 

Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 15 October 2024 |
| Team ID | 740663 |
| Project Title | Predicting Diamond Prices With ANN Using Deep Learning |
| 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** |
| Kaggle Dataset | Missing values in columns such as cut, colour, or clarity. | Moderate | Impute missing values using statistical methods (e.g., mode for categorical variables, median for numerical). |
| Kaggle Dataset | Outliers in numerical columns like carat, price, and depth. | High | Use statistical techniques like IQR to identify and  remove/handle outliers. |
| Kaggle Dataset | Incorrect or inconsistent data  types for columns (e.g., carat or price stored as strings). | Moderate | Convert columns to appropriate data types (e.g., float for carat, integer for price). |

|  |  |  |  |
| --- | --- | --- | --- |
| Kaggle Dataset | Redundant columns or features with high multicollinearity (e.g., x, y, z vs. volume). | Moderate | Perform feature selection; calculate correlations to drop redundant features. |
| Kaggle Dataset | Sparse data in columns like depth or table which may not  provide significant predictive power. | Low | Conduct feature importance analysis and drop columns contributing minimally to the prediction model. |