**Gestational Diabetes - CLEANING REPORT**

**CLEANING STEPS :**

* Removed redundant and irrelevant columns and columns with no data (all null)
* Renamed columns appropriately
* Removed and cleaned irrelevant texts, symbols and data in columns
* Created new columns to segregate data depending on the data type
* Changed to appropriate data type, rounded off decimals, changed No and Yes to 0 and 1
* Created new data model which is more readable for the dataset
* Removed empty rows from sub table

**NOISES :**

|  |  |  |
| --- | --- | --- |
| **Noise Type** | **Noises Detected** | **Cleaning Approach** |
| Data as Noise | GCT - many patient values are missing | Dropped the column. |
| Many duplicate columns with same name, but one column with yes and no and other with 1 and 0 | Dropped text columns and kept value columns. |
| Columns with no data (all null values) | Dropped the column. |
|  | Gestational Age data was as 'week + days' | Split the values after delimiter. Divide the second value with 7(to convert no. of days to week in decimal).Added the 2 columns. |
| Feature as Noise | Date of data collection had data with unrealistic ranges | Dropped the column. |
| Column names misspelled | Renamed the columns. |
| Observation as Noise | Date of visit column contained visit status details | Duplicated Date of Visit 3 and parse dates into another column and renamed the new column to Date of V3. Replaced Error with null. Delete dates from Date of Visit 3 and renamed to Visit Status V3 and replaced dates to Visited, Cleaned the text values in Visit 3 Status. |
| Units not mentioned for various test result columns | Research about units for all test results. |
| Unrelated column - eg. List used | Dropped the column. |
| Attribute Noise | Faulty data collection by human errors (unwanted symbols, texts, spaces, case sensitive) | Cleaned the columns using features like filters, splits, extract and changed data types |
| Data contained entries like Lab records lost, sample lost, ?where delivered etc | Removed or replaced with appropriate data and values |

**CHALLENGES ENCOUNTERED :**

* Understanding the meaning and purpose of some columns
* Multiple entries were made for twins field to represent both the babies information
* Date format was changing between excel and power bi
* Changed data type initially & after importing the cleaned data the changes were not reflecting for few columns
* Filtering out the special characters in certain columns (split by delimiter didn’t help) (Eg. U Albumin, 25 OHD V1)
* Additional empty rows uploaded at the bottom while importing excel files
* Missing units for some columns - unable to decide data range and units (Eg. U Albumin, WCC, Hb etc)
* No clear information about the status of patients

**KEY FINDINGS** :

* Missing Age column (might help with detailed access - only 30+ available)
* Minimal Visit 2 information provided
* Delivery date column was empty
* All factors that contribute to GDM and its effect on the baby