATA SCIENTISTS (https://www.youtube.com/watch?v=WQbKbLRKOsk&list=PL9Kc1zSa4_6OxzJqQEJa-qI55CtLZxFl2v)	For best-firms-in-india-data-scientists-to-work-for-2021/ (https://analyticsindiamag.com/best-firms-in-india-for-data-scientists-to-work-for-2021/)
Weekly Newsletter (https://recruits.analyticsindia.com/)	AIM Recruits (https://recruits.analyticsindia.com/)	Machine Learning Developers Summit (http://mlds.analyticsindia.com/)	Dating Scientists (https://www.youtube.com/watch?v=gvZfaeVvbGE&list=PL9Kc1zSa4_6OwKuqj8W6vz-V5DucIU4Y)	Chartered Data Scientist(TM) (https://www.adasci.org/cds-most-influential-data-scientists/)	Top Leaders (https://analyticsindiamag.com/most-influential-data-scientists/)
Write for us (https://analyticsindiamag.com/write-for-us/)	AWARDS (https://analyticsindiamag.com/awards/)	The Rising plugin (https://rising.analyticsindia.com/plugin/)	Podcasts – Simulated Reality (https://www.youtube.com/watch?v=gvZfaeVvbGE&list=PL9Kc1zSa4_6OwKuqj8W6vz-V5DucIU4Y)	Machine (https://www.adasci.org/latent-top-10-data-scientists/)	Data Scientists (https://analyticsindiamag.com/latent-top-10-data-scientists/)
Careers (https://www.linkedin.com/company/analytics-india-magazine/jobs/)	Analytics100 (https://plugin.analyticsindia.com/40-under-40-data-scientists/)	EVENTS	Analytics India Guru (https://www.youtube.com/watch?v=oQDZMdeyzgw&list=PL9Kc1zSa4_6OwKuqj8W6vz-V5DucIU4Y)	Continuous Learning (https://www.adasci.org/continuous-learning-startups/)	Scientists (https://analyticsindiamag.com/continuous-learning-startups/)
Contact Us (https://mlds.analyticsindia.com/awards/)	Data Science Excellence (https://analyticsindiaevents.com/aim-excellence/)	AIM Custom Events (https://analyticsindiaevents.com/aim-custom-award-events/)	AIM Custom Events (https://www.youtube.com/watch?v=Q7diUR_PRGg&list=PL9Kc1zSa4_6OwKuqj8W6vz-V5DucIU4Y)	Career Center (https://www.adasci.org/career-trends/)	Emerging + Analytic (https://analyticsindiamag.com/career-trends/)
MENTORSHIP	Women in AI Leadership (https://rising.analyticsindia.com/mentorship-circle/)	AIM Virtual (https://analyticsindiaevents.com/aim-virtual/)	Deeper Insights with Deep Dive (https://www.youtube.com/watch?v=Z3Z7Riw1G9o&list=PL9Kc1zSa4_6Oyv8tAFzC22cuLXNRUZvXAG)	Membership (https://www.adasci.org/members-benefits/)	Trends & Category (https://analyticsindiamag.com/members-benefits/)
Assisted Mentoring (https://analyticsindiamag.com/mentorship-circle/assisted-mentoring/)			Curiosum – AI Storytelling (https://www.youtube.com/watch?v=AfsqH5EzjIg&list=PL9Kc1zSa4_6OzOCjo2YUNym6thlDDoFozc)		

[ABOUT US\(HTTPS://ANALYTICSINDIAMAG.COM/ABOUT/\)](#)

[ADVERTISE\(HTTPS://ANALYTICSINDIAMAG.COM/ADVERTISE-WITH-US/\)](#)

[WRITE FOR US\(HTTPS://ANALYTICSINDIAMAG.COM/WRITE-FOR-US/\)](#)

[COPYRIGHT\(HTTPS://ANALYTICSINDIAMAG.COM/COPYRIGHT-TRADEMARKS/\)](#)  (<https://analyticsindiamag.com/>)

[PRIVACY\(HTTPS://ANALYTICSINDIAMAG.COM/PRIVACY-POLICY/\)](#)

[TERMS OF USE\(HTTPS://ANALYTICSINDIAMAG.COM/TERMS-USE/\)](#)

[CONTACT US\(HTTPS://ANALYTICSINDIAMAG.COM/CONTACT-US/\)](#)

 (<https://facebook.com/analyticsindiamagazine>)

 (<https://twitter.com/analyticsindia>)

 (<https://instagram.com/analyticsindiamagazine>)

 (<https://pinterest.com/analyticsindia>)

 (<https://youtube.com/channel/UCAlwrsgeAVgvwQsfouma>)

 (<https://medium.com/analytics-india-magazine>)

 (<https://www.linkedin.com/company/analytics-india-magazine>)

 (<https://t.me/nlxnhzb7gkx3cpvis9qgo>)