1. Search for another dataset, as my dataset is generated using AI and ChatGPT, there's a strong possibility that the sentences are somewhat similar and don't differ a lot(somehow rephrased and so on).
2. To enhance the model, we should diversify the topics within the dataset. This means using texts that discuss mountains across various contexts, such as books, magazines, blogs, and scientific articles and others. By doing so, we ensure that the model encounters a broader range of language and contexts, leading to improved performance.
3. Additionally, it's essential to balance the classes in our dataset. Right now, the distribution of labels (label 1 and label 0) is skewed, the model may become biased toward the majority class. By ensuring a more balanced representation of both classes, we can train a more robust model that generalizes better across different scenarios.