**Q1: Essay**

**Read “DefiningInsight for Visual Analytics”by Changet al.**

**1.1 Summarize what are insights of visual analytics defined in this paper?**

Summary :

In this paper, the researchers addressed the concerns of defining insight in the standpoint of visual analytics. They have presented the definition of insight for visual analytics as an amalgamation of two types of insight that are listed below.

1. Spontaneous insight: This concept is borrowed from cognitive neuroscience that explains insight as a spontaneous event that takes a solver from not knowing to solve a problem to knowing how to solve it. This occurs through a unique neural activity that gives the solver a sensation of the "aha" moment.

2. Knowledge-building insight: This insight is based on the knowledge-building process, where the solver builds an insight by gathering small pieces of information that builds upon their existing knowledge.

For spontaneous insight, the amount of pre-existing domain knowledge acts as a catalyst that facilitates the neural activity to unexpected rewiring of semantic knowledge to form an insight or even a deeper insight. Since spontaneous insight requires the presence of knowledge about the problem, the authors denote that the two insights are intertwined.

However, the authors ascribe the two insights are in fact distinct, as the trajectories in which each insight is formed are different. This phenomenon was explained more clearly by discussing the brain activity of participants while solving a problem. It was seen that parts of the brain responsible for working memory and long-term memory were active during problem solving. But, when a spontaneous insight occured, a sharp burst of activity was seen in the right part of the long-term memory. Moreover, spontaneous insight is used for more complex problems or incomprehensible tasks, whereas knowledge-building insight is a learning process that is used for regular problem-solving tasks.

Finally, the authors suggest spontaneous and knowledge-building insights to be considered for defining insight for visual analytics. As these two terms together represent a moment of enlightenment and provide a broader term to mean an advance in the knowledge.

**1.2 This article was published in 2009. Astherapid rising of data mining and machine learning techniques, are there any important changes on the insights of visual analytics?Make an argument for this. If there are, describe the changes. If there are not, discuss your reasons. Use examples from class or elsewhere to support your argument.**

In 2005, Saraiya et al.[2] define insight “as an individual observation about the data by the

participant, a unit of discovery”. North [3] points out insights are complex, deep, unexpected, qualitative, and relevant. Yi et al.[4] posit that insights are not just the end results but may open doors for further exploration processes. Chang et al.[1] gave two definitions for insight which is elaborated in the previous answer. Sacha et al.[5] present a knowledge discovery model where they use the definition of insight provided by Chang et al.[1]. The proposed model is divided into four stages. First stage is called a computer which consists of three components: data (in structured, semi-, or unstructured manner), model (data mining algorithm), and Visualization. Then in the second stage, an exploration loop is present with two components namely, “action” and “finding”. The “action” component involves the user to interact with the data either by changing model parameters or interacting with the visualization. During this, the user typically finds some pattern or unusual behaviors such as missing data, this is referred to as the “finding” stage. The third stage is the verification loop where the user forms a hypothesis and tries to disprove it by finding evidence. This process often leads to knowledge discovery and to form insights. Here the insights definition is an extended version of [1], and the process of forming insight is given as “user interprets a finding, often with previous domain knowledge, to generate a unit of information”. Insights are formed when a hypothesis is backed with strong evidence, where the isight could be small or something significant.

In 2018, the authors of Zgraggen et al.[8] define insight as “an observation, hypothesis or generalization that could be directly extracted from the data and that did not require prior knowledge or domain expertise”. This contradicts the definition of Chang et al. which says domain knowledge or expertise is essential in forming insights. On the other hand, in 2020, Law et al.[11] gave seven characteristics to define a visual insight. In terms of game development, an insight could be finding a design constraint through a feedback loop from a rule system [10].

In the machine learning and visual analytics paradigm, an insight is formed when the user optimizes the algorithm by leveraging visual user interfaces (VUI). In most cases, insight is formed in a feedback loop where the user makes an assumption, collects evidence, forms insights that helps in reducing the hypothesis or make a unique finding. In Bögl et al.[5], the authors denote that an insight is gained when a user interprets fitness of underlying model, optimal parameters for the model, and refines the set of existing hypotheses. Brooks et al.[7] introduce “FeatureInsight” an interactive visual analytics tool that provides visual summaries for feature ideation. However, the definition of insight is not clear in [9]. The clarity for the definition of insight is a persisting problem, as discussed by the authors of Kandogan et al.[9], that “more developments have to be made in terms of formalisms on vocabulary, language, and an algebra of insight”.

From the above observations, we can see that the definition of insight is molded based on the problem statement and does not have a singular meaning. The examples provided are not exhaustive and give an overview over some domains. We also see that the purpose of insight in machine learning/ neural networks could be just finding the optimal parameters for increasing accuracy or adjusting the parameters of the model to find something unknown.

References for 1.2:

[1] Chang, R., Ziemkiewicz, C., Green, T. M., & Ribarsky, W. (2009). “Defining insight for visual analytics”. IEEE Computer Graphics and Applications, 29(2), 14-17.

[2] Saraiya, P., North, C., & Duca, K. (2005). “An insight-based methodology for evaluating bioinformatics visualizations”. IEEE Transactions on Visualization and Computer Graphics, 11(4), 443-456.

[3] North, C. (2006). Toward Measuring Visualization Insight. IEEE Computer Graphics and Applications, 26(3), 6-9.

[4] Yi, J. S., Kang, Y. A., Stasko, J. T., & Jacko, J. A. (2008). “Understanding and characterizing insights: How do people gain insights using information visualization?”. In Proceedings of the 2008 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (pp. 1-6).

[5] Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., & Keim, D. A. (2014). “Knowledge Generation Model for Visual Analytics”. IEEE Transactions on Visualization and Computer Graphics, 20(12), 1604-1613.

[6] Bögl, M., Aigner, W., Filzmoser, P., Lammarsch, T., Miksch, S., & Rind, A. (2013). Visual analytics for model selection in time series analysis. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2237-2246.

[7] Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., & Simard, P. (2015, October). FeatureInsight: Visual support for error-driven feature ideation in text classification. In 2015 IEEE Conference on Visual Analytics Science and Technology (VAST) (pp. 105-112). IEEE.

[8] Zgraggen, E., Zhao, Z., Zeleznik, R., & Kraska, T. (2018, April). Investigating the effect of the multiple comparisons problem in visual analysis. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

[9] Kandogan, E., & Engelke, U. (2018, June). Towards a Unified Representation of Insight in Human-in-the-Loop Analytics: A User Study. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics (pp. 1-7).

[10] Smith, A., Nelson, M., & Mateas, M. (2009, October). Prototyping games with biped. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 4, No. 1).

[11] Law, P. M., Endert, A., & Stasko, J. (2020). What are Data Insights to Professional Visualization Users?. arXiv preprint arXiv:2008.13057.

**Q2: Development and Analysis**

**a. Which columns do you choose for performing the clustering?**

* Displacement, Cylinders, MPG, Acceleration

**b. What is your rule for deciding whether a car belongs to a cluster or not?**

* A dimension reduction algorithm, t-distributed stochastic neighbor embedding (t-SNE) algorithm was applied to condense the four columns into two components. The t-SNE algorithm first maps a probability distribution function over pairs of high-dimensional objects such that similar objects are assigned a higher probability. Then it defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map. (source: https://en.wikipedia.org/wiki/T-distributed\_stochastic\_neighbor\_embedding)
* After the condensed data is given as input to k-means clustering algorithm which assigns cluster labels to each data point. K-means algorithm partitions data into the predefined number of clusters by assigning each data point to the cluster, whose centroid has similar mean. (source: https://en.wikipedia.org/wiki/K-means\_clustering)
* After obtaining the cluster labels, the data points are plotted on a scatter plot where the two axes represent each of the condensed columns and the color represents the cluster that each data point belongs to.
* Since both algorithms, t-SNE and k-means use statistical transformations, it is hard to interpret the underlying rules. Therefore a decision tree classifier was trained by giving the input as the four selected columns data and the labels as the k-means cluster labels. By specifying the depth of the decision tree as three, an explainable model was created that can approximate the rules in assigning cluster labels. (show in Figure 1)
* The final plot is shown in Figure 2.

**c. How many cars are there in the largest cluster?**

* 205

**d. Do you find any car(s) falsely clustered? For example, it is more reasonable that car A should be in cluster 2, but it is displayed in cluster 1. If so, you need to explain why this happens.**

* There are four points close to cluster 1 that belong to cluster 2. Upon looking at their specifications, it looks like they have similar MPG as cluster 3 and similar cylinders as cluster 1. But the overall specifications look similar to cluster 3, therefore we can say they are not falsely clustered but have some resemblances to cluster 1 as well.

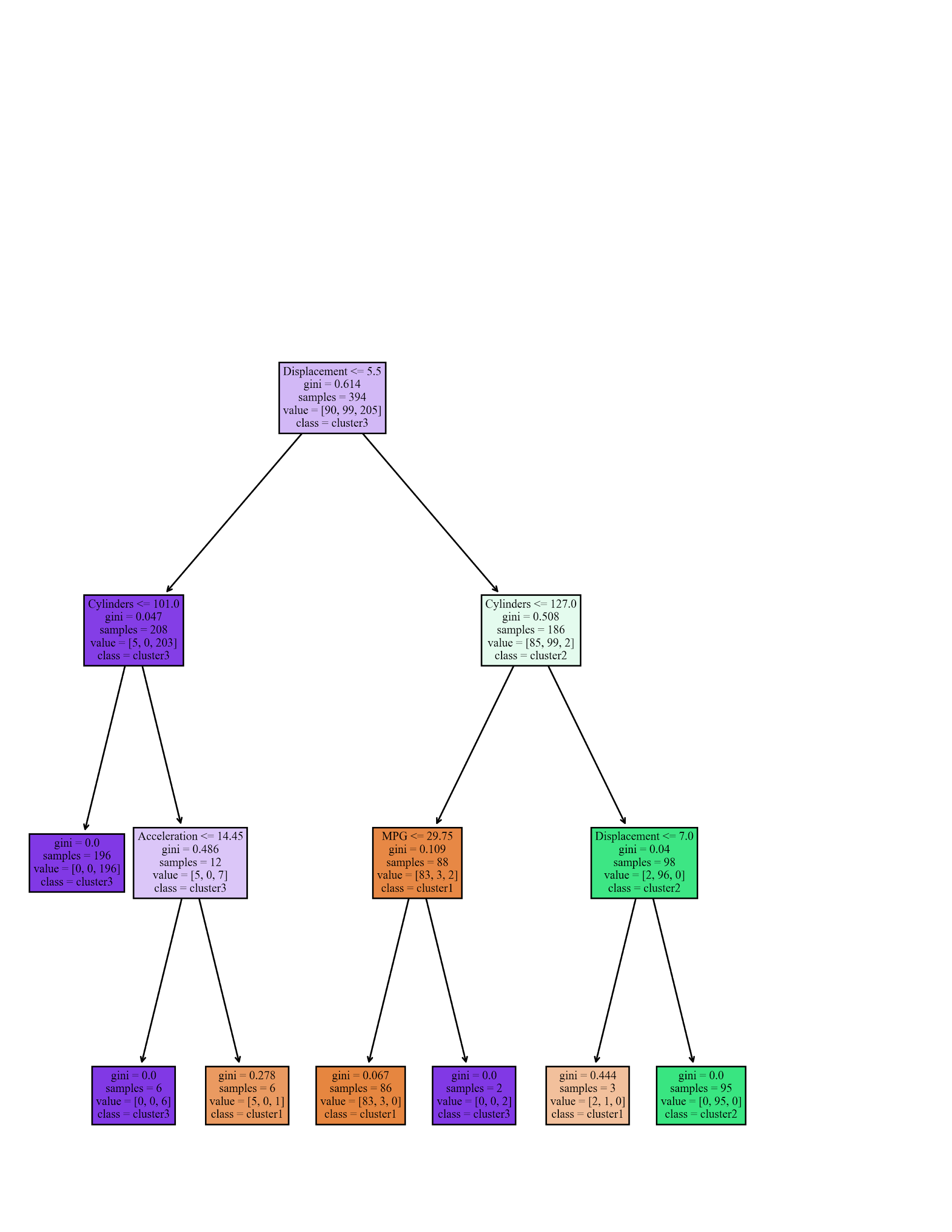
Figure1: Creating an explainable model using a decision tree classifier to interpret the rules behind assigning clusters by the combination of t-SNE and k-means algorithms.

Figure 2 : Final plot

