Important Interview Questions from Project Work

1. What is imputation? Explain the KNN imputation technique.

2. How do we perform imputation for categorical and numerical features?

3. What are the different methods to detect outliers in a dataset?

4. What is the difference between print(df[i].unique()) and print(set(df[i].tolist()))?

5. Explain the usage of: lambda x: 'ckd' if x == 'ckd\t' else x.

6. When should we use .replace() and when should we use .apply() with a lambda function?

7. What is the reason for using multiple .apply() statements like:

```
df['dm'] = df['dm'].apply(lambda x: 'no' if x == '\tno' else x)
df['dm'] = df['dm'].apply(lambda y: 'yes' if y == '\tyes' else y)
df['dm'] = df['dm'].apply(lambda z: z.lstrip())
```

8. Why is .lstrip() not working and giving a float datatype error even though previous lines treated 'dm' as object datatype?

9. How do we convert a column of type 'object' to a numeric type in pandas?

10. What are the best practices to handle null values in a dataset?

11. Why do we use double square brackets [[ ]] in fit_transform()? Explain each part of the code.

12. What is a distplot?

13. How can we visualize both line and curve in a distplot? What might cause it to not show?

14. How can we detect outliers using a distplot?

15. When should we use Label Encoding vs One-Hot Encoding? Consider different scenarios and model types.

16. When should we use normalization and when should we use standardization?

17. What is the empirical rule? How does it relate to z-scores?

18. Explain the classification report and confusion matrix in detail, including all terminology.

19. What is the real-world significance of confusion matrix metrics like precision, recall, and F1-score?

20. Is there a shorter or more efficient way to write: df_imputed['wc'] = df_imputed['wc'].apply(lambda x: '6200' if x == '\t6200' else x)?