Dokumen Analisis STKI

**Kelompok 2 STBI Magister Ilmu Komputer UGM**  
Anggota:

* Salsabila Nurulfarah Mahmudah (21/475591/PPA/06131)
* Dian Novitaningrum (21/483928/PPA/06177)
* Kartika Rizqi Nastiti (21/485636/PPA/06220)

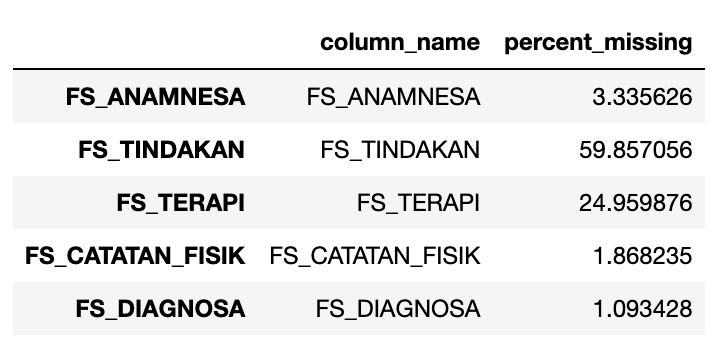
# **Data**

Rekam Medis SOAP

* Banyak kolom : 5 (FS\_ANAMNESA, FS\_TINDAKAN, FS\_TERAPI,

FS\_CATATAN\_FISIK, dan FS\_DIAGNOSA)

* Banyak baris : 126.483
* Format *file* : .xlsx
* Persentase *missing* :



# **Metodologi**

Berikut adalah proses implementasi sebagai analisis pada STKI:

1. *Preprocessing*
   1. *Lowercase,* membuat semua huruf menjadi huruf kecil.
   2. *Remove certain characters,* seperti *tag* HTML, tanda baca, dan karakter lainnya.
   3. *Remove* NaN berdasarkan tiap kolom FS\_ANAMNESA dan FS\_DIAGNOSA.
   4. *Remove stopwords,* seperti kata sambung, kata depan, atau kata-kata lain yang tidak mengandung makna signifikan.
2. Implementasi *Inverted Index,* menggunakan data dari kolom FS\_ANAMNESA dan FS\_DIAGNOSA. Selain itu, kami juga membuat 1 *database* yang berisi semua kolom tanpa melakukan *preprocessing.*
3. Implementasi penghitungan nilai *similarity* antar *terms* menggunakan *Vector Space Model (VSM).* Formulanya didasarkan pada *cosine similarity.*
4. Implementasi penghitungan bobot TF-IDF untuk tiap *term* dari tiap dokumen.
5. Implementasi penentuan dokumen yang mirip dengan *input query* dengan membandingkan nilai *vector space.*
6. *Input* *query* berupa *word* atau *sentences* dan *output* berupa dokumen rekam medis.
7. Evaluasi
   1. *Computation time* (proses *indexing* dan *querying*),
   2. *Precision,* dan
   3. *Recall.*