**Peer Review Report: Application of Machine Learning K-Means Clustering and Linear Regression in Determining the Risk Level of Pulmonary Tuberculosis**

**Authors:** Ziaul Islam Bablu, Sowmik Barua, Mowmita Tajnin Jiba, Towhidul Haque Limon, Piyal Dey, Abhijit Pathak

**Overall Impression:**

This paper explores the use of K-means clustering and linear regression to analyze pulmonary tuberculosis (TB) risk in Bangladesh. While the topic is important and the chosen methods have potential, the paper suffers from several weaknesses:

* **Limited clarity and depth in methodology:** The explanations of both K-means and linear regression are superficial, lacking sufficient detail for reproducibility.
* **Questionable data and analysis:** The dataset used is small and lacks information about its origin and validity. The choice of variables (population, area, TB cases) is simplistic and ignores crucial factors influencing TB risk.
* **Weak interpretation and overstated conclusions:** The interpretation of results is often vague and lacks nuanced discussion of limitations. The conclusions drawn are too strong given the limited analysis.

**Major Comments:**

1. **Methodology:**
   * **K-means:**
     + The description of how the number of clusters (k=3) was determined is missing. Explain the rationale and any methods used (e.g., elbow method, silhouette analysis).
     + Provide more details about the implementation: distance metric used, initialization method for centroids, number of iterations.
   * **Linear Regression:**
     + The choice of population density as the sole predictor for TB cases is overly simplistic. Include other relevant socio-economic and health-related factors to improve the model's explanatory power.
     + Discuss potential issues like multicollinearity if adding more variables.
2. **Data:**
   * The dataset used needs more context: source, time period, reliability, and any limitations.
   * The limited number of districts (17) raises concerns about the representativeness and generalizability of the findings.
   * Justify the choice of variables and acknowledge their limitations in capturing the complexity of TB risk. Consider factors like poverty, access to healthcare, malnutrition, HIV prevalence, etc.
3. **Results and Discussion:**
   * Instead of simply listing cluster assignments, provide a more insightful analysis of the characteristics of each cluster. What distinguishes high-risk areas from low-risk ones?
   * The interpretation of R-squared (0.5740) needs to be more nuanced. While it indicates some explanatory power, it also means that 43% of the variation in TB cases is unexplained by population density.
   * Discuss the limitations of the study in detail, including the small sample size, simplified model, and potential biases in the data.
4. **Writing and Presentation:**
   * The writing needs improvement in terms of clarity, grammar, and flow.
   * The introduction lacks a strong research question and clear articulation of the study's significance.
   * The "Related Works" section is not well-integrated and lacks a clear connection to the current study.
   * The conclusion should summarize the key findings and limitations concisely and avoid overstating the conclusions.

**Minor Comments:**

* **Figure 1:** Improve the figure's quality and clarity. Use a more descriptive title and label the components clearly.
* **Table 8:** Provide a title and caption explaining the table's content.
* **Figure 3:** Improve the visual presentation of the cluster graph (e.g., use distinct colors or markers for different clusters).
* **Figure 4:** Label the axes appropriately (e.g., "Population Density" and "Number of TB Cases").

**Recommendations for Revision:**

1. **Substantially revise the methodology section** to provide more details and address the points raised above.
2. **Strengthen the data section** by providing more context about the dataset and addressing its limitations. Consider acquiring a more comprehensive dataset with additional relevant variables.
3. **Re-analyze the data** using a more robust and nuanced approach, incorporating additional relevant factors influencing TB risk.
4. **Revise the discussion and conclusion** to acknowledge the study's limitations and avoid overstating the findings.
5. **Improve the overall clarity and presentation** of the paper by addressing the writing and formatting issues highlighted above.

**Overall, while the paper addresses an important topic, it requires significant revisions to meet the standards of a publishable manuscript. The authors need to address the methodological limitations, strengthen the data analysis, and present their findings and conclusions more cautiously.**

**Additional Suggestions for Figures:**

Here are some more specific suggestions for improving the figures in your paper:

**Figure 1:**

* **Replace with a flowchart:** Instead of a basic diagram, consider using a flowchart to illustrate the research methodology. This will help to clearly depict the sequence of steps and the flow of information.
* **Use standard flowchart symbols:** Employ standard flowchart symbols (e.g., rectangles for processes, diamonds for decisions) for better readability and comprehension.
* **Add a brief explanation below the figure:** Provide a concise caption explaining the figure's purpose and key elements.

**Figure 3:**

* **Improve visual clarity:** Use distinct colors or markers to differentiate the three clusters clearly.
* **Label clusters directly:** Instead of relying solely on a legend, label each cluster directly on the graph (e.g., "Cluster 1", "Cluster 2", "Cluster 3").
* **Choose appropriate axis labels:** Ensure that the axes are clearly labeled with the specific variables being plotted and their units of measurement. Consider adding a title that summarizes the information conveyed by the graph.

**Figure 4:**

* **Add a trend line:** Overlay the scatter plot with the linear regression trend line to visually represent the relationship between population density and TB cases.
* **Include the regression equation:** Display the derived linear regression equation (Y = 2.584 + 0.006X) directly on the graph to provide a clear representation of the mathematical relationship.
* **Consider using a more informative graph type:** Instead of a simple scatter plot, consider using a bubble plot where the size of the bubble represents the population density, allowing for visualization of three variables at once.

**General Figure Guidelines:**

* **Ensure high resolution:** All figures should be high resolution (at least 300 dpi) to ensure clarity when printed or displayed electronically.
* **Use a consistent style:** Maintain consistency in font size, style, and formatting across all figures and throughout the manuscript.
* **Refer to figures in the text:** Refer to each figure in the text to guide the reader and explain the key findings illustrated by the figure.

By implementing these suggestions, you can significantly enhance the clarity, informativeness, and overall impact of the figures in your paper. Remember, well-designed figures can greatly aid in communicating your research findings effectively to your audience.