**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 July 2024 |
| Team ID | 739761 |
| Project Title | Doctors Annual Salary Prediction |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Dataset | Missing values in the salary column | High | Use imputation techniques such as mean, median, or mode to fill in missing values. |
| Dataset | Outliers in the salary data | Moderate | Use statistical methods or machine learning models to detect and handle outliers (e.g., capping, removal). |
| Dataset | Inconsistent job titles or specialties | Low | Standardize job titles and specialties using a predefined list or mapping. |
| Dataset | Data entry errors (e.g., typos, incorrect salary figures) | Moderate | Implement data validation rules and use scripts to detect and correct errors. |
| Dataset | Incomplete data regarding work experience or education | High | Collect additional data if possible or use imputation techniques. |
| Dataset | Unbalanced dataset with more data from certain regions or specialties | Moderate | Use techniques like oversampling, undersampling, or synthetic data generation to balance the dataset. |
| Dataset | Irrelevant features that do not contribute to predicting annual salary | Low | Perform feature selection using techniques like correlation analysis or feature importance from a machine learning model. |