CAS CUSTOM SERVICES℠

IMPACT OF DATA QUALITY ON MACHINE LEARNING RESULTS

Solution success story

The challenge: Prediction accuracy for machine learning

While tangible wins are proving valuable for AI in R&D, data continues to be a key stumbling block with AI for many firms. Recently, a paper published in the Proceedings of the National Academy of Science classified almost 10,000 chemical compounds using Morgan fingerprints (an established and commonly used descriptor) to represent chemical compounds.

A machine learning algorithm was required to help classify the compounds as either active or inactive against a series of five different targets based off their predicted biological activities. Once complete, prediction accuracy was only 10% across targets.

CAS
A division of the
American Chemical Society

The solution: Quality data dramatically improved success

CAS hypothesized: does the quality of chemical descriptions utilized by a given algorithm improve prediction accuracy?

After mapping nearly 10,000 chemical compounds to CAS proprietary fingerprint descriptors, CAS leveraged the higher quality descriptors for improved prediction results utilizing an SVM algorithm.

By improving the quality of the data, CAS increased prediction accuracy from 10% to 45% across targets with a median of 33% and a mean of 31%. This improvement in prediction accuracy meant that teams could explore vast landscapes, minimize false positives, and focus their efforts on top compounds in hours instead of weeks.

Data scientists often focus on the complexity and nuances of algorithms because they can be truly unique. However, performance can also be significantly improved just by utilizing higher quality data to train algorithms.

"More data beats clever algorithms, but better data beats more data"

Peter Norvig,
PHD Director of research at Alphabet,
notable author and expert of AI

Quality data improves AI prediction accuracy by 31%

CAS Comparative Study

Find out how CAS Custom Services can help you transform scientific data into actionable, evidence-based insights that maximize investment and fuel success at **cas.org.**

CAS
A division of the
American Chemical Society

SVCGENENGBRO101267230721