# Explainable Artificial Intelligence as a Generic Data Analysis Tool

*Keywords: Explainable AI, Data science, Regression analysis, Clustering, Novel approach*

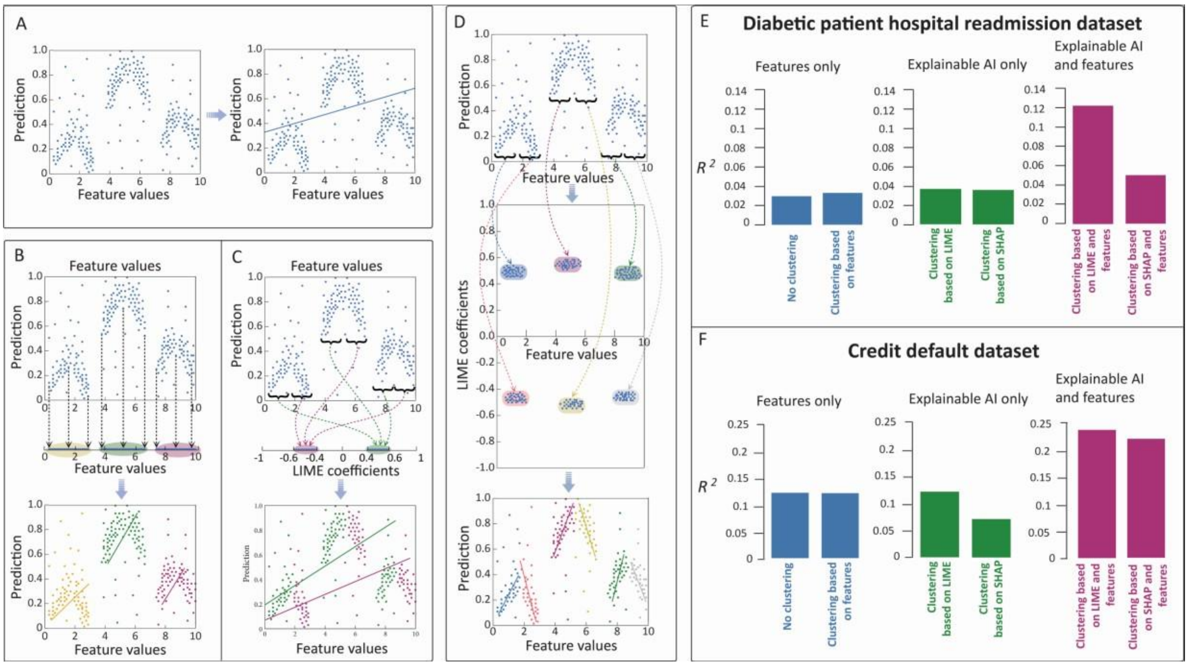
## Extended Abstract

Artificial Intelligence (AI) in general, and Machine Learning (ML) in particular, have garnered significant attention in recent years due to the growing number of applications. ML models are now being used in various aspects of life, including judiciary applications, health-related applications, and financial applications. Unfortunately, many of these models are hard to interpret by humans, which is problematic in high-stake applications that affect people's well-being, e.g., when the outcome of a model influences the decision of whether a customer should receive a loan, a patient should have surgery, or a suspect should be sent to prison. In all these applications, the person making the decision (i.e., the banker, the doctor, or the judge), as well as the person affected by this decision (i.e., the customer, the patient, or the suspect) would benefit from a human-interpretable explanation of the ML and/or its outcome. These concerns have inspired the rapidly growing field of Explainable AI (or XAI for short).

Motivated by the wealth of innovative XAI techniques that were proposed in recent years, our goal is to explore the possibility of using these techniques outside of their originally-intended scope. In particular, we are interested in investigating whether XAI techniques can provide insights into the data set itself rather than explaining ML models and their outcomes. If this is indeed possible, then the growing literature on XAI can support existing data analysis methods. Arguably, one of the most widely-applied such methods is regression analysis, which is typically used to estimate the relationship between the dependent variable and one or more independent features.

There are few XAI tools suitable for any classifier and capable of handling tabular data. Among these, Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and Shapley Additive Explanations (SHAP) (Lundberg, et al., 2020) are the most widely used ones and are also used in this research. Notably, the way in which regression analysis works is fundamentally different from the way in which LIME and SHAP work. The former provides a global view of the data, whereas the latter two provide a local view. More specifically, LIME and SHAP provide insights on a specific input instance, whereas regression estimates a high-level relationship, taking into consideration the entire data set. Thus, our key observation is that the high-level view (which is often the primary concern of data analysts) may benefit from additional details obtained by zooming into the different instances, i.e., by applying LIME and SHAP not on a single instance of interest as originally intended by the developers of these techniques, but rather on all instances in the data set. In this research, we explore possible ways to achieve this.

Having explained how LIME and SHAP work, let us now present a way these techniques can be combined with regression analysis to support data analysts interested in understanding the relationship between a given feature and the outcome variable. Starting with LIME, we will illustrate our basic idea using a stylized example corresponding to a hypothetical data set; see the left panel of Figure 1. A. To understand the relationship between the feature and the output variable, an analyst may decide to run a regression, thereby fitting a straight line to the data; a hypothetical such line is depicted in the right panel of Figure 1. A. However, the goodness of fit of such a model will probably be poor since the relationship at hand is clearly not linear. Instead, given that the data points are clustered, the analyst may first decide to run a clustering algorithm and then run a regression in each cluster. Suppose that the resulting fitted lines turn out to be as illustrated in Figure 1. B, where every color represents a separate cluster (note that the data points are clustered based solely on the x-axis values, since this axis represents the only feature in our data set). Again, the resulting goodness of fit will probably be poor since the relationship in each cluster is also not linear. Clustering the data based on the LIME/SHAP coefficients (Figure 1. C) implies that the points around which the slope is positive are grouped together (highlighted in green), and those around which the slope is negative are grouped together (highlighted in pink). Still, fitting a straight line in each cluster fails to capture the overall pattern. We propose clustering the data based on the feature values, and LIME/SHAP coefficients, which results in six clusters, and fitting a straight line in each cluster captures the overall pattern adequately. As can be seen, if the data analyst runs a separate regression in each of these clusters, the goodness of fit is likely to be greater than that obtained by any of the previous methods. We demonstrate in the experiments that clustering the data points this way significantly improves the average R$^2$ of regression run on all the clusters (Figure 1. E and 1. F demonstrate the obtained R$^2$ on two different data sets). The intuition behind this is that instead of relying only on features or only on instance level feature rankings, we are using more information in terms of the actual instance and its slope at the local level, therefore making the clustering more informed.



## References

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., . . . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 56-67.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Association for Computing Machinery*, 1135-1144.