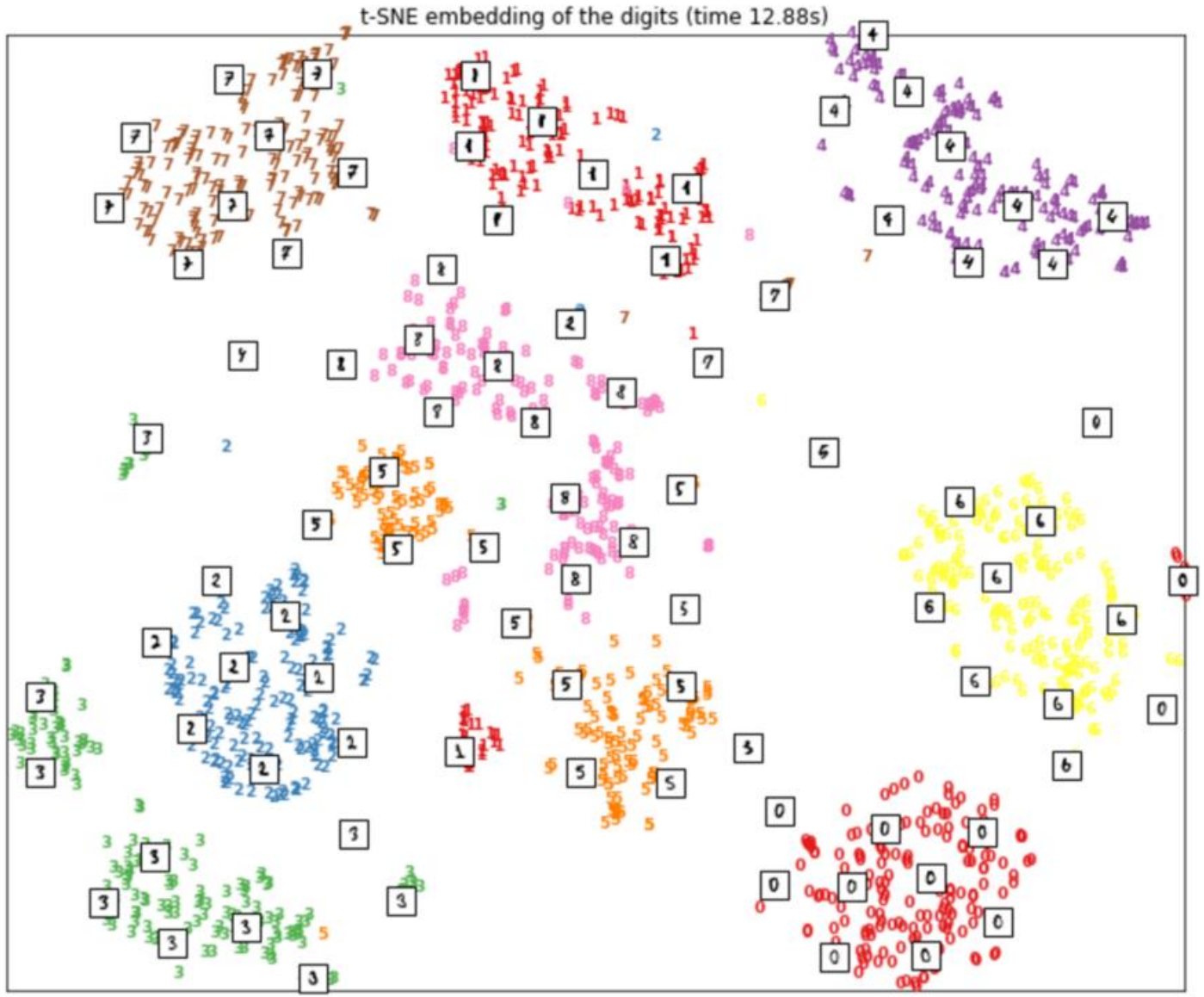
Exploratory Analysis and Visualization

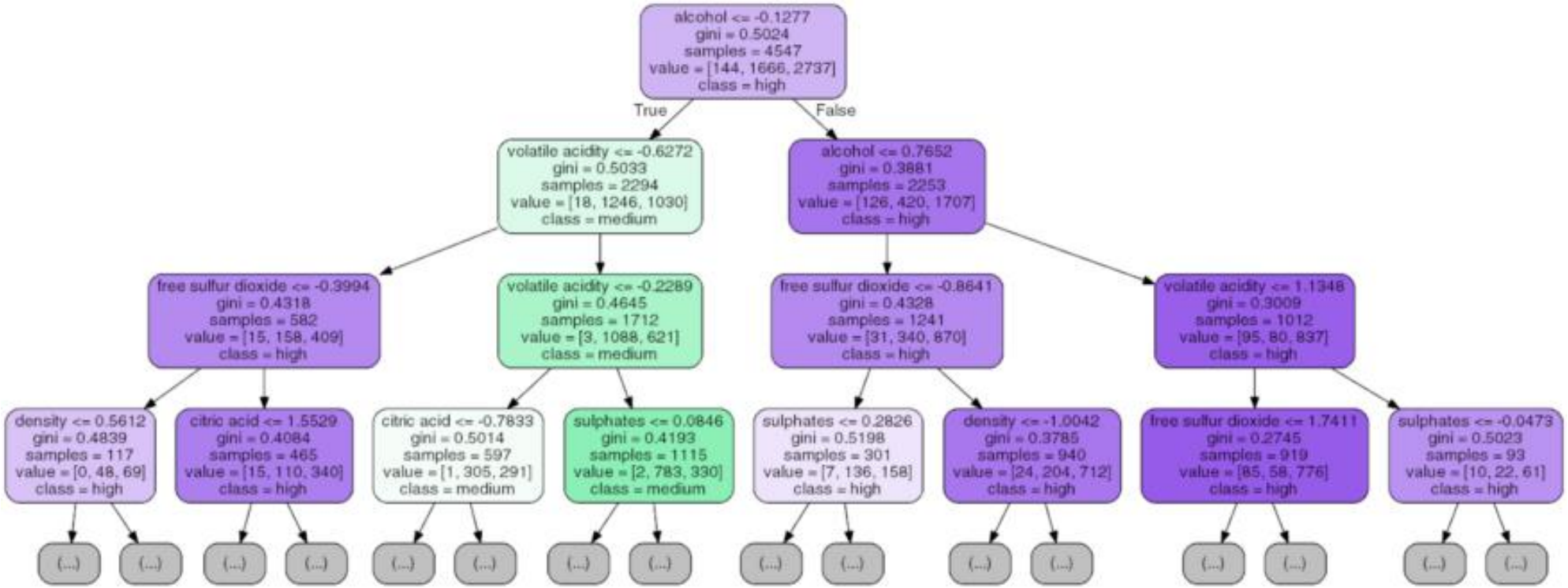
The idea of exploratory analysis is not something entirely new. For years now, data visualization has been one of the most effective tools for getting latent insights from data. Some of these techniques can help us in identifying key features and meaningful representations from our data which can give an indication of what might be influential for a model to take decisions in a human-interpretable form.

Dimensionality reduction techniques are very useful here since often we deal with a very large feature space (curse of dimensionality) and reducing the feature space helps us visualize and see what might be influencing a model to take specific decisions. Some of these techniques are as follows.

- **Dimensionality reduction:** Principal Component Analysis (PCA), Self-organizing maps (SOM), Latent Semantic Indexing
- **Manifold Learning:** t-Distributed Stochastic Neighbor Embedding (t-SNE)
- **Variational autoencoders:** An automated generative approach using variational autoencoders (VAE)
- **Clustering:** Hierarchical Clustering


blasting
children
congregation
peter
afflictions
faith
aaron
flowing
ham
famine
japheth
pestilence
shem
noah
moses
godly
egyptians
bondage
pharaoh
egypt
apostles
ordinance
lord
persecutions
morever
glory
promised
rod
jesus
james
immediately
john
god
christ
peril
baptist
peter
gospel
zebedee
messias
mildew
noe
neverthless
henoch
doctor
galilee
worldly
reasonable


t-SNE embedding of the digits (time 12.88s)



However, like we discussed we may not get these rules for other models which are not as interpretable as tree based models. Also huge decision trees always become very difficult to visualize and interpret.

Conclusion

This article should help you take more definitive steps on the road towards Explanable AI (XAI). You now know the need and importance of model interpretation. The issues with bias and fairness from the first article. Here we have taken a look at traditional techniques for model interpretation, discussed their challenges and limitations and also covered the classic trade-off between model interpretability and prediction performance. Finally, we looked at the current state-of-the-art model interpretation techniques and strategies including feature importances, PDPs, global surrogates, local surrogates and LIME, shapley values and SHAP. Like I have mentioned before, Let's try and work towards human-interpretable machine learning and XAI to demystify machine learning for everyone and help increase the trust in model decisions.