Methodology for Extracting Dark Matter Distribution from Gaia Data

1. **Introduction**
   The goal of this project is to transition from basic data analysis of Gaia astronomical data to deriving a *scientifically meaningful distribution of dark matte*r. This document outlines the step-by-step methodology that will be followed to achieve this objective.

   Why Gaia Data Over SDSS?
   Gaia data was chosen over SDSS due to its superior astrometric accuracy, particularly in measuring stellar distances and proper motions. While SDSS provides valuable spectroscopic and photometric data, it lacks the precise parallax and motion measurements necessary for kinematic studies. Gaia's dataset enables a more accurate analysis of stellar velocities, which is crucial for inferring mass distributions and identifying dark matter signatures within the Milky Way.

2. Data Acquisition and Pre-processing

- Utilizing *Gaia DR3 dataset*, specifically extracting columns: **source_id, ra, dec, parallax, pmra, pmdec, and phot_g_mean_mag.**
- The dataset columns represent the following:
  - **source_id**: Unique identifier for each celestial object.
  - **ra**: Right ascension, the celestial equivalent of longitude.
  - **dec**: Declination, the celestial equivalent of latitude.
  - **parallax**: The apparent shift in the position of a star due to Earth's movement, used to estimate distance.
  - **pmra**: Proper motion in right ascension, indicating movement across the sky.
  - **pmdec**: Proper motion in declination, indicating movement across the sky.
  - **phot_g_mean_mag**: Mean magnitude in the Gaia G-band, representing brightness.
- Filter data for quality by removing objects with high parallax uncertainty or negative parallax values.
- Engineer new features:
  - **Total Proper Motion**:

$$\boldsymbol{pm_{total} = \sqrt{pmra^2 + pmdec^2}}$$

  - **Distance Estimation**:

$$\boldsymbol{d = \frac{1}{parallax}}$$

for positive parallax values.

- Normalize and scale data for consistency in training models.

3. Kinematic Analysis and Galactic Dynamics

- Transform celestial coordinates (ra, dec, parallax) into 3D Cartesian positions.
- Compute velocity components from proper motion and radial velocity *(if available).*
- Analyse velocity dispersion as a function of galactic radius to identify mass discrepancies.

4. Modelling Gravitational Effects

- Apply *Jeans equations* to estimate the total gravitational mass distribution of the Milky Way.
- Compare observed stellar velocities with expected values from visible mass to infer the presence of dark matter.

5. Machine Learning for Mass Distribution Estimation

- Train regression models (Random Forest, Neural Networks) to predict mass density from kinematic data.
- Use clustering techniques (K-Means, Gaussian Mixture Models) to detect substructures and anomalies.
- Validate model predictions against existing Milky Way mass distribution models.

6. Dark Matter Density Estimation

- Apply *Bayesian inference* to refine dark matter profile predictions.
- Compare results with known dark matter halo models (e.g., Navarro-Frenk-White profile).
- Generate spatial density maps of dark matter using inferred mass discrepancies.

7. Expected Outcomes and Conclusion

- A data-driven estimation of the dark matter density distribution in the Milky Way.
- Identification of regions with significant deviations from expected mass distributions.
- Comparison with existing astrophysical models to validate findings.

8. Tools and Technologies Used

- **Programming Language**: Python
- **Data Handling**: Pandas, NumPy
- **Visualization**: Matplotlib, Seaborn, Tableau
- **Machine Learning**: Scikit-learn, TensorFlow/PyTorch

- **Astrometric Calculations**: Astropy
- **Computational Framework**: Jupyter Notebook, VS Code

This methodology will ensure a rigorous transition from raw data analysis to extracting meaningful scientific insights into the nature and distribution of dark matter.