



**TESIS**

**SISTEM REKOMENDASI AYAT AL-QUR'AN  
BERBASIS K-MEANS CLUSTERING DAN  
COSINE-SIMILARITY DENGAN REDUKSI  
DIMENSI PCA**

**AMARUL AKBAR**

NPM 23066020015

**DOSEN PEMBIMBING**

Dr. Basuki Rahmat, S.Si., MT

Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT

**KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI**

**UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR**

**FAKULTAS ILMU KOMPUTER**

**PROGRAM MAGISTER TEKNOLOGI INFORMASI**

**SURABAYA**

**2026**

*(Halaman ini sengaja dikosongkan)*

**PENGAJUAN TESIS**

**SISTEM REKOMENDASI AYAT AL-QUR’AN BERBASIS**  
**K-MEANS CLUSTERING DAN COSINE-SIMILARITY**  
**DENGAN REDUKSI DIMENSI PCA**

Tesis

Sebagai Salah Satu Syarat Untuk Mencapai Gelar Magister  
Program Studi Magister Teknologi Informasi

Disusun dan diajukan oleh

**AMARUL AKBAR**

**23066020015**

**FAKULTAS ILMU KOMPUTER**  
**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”**  
**JAWA TIMUR**  
**SURABAYA**  
**2026**

*(Halaman ini sengaja dikosongkan)*

**SISTEM REKOMENDASI AYAT AL-QUR'AN BERBASIS  
K-MEANS CLUSTERING DAN COSINE-SIMILARITY  
DENGAN REDUKSI DIMENSI PCA**

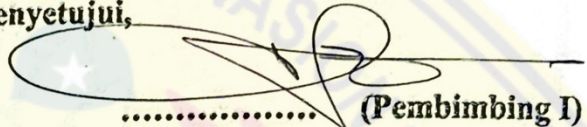
Oleh:

**AMARUL AKBAR**  
**23066020015**

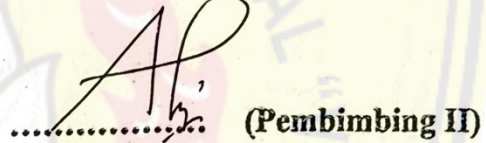
Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Ujian Tesis  
Prodi Magister Teknologi Informasi Fakultas Ilmu Komputer Universitas  
Pembangunan Nasional Veteran Jawa Timur  
Pada Tanggal 12 Januari 2026

Menyetujui,

**Dr. Basuki Rahmat, S.Si., MT.**  
NIP. 19690723 202121 1 002

  
..... (Pembimbing I)

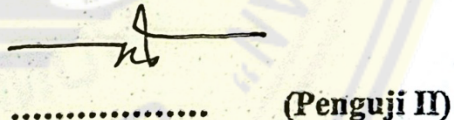
**Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT**  
NPT. 22219860816400

  
..... (Pembimbing II)

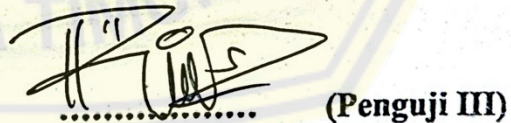
**Dr. Eng. Agussalim, S.Pd., M.T.**  
NIP. 198508112019031005

  
..... (Penguji I)

**Dr. Ir. Mohammad Idhom, S.P., S.Kom., M.T.**  
NIP. 198303102021211006

  
..... (Penguji II)

**Dr. Rizky Parlika, S.Kom, M.Kom.**  
NIP. 198405182021211003

  
..... (Penguji III)

Mengetahui,

Dekan Fakultas Ilmu Komputer

  
**Prof. Dr. Ir. Novirina Hendrasarie, M.T.**  
NIP. 196811261994032001

*(Halaman ini sengaja dikosongkan)*

**SISTEM REKOMENDASI AYAT AL-QUR'AN BERBASIS  
K-MEANS CLUSTERING DAN COSINE-SIMILARITY  
DENGAN REDUKSI DIMENSI PCA**

**Disusun Oleh:**

**AMARUL AKBAR**  
**23066020015**

**Telah disetujui untuk mengikuti Ujian Tesis**

**Menyetujui,**

**Pembimbing I**

**Pembimbing II**



**Dr. Basuki Rahmat, S.Si., MT**  
**NIP. 196907232021211002**



**Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT**  
**NPT. 22219860816400**

**Mengetahui,**

**KOORDINATOR PRODI**



**Dr. Eng. Agussalim, S.Pd., M.T.**  
**NIP. 198508112019031005**

*(Halaman ini sengaja dikosongkan)*



## SURAT PERNYATAAN BEBAS PLAGIASI

Yang bertanda tangan di bawah ini

Nama : Amarul Akbar  
NPM : 23066020015  
Program : Magister (S2)  
Program Studi : Magister Teknologi Informasi (MTI)  
Fakultas : Fakultas Ilmu Komputer

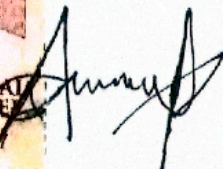
Menyatakan bahwa dalam dokumen ilmiah Tesis ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila dikemudian hari ditemukan indikasi plagiat pada Tesis ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan dipergunakan sebagaimana mestinya.

Surabaya, 23 Januari 2026

Yang Membuat Pernyataan,

  
  
**AMARUL AKBAR**  
NPM.23066020015

*(Halaman ini sengaja dikosongkan)*

## KATA PENGANTAR

Alhamdulillah, puji syukur kehadiran Allah SWT atas segala nikmat rahmat, hidayah dan karunia-Nya. Sholawat serta salam juga kepada baginda Rasulullah Shallallahu ‘Alaihi Wassalam. Penulis sangat bersyukur sehingga tesis dengan judul “SISTEM REKOMENDASI AYAT AL-QUR’AN BERBASIS *K-MEANS CLUSTERING* DAN *COSINE-SIMILARITY* DENGAN REDUKSI DIMENSI *PCA*” dapat terselesaikan dengan baik. Bukan hal yang mudah untuk mewujudkan gagasan ini dalam sebuah susunan tesis. Oleh karena itu, atas bimbingan, arahan dan motivasi dari berbagai pihak, tesis ini dapat diselesaikan sesuai kaidah-kaidah yang telah ditetapkan dengan baik. Untuk itu, penulis mengucapkan terima kasih kepada:

1. Orang tua dan keluarga tercinta yang senantiasa memberikan support, bimbingan, nasihat, do’a dan banyak lagi.
2. Istri Meiga Rahmania dan anak Zakiya Maida Al Ghaziyah yang menjadi support sytem terbaik.
3. Dr. Basuki Rahmat, S.Si., MT, selaku dosen pembimbing pertama dan Wakil Dekan III Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur, yang memberikan saran dan sebagainya dalam penyelesaian tesis ini.
4. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT, selaku dosen pembimbing kedua atas bantuan, kritik, dan saran yang membangun juga.
5. Prof. Dr. Ir. Akhmmad Fauzi, MMT., IPU. selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
6. Prof. Dr. Ir. Novirina Hendrasarie, MT., selaku Dekan Fakultas Ilmu Komputer UPN “Veteran” Jawa Timur.
7. Teman-teman seperjuangan di MTI UPN “Veteran” Jawa Timur yang membantu dan mendukung penulis.
8. Dr. Eng Agussalim, S.Pd.,M.T., selaku penguji I dan Koordinator Prodi MTI, Bapak Dr. Rizky Parlika, S.Kom, M.Kom., selaku pengiji II dan Bapak Dr.Ir. Mohammad Idhom, S.P.,S.Kom., M.T.

9. Para Dosen di jajaran UPN “Veteran” JATIM yang lain yang sudah membantu banyak seperti Pak Dr. I Gede Susrama Mas Diyasa, ST., MT., dan Ibu Dr. Rr. Ani Dijah Rahajoe, ST., M.Cs.
10. Para Dosen yang sudah bersedia menjadi responden dan memberikan masukan atau saran. Bapak Dr. Fazlul Rahman, Lc., MA.Hum., Bapak Mohamad Mirwan., ST., MT. dan Bapak Letda KC Taufikurrahman, S.Pd., M.Pd. Termasuk Ustadz Khairul Auni, S.Pd yang sudah bersedia menjadi tester dan ahli.
11. Teman-teman perjuangan di Malang maupun di Lombok, baik di Komintas YukNgaji Malang, Alumni dan teman kelas Informatika UMM Malang dan teman-teman kelas angkatan SMAN 7 Mataram.
12. Pihak-pihak lain yang tidak dapat disebutkan satu per satu, yang telah membantu dalam penyelesaian tesis ini.

Penulis menyadari bahwa di dalam penyusunan tesis ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan tesis ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya, 23 Januari 2026

Amarul Akbar

## ABSTRAK

**AMARUL AKBAR.** SISTEM REKOMENDASI AYAT AL-QUR'AN BERBASIS *K-MEANS CLUSTERING* DAN *COSINE-SIMILARITY* DENGAN OPTIMASI REDUKSI DIMENSI *PCA* (dibimbing oleh **Dr. Basuki Rahmat, S.Si., MT.** dan **Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT**)

Perkembangan teknologi informasi mempermudah akses Al-Qur'an melalui berbagai platform digital, namun banyaknya ayat dan keragaman tema sering menyulitkan pengguna dalam menemukan ayat yang relevan dengan kebutuhan tertentu. Penelitian ini bertujuan mengembangkan sistem rekomendasi ayat Al-Qur'an berbasis teks terjemahan Bahasa Indonesia menggunakan pendekatan kuantitatif. Sistem dibangun dengan *TF-IDF* sebagai representasi teks, *K-Means Clustering* untuk membentuk struktur global ayat secara leksikal-tematik, serta *Cosine Similarity* sebagai mekanisme perankingan berdasarkan kemiripan kata kunci pengguna. Untuk mengatasi tingginya dimensi data hasil vektorisasi *TF-IDF*, diterapkan *Principal Component Analysis (PCA)* sebagai teknik reduksi dimensi sebelum proses *clustering*. *PCA* sebagai metode pendukung untuk menyederhanakan ruang fitur dan meningkatkan kualitas kluster secara relatif. Evaluasi dilakukan secara kuantitatif menggunakan *Silhouette Score (SS)* dan *Davies-Bouldin Index (DBI)* untuk mengukur kualitas kluster, serta *precision* untuk menilai performa rekomendasi. Hasil pengujian menunjukkan bahwa penerapan *PCA* mampu meningkatkan kualitas kluster dibandingkan model tanpa reduksi dimensi, meskipun nilai *Silhouette Score* masih berada pada kategori cukup. Sistem rekomendasi menghasilkan nilai *precision* 76% pada skenario *Top-5* rekomendasi. Meskipun sistem masih berbasis kesamaan leksikal dan belum menangkap makna semantik secara mendalam, integrasi *K-Means Clustering*, *Cosine Similarity*, dan *PCA* terbukti mampu menghasilkan sistem rekomendasi ayat Al-Qur'an yang relevan dan efisien sebagai pendukung pencarian ayat untuk kebutuhan pembelajaran dan tadabbur.

**Kata kunci :** Sistem Rekomendasi, Al-Qur'an, *K-Means Clustering*, *Cosine Similarity*, *Principle Component Analysis*, *Precision*.

*(Halaman ini sengaja dikosongkan)*

## **ABSTRACT**

**AMARUL AKBAR** *A RECOMMENDATION SYSTEM FOR QUR'ANIC VERSES BASED ON K-MEANS CLUSTERING AND COSINE SIMILARITY WITH PCA-BASED DIMENSIONALITY REDUCTION OPTIMIZATION* (Supervised by **Dr. Basuki Rahmat, S.Si., MT. dan Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT**)

*The development of information technology has facilitated access to the Qur'an through various digital platforms; however, the large number of verses and the diversity of themes often make it difficult for users to find verses that are relevant to specific needs. This study aims to develop a Qur'anic verse recommendation system based on Indonesian translated text using a quantitative approach. The system is built using TF-IDF as text representation, K-Means Clustering to construct a global lexical-thematic structure of verses, and Cosine Similarity as a ranking mechanism based on the similarity of user keywords. To address the high dimensionality of the data resulting from TF-IDF vectorization, Principal Component Analysis (PCA) is applied as a dimensionality reduction technique prior to the clustering process. PCA serves as a supporting method to simplify the feature space and relatively improve cluster quality. The evaluation is conducted quantitatively using Silhouette Score (SS) and Davies-Bouldin Index (DBI) to measure cluster quality, as well as precision to assess recommendation performance. The experimental results indicate that the application of PCA improves cluster quality compared to the model without dimensionality reduction, although the Silhouette Score remains in the moderate category. The recommendation system achieves a precision value of 76% in the Top-5 recommendation scenario. Although the system is still based on lexical similarity and has not yet captured deep semantic meaning, the integration of K-Means Clustering, Cosine Similarity, and PCA is proven to produce a relevant and efficient Qur'anic verse recommendation system to support verse retrieval for learning and contemplative (tadabbur) purposes.*

**Keyword :** *Recommendation System, Al-Qur'an, K-Means Clustering, Cosine Similarity, Principle Component Analysis, Precision.*

*(Halaman ini sengaja dikosongkan)*



## DAFTAR ISI

|                                       |        |
|---------------------------------------|--------|
| HALAMAN JUDUL .....                   | iii    |
| PENGAJUAN TESIS .....                 | v      |
| PENGESAHAN TESIS .....                | vii    |
| PERSETUJUAN TESIS .....               | ix     |
| SURAT PERNYATAAN BEBAS PLAGIASI ..... | xi     |
| KATA PENGANTAR .....                  | xiii   |
| ABSTRAK.....                          | xv     |
| <i>ABSTRACT</i> .....                 | xvii   |
| DAFTAR ISI.....                       | xix    |
| DAFTAR GAMBAR.....                    | xxii   |
| DAFTAR TABEL.....                     | xxvi   |
| DAFTAR KODE .....                     | xxviii |
| DAFTAR NOTASI.....                    | xxx    |
| BAB I PENDAHULUAN.....                | 1      |
| 1.1. Latar Belakang .....             | 1      |
| 1.2. Rumusan Permasalahan .....       | 6      |
| 1.3. Tujuan Penelitian .....          | 7      |
| 1.4. Manfaat Penelitian .....         | 7      |
| 1.5. Ruang Lingkup.....               | 8      |
| 1.6. Kontribusi Penelitian .....      | 9      |
| BAB II TINJAUAN PUSTAKA .....         | 11     |
| 2.1. Penelitian Terdahulu .....       | 11     |
| 2.2. Sistem Rekomendasi.....          | 15     |
| 2.3. Content-Based Filtering.....     | 16     |
| 2.4. K-Means Clustering.....          | 17     |
| 2.5. Cosine Similarity .....          | 18     |
| 2.6. TF-IDF Vectorizer .....          | 19     |

|                                     |                                           |    |
|-------------------------------------|-------------------------------------------|----|
| 2.7.                                | Principle Component Analysis (PCA) .....  | 21 |
| 2.8.                                | Silhouette Score .....                    | 23 |
| 2.9.                                | Davies Bouldin Indeks (DBI) .....         | 25 |
| 2.10.                               | Hyperparameter Tuning K-Means .....       | 26 |
| 2.11.                               | Data Mining .....                         | 27 |
| 2.13.                               | Preprocessing Data.....                   | 30 |
| 2.14.                               | Elbow Method.....                         | 32 |
| 2.15.                               | WordCloud.....                            | 33 |
| 2.16.                               | Python Flask .....                        | 34 |
| 2.17.                               | Jupyter Notebook.....                     | 35 |
| 2.18.                               | Visual Studio Code .....                  | 36 |
| 2.19.                               | Precision.....                            | 37 |
| BAB III METODOLOGI PENELITIAN ..... |                                           | 39 |
| 3.1.                                | Analisis Masalah & Kebutuhan Sistem ..... | 40 |
| 3.1.1.                              | Kebutuhan Fungsional .....                | 40 |
| 3.1.2.                              | Kebutuhan Non-Fungsional .....            | 41 |
| 3.2.                                | Pengumpulan Dataset.....                  | 41 |
| 3.3.                                | Preprocessing Data.....                   | 42 |
| 3.3.1.                              | Cleaning .....                            | 42 |
| 3.3.2.                              | Case Folding .....                        | 43 |
| 3.3.3.                              | Tokenizing .....                          | 44 |
| 3.3.4.                              | Stopword Removal.....                     | 45 |
| 3.3.5.                              | Stemming .....                            | 46 |
| 3.4.                                | Pemodelan K-Means.....                    | 47 |
| 3.4.1.                              | TF-IDF Vectorizer .....                   | 47 |
| 3.4.2.                              | Reduksi Dimensi Menggunakan PCA .....     | 53 |

|                                  |                                   |     |
|----------------------------------|-----------------------------------|-----|
| 3.4.2.                           | Hyperparameter Tuning.....        | 54  |
| 3.4.3.                           | Metode Elbow.....                 | 56  |
| 3.5.                             | Evaluasi K-Means Clustering.....  | 57  |
| 3.6.                             | Sistem Rekomendasi.....           | 57  |
| 3.6.1.                           | Preprocessing Keyword.....        | 59  |
| 3.6.2.                           | TF-IDF Keyword.....               | 59  |
| 3.6.3.                           | Implementasi Model K-Means.....   | 60  |
| 3.6.4.                           | Cosine Similarity.....            | 61  |
| 3.7.                             | Pengembangan Website.....         | 64  |
| 3.8.                             | Pengujian Sistem.....             | 66  |
| 3.8.1.                           | Pengujian Precision.....          | 66  |
| BAB IV HASIL DAN PEMBAHASAN..... |                                   | 69  |
| 4.1.                             | Pengumpulan Dataset.....          | 69  |
| 4.2.                             | Preprocessing Data.....           | 69  |
| 4.2.1.                           | Data Cleaning.....                | 73  |
| 4.2.2.                           | Case Folding.....                 | 74  |
| 4.2.3.                           | Tokenizing.....                   | 76  |
| 4.2.4.                           | Stopwords Removal.....            | 77  |
| 4.2.5.                           | Stemming.....                     | 78  |
| 4.3.                             | Pemodelan K-Means Clustering..... | 84  |
| 4.3.1.                           | TF-IDF Vectorizer.....            | 86  |
| 4.3.2.                           | Skenario Pengujian.....           | 93  |
| 4.4.                             | Evaluasi Model.....               | 99  |
| 4.5.                             | Sistem Rekomendasi.....           | 130 |
| 4.5.1.                           | Preprocessing Keyword.....        | 130 |
| 4.5.2.                           | TF-IDF Keyword.....               | 132 |

|                                  |                                         |     |
|----------------------------------|-----------------------------------------|-----|
| 4.5.3.                           | Implementasi Keyword Model K-Means..... | 133 |
| 4.5.4.                           | Cosine Similarity .....                 | 134 |
| 4.6.                             | Implementasi Website.....               | 143 |
| 4.7.                             | Pengujian Sistem.....                   | 151 |
| BAB V KESIMPULAN DAN SARAN ..... |                                         | 157 |
| 5.1.                             | Kesimpulan .....                        | 157 |
| 5.2.                             | Saran .....                             | 158 |
| DAFTAR PUSTAKA .....             |                                         | 161 |

## DAFTAR GAMBAR

|                                                             |    |
|-------------------------------------------------------------|----|
| <b>Gambar 2. 1</b> Arsitektur sistem CBF .....              | 16 |
| <b>Gambar 2. 2</b> Alur K-Means Clustering .....            | 17 |
| <b>Gambar 2. 3</b> Elbow menunjukkan klaster .....          | 32 |
| <b>Gambar 2. 4</b> Elbow klaster tidak jelas.....           | 33 |
| <b>Gambar 2. 5</b> Visualisasi WordCloud [3] .....          | 34 |
| <b>Gambar 2. 6</b> Logo Flask .....                         | 35 |
| <b>Gambar 2. 7</b> Logo Jupyter Notebook .....              | 35 |
| <b>Gambar 2. 8</b> Logo VSCode .....                        | 36 |
| <b>Gambar 3. 1</b> Tahapan Penelitian .....                 | 39 |
| <b>Gambar 3. 2</b> Dataset dari Kaggle .....                | 41 |
| <b>Gambar 3. 3</b> Tahapan Preprocessing Data .....         | 42 |
| <b>Gambar 3. 4</b> Tahapan Pemodelan K-Means.....           | 47 |
| <b>Gambar 3. 5</b> Alur Perhitungan TF-IDF .....            | 48 |
| <b>Gambar 3. 6</b> Tahapan Sistem Rekomendasi.....          | 58 |
| <b>Gambar 3. 7</b> Arsitektur Sistem Rekomendasi .....      | 61 |
| <b>Gambar 3. 8</b> Wireframe Tampilan Utama.....            | 64 |
| <b>Gambar 3. 9</b> Wireframe Page Hasil Rekomendasi.....    | 65 |
| <b>Gambar 3. 10</b> Wireframe Melihat Detail.....           | 65 |
| <b>Gambar 4. 1</b> Melihat isi file surah CSV .....         | 69 |
| <b>Gambar 4. 2</b> Info Dataset Preprocessing .....         | 72 |
| <b>Gambar 4. 3</b> Seleksi atribut yang dipakai.....        | 73 |
| <b>Gambar 4. 4</b> Hasil Proses Cleaning.....               | 74 |
| <b>Gambar 4. 5</b> Hasil Proses Case Folding .....          | 75 |
| <b>Gambar 4. 6</b> Hasil Proses Tokenizing .....            | 77 |
| <b>Gambar 4. 7</b> Hasil Proses Stopword Removal .....      | 78 |
| <b>Gambar 4. 8</b> Hasil Proses Stemming.....               | 80 |
| <b>Gambar 4. 9</b> Kata frekuensi terbanyak di Dataset..... | 81 |
| <b>Gambar 4. 10</b> Common Words.....                       | 82 |
| <b>Gambar 4. 11</b> Visualisasi WordCloud Dataset.....      | 84 |
| <b>Gambar 4. 12</b> Melihat Hasil Preprocessing.....        | 86 |

|                                                                     |     |
|---------------------------------------------------------------------|-----|
| <b>Gambar 4. 13</b> Hasil Vektorisasi TF-IDF .....                  | 88  |
| <b>Gambar 4. 14</b> Hasil TF pada Dataset .....                     | 89  |
| <b>Gambar 4. 15</b> Hasil IDF pada Dataset.....                     | 89  |
| <b>Gambar 4. 16</b> Hasil TF-IDF .....                              | 89  |
| <b>Gambar 4. 17</b> Membuka file TF-IDF CSV .....                   | 90  |
| <b>Gambar 4. 18</b> Nilai TF-IDF kata pada dokumen.....             | 92  |
| <b>Gambar 4. 19</b> Scree Plot Explained Variance Ratio .....       | 94  |
| <b>Gambar 4. 20</b> Hasil SS dan DBI Baseline Model .....           | 100 |
| <b>Gambar 4. 21</b> Nilai Elbow K-Means Tanpa PCA .....             | 100 |
| <b>Gambar 4. 22</b> BarChart Klaster Baseline Model.....            | 101 |
| <b>Gambar 4. 23</b> Hasil reduksi dengan 6 komponen .....           | 101 |
| <b>Gambar 4. 24</b> Hasil uji hyperparameter 6 komponen .....       | 103 |
| <b>Gambar 4. 25</b> Hasil perbandingan konfigurasi SS & DBI .....   | 105 |
| <b>Gambar 4. 26</b> Konfigurasi terbaik SS atas PCA 6 komponen..... | 106 |
| <b>Gambar 4. 27</b> Hasil evaluasi klaster terbaik PCA 6 .....      | 108 |
| <b>Gambar 4. 28</b> Metode elbow untuk PCA 6 .....                  | 109 |
| <b>Gambar 4. 29</b> Diagram batang PCA 6 komponen.....              | 110 |
| <b>Gambar 4. 30</b> Hasil sebaran PCA 6 komponen.....               | 112 |
| <b>Gambar 4. 31</b> Hasil reduksi dimensi PCA 4 komponen.....       | 113 |
| <b>Gambar 4. 32</b> Hasil uji hyperparameter 4 komponen .....       | 114 |
| <b>Gambar 4. 33</b> Perbandingan SS & DBI PCA 4 komponen .....      | 115 |
| <b>Gambar 4. 34</b> Konfigurasi terbaik PCA 4 komponen .....        | 116 |
| <b>Gambar 4. 35</b> Hasil evaluasi klaster terbaik PCA 4 .....      | 116 |
| <b>Gambar 4. 36</b> Metode elbow untuk PCA 4 .....                  | 117 |
| <b>Gambar 4. 37</b> Diagram batang PCA 4 komponen.....              | 118 |
| <b>Gambar 4. 38</b> Sebaran klaster PCA 4 komponen.....             | 118 |
| <b>Gambar 4. 39</b> Hasil reduksi dimensi PCA 2 komponen.....       | 121 |
| <b>Gambar 4. 40</b> Hasil uji hyperparameter 2 komponen .....       | 121 |
| <b>Gambar 4. 41</b> Hasil perbandingan SS & DBI PCA 2.....          | 122 |
| <b>Gambar 4. 42</b> Konfigurasi terbaik PCA 2 komponen .....        | 123 |
| <b>Gambar 4. 43</b> Hasil evaluasi klaster terbaik PCA 2 .....      | 123 |
| <b>Gambar 4. 44</b> Metode elbow PCA 2 komponen .....               | 124 |

|                     |                                                      |     |
|---------------------|------------------------------------------------------|-----|
| <b>Gambar 4. 45</b> | Diagram batang PCA 2 komponen.....                   | 125 |
| <b>Gambar 4. 46</b> | Sebaran klaster PCA 2 komponen.....                  | 126 |
| <b>Gambar 4. 47</b> | Wordcloud Cluster PCA 2 Komponen.....                | 129 |
| <b>Gambar 4. 48</b> | Wordcloud Klaster Tertinggi PCA 2 Komponen .....     | 129 |
| <b>Gambar 4. 49</b> | DataFrame Lima Rekomendasi Tertinggi .....           | 137 |
| <b>Gambar 4. 50</b> | Nilai TF-IDF dari hasil stemming kata kunci.....     | 139 |
| <b>Gambar 4. 51</b> | Detail Perhitungan TF-IDF dan Cosine Similarity..... | 142 |
| <b>Gambar 4. 52</b> | Isi pada folder aplikasi rekomendasi .....           | 145 |
| <b>Gambar 4. 53</b> | Running Server.py .....                              | 150 |
| <b>Gambar 4. 54</b> | Halaman Home .....                                   | 150 |
| <b>Gambar 4. 55</b> | Page Sistem Rekomendasi.....                         | 151 |
| <b>Gambar 4. 56</b> | Bagian detail ayat di web .....                      | 151 |
| <b>Gambar 4. 57</b> | Kata Berhasil Ditemukan .....                        | 152 |
| <b>Gambar 4. 58</b> | Kata Tidak Ditemukan .....                           | 152 |
| <b>Gambar 4. 59</b> | Perbedaan dengan Pencarian Sederhana .....           | 152 |
| <b>Gambar 4. 60</b> | Pencarian Umum pada Aplikasi Al-Qur'an .....         | 153 |

*(Halaman ini sengaja dikosongkan)*



## DAFTAR TABEL

|                                                                |     |
|----------------------------------------------------------------|-----|
| <b>Tabel 2. 1</b> Penelitian Terdahulu .....                   | 11  |
| <b>Tabel 2. 2</b> Silhouette Score Range .....                 | 25  |
| <b>Tabel 2. 3</b> Davies Bouldin Index Score.....              | 26  |
| <b>Tabel 3. 1</b> Teks Sebelum Preprocessing.....              | 42  |
| <b>Tabel 3. 2</b> Hasil Setelah Cleaning .....                 | 43  |
| <b>Tabel 3. 3</b> Hasil Case Folding .....                     | 43  |
| <b>Tabel 3. 4</b> Hasil Tokenizing .....                       | 44  |
| <b>Tabel 3. 5</b> Hasil Stopword Removal .....                 | 45  |
| <b>Tabel 3. 6</b> Hasil Stemming Data .....                    | 46  |
| <b>Tabel 3. 7</b> Menghitung TF Pada Dokumen .....             | 49  |
| <b>Tabel 3. 8</b> Menghitung IDF Pada Dokumen .....            | 50  |
| <b>Tabel 3. 9</b> Menghitung Nilai TF-IDF Raw .....            | 51  |
| <b>Tabel 3. 10</b> Menghitung TF-IDF Final .....               | 52  |
| <b>Tabel 3. 11</b> Skenario Pengujian.....                     | 55  |
| <b>Tabel 3. 12</b> Nilai TF-IDF Pada Kata Kunci .....          | 62  |
| <b>Tabel 3. 13</b> Gambaran Nilai Cosine Similarity .....      | 62  |
| <b>Tabel 4. 1</b> Pengujian Precision Sistem Rekomendasi ..... | 153 |

*(Halaman ini sengaja dikosongkan)*

## DAFTAR KODE

|                                                                  |     |
|------------------------------------------------------------------|-----|
| <b>Kode 4. 1</b> Import Library Preprocessing .....              | 70  |
| <b>Kode 4. 2</b> Cleaning Data .....                             | 73  |
| <b>Kode 4. 3</b> Case Folding Data.....                          | 75  |
| <b>Kode 4. 4</b> Tokenizing Data.....                            | 76  |
| <b>Kode 4. 5</b> Stopwords Removal Data .....                    | 77  |
| <b>Kode 4. 6</b> Stemming Data .....                             | 79  |
| <b>Kode 4. 7</b> Menghitung Frekuensi Kata Terbanyak .....       | 80  |
| <b>Kode 4. 8</b> Visualiasi BarChart Frekuensi Kata .....        | 82  |
| <b>Kode 4. 9</b> Menampilkan Visualisasi WordCloud.....          | 83  |
| <b>Kode 4. 10</b> Library Tahap Pemodelan K-Means.....           | 85  |
| <b>Kode 4. 11</b> TF-IDF Vectorizer Model.....                   | 87  |
| <b>Kode 4. 12</b> Simpan Hasil TF-IDF ke Format CSV .....        | 90  |
| <b>Kode 4. 13</b> Grid nilai uji hyperparameter.....             | 91  |
| <b>Kode 4. 14</b> Simpan TF-IDF menggunakan pickle.....          | 92  |
| <b>Kode 4. 15</b> Reduksi dimensi dengan komponen PCA .....      | 93  |
| <b>Kode 4. 16</b> Grid Parameter .....                           | 95  |
| <b>Kode 4. 17</b> Mengkombinasikan Parameter .....               | 96  |
| <b>Kode 4. 18</b> Evaluasi Matriks SS & DBI.....                 | 97  |
| <b>Kode 4. 19</b> Visualisasi Metode Elbow .....                 | 98  |
| <b>Kode 4. 20</b> Membandingkan SS & DBI.....                    | 104 |
| <b>Kode 4. 21</b> Memilih Parameter Terbaik .....                | 105 |
| <b>Kode 4. 22</b> Menerapkan parameter terbaik untuk model ..... | 106 |
| <b>Kode 4. 23</b> Mengevaluasi Kembali Klaster Terbaik .....     | 107 |
| <b>Kode 4. 24</b> Visualisasi Data Menggunakan Bar Chart.....    | 109 |
| <b>Kode 4. 25</b> Visualisasi Sebaran PCA Dua Dimensi .....      | 111 |
| <b>Kode 4. 26</b> Menyimpan Model K-Means dengan Pickle .....    | 127 |
| <b>Kode 4. 27</b> Analisis wordcloud pada klaster .....          | 128 |
| <b>Kode 4. 28</b> Preprocessing Pada Keyword .....               | 130 |
| <b>Kode 4. 29</b> Memasukkan Kata Kunci Uji Model .....          | 132 |
| <b>Kode 4. 30</b> Memberikan TF-IDF Pada Keyword.....            | 132 |

|                                                                          |            |
|--------------------------------------------------------------------------|------------|
| <b>Kode 4. 31 Implementasi Model Pada Sistem Rekomendasi.....</b>        | <b>133</b> |
| <b>Kode 4. 32 Penerapan Cosine Similarity.....</b>                       | <b>134</b> |
| <b>Kode 4. 33 Mengambil Rekomendasi Top-N.....</b>                       | <b>135</b> |
| <b>Kode 4. 34 Analisis TF-IDF antara Keyword dan Hasil.....</b>          | <b>137</b> |
| <b>Kode 4. 35 Memperlihatkan nilai TF-IDF pada Keyword.....</b>          | <b>138</b> |
| <b>Kode 4. 36 Detail Perhitungan TF-IDF pada Hasil Rekomendasi .....</b> | <b>139</b> |
| <b>Kode 4. 37 Detail Cosine Similarity pada Hasil Rekomendasi.....</b>   | <b>141</b> |
| <b>Kode 4. 38 Server Flask .....</b>                                     | <b>143</b> |
| <b>Kode 4. 39 Home Flask VSCode .....</b>                                | <b>144</b> |
| <b>Kode 4. 40 Library pada Rekomendasi Py.....</b>                       | <b>146</b> |
| <b>Kode 4. 41 Bagian Index dari Rekomendas Py.....</b>                   | <b>147</b> |
| <b>Kode 4. 42 Bagian Preprocessing Rekomendasi Py.....</b>               | <b>149</b> |

## DAFTAR NOTASI

|                  |                                                            |
|------------------|------------------------------------------------------------|
| $x$              | : Vektor data                                              |
| $n$              | : Jumlah data                                              |
| $k$              | : Jumlah klaster                                           |
| $d$              | : Dimensi data                                             |
| $i, j$           | : Indeks data atau klaster                                 |
| $x_i$            | : Data ke- $i$                                             |
| $c_j$            | : Centroid klaster ke- $j$                                 |
| $\ x_i - c_j\ $  | : Jarak Euclidean antara data ke- $i$ dan centroid ke- $j$ |
| $d(x_i, c_j)$    | : Jarak antara data dan centroid                           |
| $\Sigma$         | : Operator penjumlahan                                     |
| $n_j$            | : Jumlah data pada klaster ke- $j$                         |
| $A$              | : Vektor dokumen/ayat                                      |
| $B$              | : Vektor query                                             |
| $Q$              | : Vektor TF-IDF query                                      |
| $D$              | : Vektor TF-IDF dokumen                                    |
| $Q \cdot D$      | : Perkalian dot product                                    |
| $\ Q\ $          | : Panjang (norma) vektor query                             |
| $\ D\ $          | : Panjang (norma) vektor dokumen                           |
| $\cos(\theta)$   | : Nilai kemiripan Cosine                                   |
| $t$              | : Kata atau term                                           |
| $d$              | : Dokumen                                                  |
| $f(t, d)$        | : Frekuensi term $t$ dalam dokumen $d$                     |
| $tf(t, d)$       | : Term Frequency (TF)                                      |
| $idf(t)$         | : Inverse Document Frequency (IDF)                         |
| $N$              | : Jumlah seluruh dokumen                                   |
| $n_t$            | : Jumlah dokumen yang mengandung term $t$                  |
| $tfidf(t, d)$    | : Nilai bobot TF-IDF                                       |
| $\log$           | : Fungsi logaritma                                         |
| $\Sigma tfidf^2$ | : Normaslisasi TF-IDF                                      |

|                   |                                                    |
|-------------------|----------------------------------------------------|
| $X$               | : Matriks data asli                                |
| $\bar{x}$         | : Rata-rata data                                   |
| $S$               | : Matriks kovarians                                |
| $\lambda$         | : Eigenvalue                                       |
| $v$               | : Eigenvector                                      |
| $Z$               | : Data hasil transformasi PCA                      |
| $PC$              | : Principle Component                              |
| $k$               | : Jumlah komponen utama                            |
| $ S - \lambda I $ | : Determinan matriks karakteristik                 |
| $a_i$             | : Rata-rata jarak intra-klaster                    |
| $b_i$             | : Rata-rata jarak terdekat ke klaster lain         |
| $S_i$             | : Nilai Silhouette data ke- $i$                    |
| $\max(a_i, b_i)$  | : Nilai maksimum antara kohesi dan separasi        |
| $SS$              | : Silhouette Score rata-rata                       |
| $n$               | : Jumlah klaster                                   |
| $S_i$             | : Rata-rata jarak data ke centroid klaster ke- $i$ |
| $d(c_i, c_j)$     | : Jarak antara centroid klaster $i$ dan $j$        |
| $DBI$             | : Nilai Davies Bouldin Index                       |
| $max$             | : Nilai maksimum rasio antar klaster               |
| $\frac{1}{n}$     | : Rata-rata keseluruhan klaster                    |