METHODOLOGY

1. Data Collection and Dataset Preparation

The system uses two structured datasets: a **resume dataset** and a **job description dataset**.
The resume dataset contains candidate information such as skills, experience, education, certifications, and professional summaries.
The job description dataset includes job roles, required skills, experience requirements, employment type, and detailed job descriptions.

To ensure sufficient data for model training and testing, **synthetic datasets** were generated using Python scripts. This approach allows controlled variation in skills, roles, and experience levels while maintaining realistic HR data patterns.

2. Data Preprocessing

Raw textual data is preprocessed to improve the effectiveness of the matching algorithm.
The preprocessing steps include:

- Conversion of text to lowercase for uniformity
- Removal of special characters and extra whitespace
- Cleaning and normalization of skills and descriptions
- Combination of multiple text fields into a single representative text for resumes and jobs

The cleaned data is stored in processed files and used as input for the matching model.

3. Feature Extraction

To convert textual information into a machine-readable format, **Natural Language Processing (NLP)** techniques are applied.
The combined resume text and job description text are transformed into numerical vectors using **Term Frequency–Inverse Document Frequency (TF-IDF)**.
This representation highlights important skills and keywords while reducing the impact of commonly occurring terms.

4. Resume–Job Matching Algorithm

The matching process is based on **cosine similarity**, which measures the semantic similarity between resume vectors and job description vectors.
For each resume, similarity scores are calculated against available job descriptions.
Higher similarity scores indicate a stronger match between candidate skills and job requirements.

5. Ranking and Recommendation

Based on the computed similarity scores, resumes are ranked for each job role.
The system identifies and recommends the **top-matching candidates** for HR review.

This ranking mechanism reduces manual effort and improves objectivity in the screening process.

6. System Integration

The matching logic is integrated into a backend service that can be accessed through an HR portal.
The modular design separates data processing, model training, and prediction, allowing easy maintenance and future enhancements.

7. Evaluation and Validation

The system’s performance is evaluated by analyzing similarity scores and reviewing the relevance of matched resumes.
Sample test cases are used to validate whether the recommended candidates align with job requirements.