BAB II

LANDASAN TEORI

2.1. Penelitian Terdahulu

Penelitian yang dilakukan oleh Agus Nursikuwagus dan Tono Hartono dalam judul "*A Decision Support System to Custer a Priority Development Sub Town in Education Field with K-Means Clustering Algorithm (Case Study Center Jawa Province of Indonesia)"* membahas mengenai sebuah sistem pendukung keputusan untuk mengelompokkan prioritas pembangunan kecamatan dalam bidang pendidikan menggunakan algoritma k-means. Tujuan dari penelitian ini adalah untuk mengimplementasikan algoritma k-means terhadap sebuah keputusan yang melibatkan populasi, jumlah ruangan kelas, jumlah guru di dalam sebuah wilayah untuk memprioritaskan bantuan pendidikan. *Output* yang diharapkan dari penelitian ini adalah klasterisasi dari wilayah di Jawa Tengah yang harus diprioritaskan pemerintah untuk membantu pengembangan pendidikan di sana. Hasil dari penelitian ini adalah didapatkannya sebelas daerah yang menjadi prioritas utama untuk dibantu di dalam bidang pendidikan. [5]

Persamaan penelitian yang dilakukan Agus Nursikuwagus dan Tono Hartono dengan penelitian penulis adalah algoritma yang digunakan. Algoritma yang digunakan adalah algoritma k-means untuk melakukan klasterisasi. Sedangkan perbedaannya adalah objek penelitian yang dilakukan. Di mana dalam penelitian ini hal yang diangkat adalah mengenai wilayah yang akan diklaster untuk melihat daerah yang harus diprioritaskan terlebih dahulu. Sedangkan

penelitian penulis mengenai klasterisasi alasan mahasiswa baru masuk ke Universitas Komputer Indonesia.

Penelitian berikutnya membahas penerapan *machine learning* dalam penentuan segmentasi mahasiswa baru. Penelitian ini ditulis oleh Devi Suganti, Hermanus Wim Hapsoro, dan Wahyu Setiono. Algoritma yang dipakai dalam penelitian ini adalah *k-modes*. Data yang digunakan adalah data pendaftaran dari 4 program studi pada tahun 2019 sampai dengan 2021. Total data yang dikumpulkan sebanyak 1.219 pendaftar. Lalu ada 109 variabel yang dilakukan untuk input data pendaftaran. Klaster yang dihasilkan berjumlah 2 buah. Hasil yang didapatkan adalah program studi teknik informatika didominasi mahasiswa berjenis kelamin laki-laki, asal Pekalongan, pekerjaan ayah PNS, dan asal sekolah SMA. Lalu untuk program studi sistem informasi didominasi mahasiswa berjenis kelamin perempuan, asal kota Batang, pekerjaan ayah wiraswasta, dan asal sekolah SMK. [6] Persamaan dari penelitian penulis dengan penelitian ini yaitu objek yang diteliti adalah sama yaitu mahasiswa baru dari sebuah universitas. Perbedaannya terdapat pada algoritma yang digunakan, di mana penulis menggunakan algoritma *k-means*.

Penelitian yang dilaksanakan oleh Zia Tabaruk dan Sultan Bacharuddin Yusuf Hidayat yang berjudul “Strategi Promosi Menjaring Mahasiswa Baru Berdasakan Segmentasi Data PPMB Menggunakan K-Means” membahas mengenai strategi promosi menjaring mahasiswa baru yang didapatkan dari segmentasi data PPMB menggunakan algoritma *k-means*. Data yang digunakan terdiri dari data PPMB pada tahun 2018-2022. Di mana atribut yang digunakan

ada 4, yaitu jenis kelamin, asal sekolah, alamat mahasiswa, dan program studi. Penelitian ini menggunakan *software* WEKA untuk memproses datanya. Klaster yang didapatkan sebanyak 3 buah. Hasil penelitian yang didapatkan adalah mahasiswa baru tahun ajaran 2018-2022 jenis kelamin yang mendominasi adalah laki-laki dengan sekolah asal SMK yang berasal dari Kabupaten Bekasi dan memilih jurusan S-1 Industri. [7]

Persamaan penelitian penulis dengan penelitian Zia Tabaruk dan Sultan Bacharuddin Yusuf Hidayat adalah menggunakan algoritma *k-means* dan sumber data yang sama yaitu data penerimaan mahasiswa baru. Perbedaannya terletak pada *tools* yang digunakan, penulis menggunakan bahasa pemrograman *python* dan google colabs sedangkan penelitian ini menggunakan WEKA *software.*

Penelitian selanjutnya ditulis oleh Adi Sucipto yang berjudul "Klasterisasi Calon Mahasiswa Baru Menggunakan Algoritma K-Means". Penelitian ini bermaksud untuk menggambarkan klaster yang terdapat pada pola nilai CBT dan wawancara di Sekolah Tinggi Multi Media. Algoritma *data mining* yang digunakan pada penelitian ini adalah algoritma *k-means*. Penggunaan algoritma ini dengan nilai k berjumlah 4. Data yang digunakan adalah data calon mahasiswa yang telah mengikuti tahapan pendaftaran sampai dengan registrasi, yang berjumlah sebanyak 5560 calon mahasiswa. Atribut yang digunakan ada 4 yaitu nomor, nomor *test*, CBT, wawancara, dan total nilai. Di mana untuk CBT dan wawancara itu adalah nilai yang didapatkan masing-masing calon mahasiswa. Hasil dari penelitian didapatkan bahwa dari empat klaster yang dibentuk, terdapat 2 klaster, yaitu klaster pertama dan ketiga yang menunjukkan jumlah diterimanya

signifikan sehingga bisa dijadikan pedoman untuk seleksi calon mahasiswa baru. [8]

Persamaan penelitian penulis dengan penelitian yang ditulis oleh Adi Sucipto ini terdapat pada kesamaan algoritma yang akan dipakai. Di mana penulis dan Adi Sucipto menggunakan *k-means* sebagai algoritma utama. Lalu untuk data yang digunakan sama yaitu data calon mahasiswa baru. Perbedaannya terletak pada batasan data yang digunakan, di mana pada penelitian Adi Sucipto ini data yang digunakan meliputi berbagai program studi sedangkan penulis membatasi data yang akan digunakan pada satu program studi saja yaitu sistem informasi.

Penelitian selanjutnya membahas mengenai pola segmentasi konsumen pemegang kartu kredit. Penelitian ini ditulis oleh Sakshi Priyadarshni, Rakshan Fathima, Siddhaling Urolagin, Anupkumar M. Bongale, and Deepak Sudhakar Dharrao. Judul dari penelitian ini adalah "*Unveiling Customer Segmentation Patterns in Credit Card Data using K-Means Clustering: A Machine Learning Approach*". Penelitian ini diproses menggunakan algoritma *k-means* dan *python* 3.10. *Libraries* yang dipakai yaitu NumPy, Pandas, dan Scikit-Learn. Nilai k yang digunakan untuk klasterisasi yaitu 3. Penelitian ini menyoroti pentingnya untuk terus menyempurnakan dan mengembangkan strategi segmentasi untuk beradaptasi dengan perubahan kebutuhan nasabah dan dinamika pasar. Penggunaan *k-means* dalam penelitian ini memberikan pemahaman yang lebih tentang pelanggan mereka, mengidentifikasi peluang pasar yang belum dimanfaatkan dan mengoptimalkan alokasi sumber daya. [9]

Persamaan antara penelitian penulis dengan penelitian yang dijabarkan yaitu pada algoritma yang digunakan (*k-means*). Persamaan lainnya adalah mengenai tujuan segmentasi agar mendapatkan pemahaman yang lebih baik lagi untuk mengoptimalkan sumber daya yang ada dan memberikan pelayanan yang terbaik kepada pelanggan (dalam hal ini mahasiswa baru). Sedangkan perbedaannya terdapat pada data yang digunakan. Dalam penelitian ini data yang digunakan adalah data penggunaan kartu kredit oleh nasabah bank. Sementara itu, penelitian penulis menggunakan data penerimaan mahasiswa baru.

2.2. Teori Pendukung

Teori pendukung adalah dasar teori dalam penelitian yang digunakan sebagai acuan dalam melakukan penelitian oleh penulis.

2.2.2. Data Mining

Data mining merupakan perluasan dari analisis data tradisional dan pendekatan statistik dalam hal menggabungkan teknik analisis yang diambil dari berbagai disiplin ilmu termasuk, namun tidak terbatas pada analisis numerik, pencocokan pola dan bidang kecerdasan buatan seperti pembelajaran mesin, dan jaringan saraf dan algoritma genetik. Meskipun banyak tugas penggalian data mengikuti pendekatan analisis data tradisional berbasis hipotesis, namun biasa untuk menggunakan pendekatan oportunistis, pendekatan berbasis data yang mendorong pola algoritma pendeteksian untuk menemukan tren, pola, dan hubungan yang berguna. [10]

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data. [11]

Data mining merupakan proses semi otomatik yang menggunakan teknik statistika, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang tersimpan di dalam *database* besar. *Data mining* adalah bagian dari proses KDD (*Knowledge Discovery in Databases*) yang terdiri dari beberapa tahapan seperti pemilihan data, pra pengolahan, transformasi, *data mining*, dan evaluasi hasil. KDD secara umum juga dikenal sebagai pangkalan data. [12]

Mining diartikan dari sejumlah besar material dasar. *Data mining* adalah untuk mengolah data sehingga menghasilkan informasi baru yang bermanfaat. Proses penggalian informasi dari sebuah *dataset* atau kumpulan data disebut dengan *data mining. Data mining* digunakan dalam aplikasi yang lebih luas. Salah satu metode yang termasuk *data mining* adalah *clustering* atau pengelompokan adalah alat yang digunakan dalam ilmu data . [13]

Dari penjelasan di atas dapat kita simpulkan bahwa *data mining* adalah proses menggunakan teknik-teknik komputer untuk menganalisis dan mengekstraksi pengetahuan dari basis data secara otomatis. Tujuannya adalah menggali informasi berharga yang tidak diketahui sebelumnya secara manual. Proses ini melibatkan teknik statistika, matematika, kecerdasan buatan, dan

machine learning. Ini adalah bagian dari KDD dan digunakan untuk menghasilkan informasi baru yang berguna dari data yang ada.

2.2.2.1. Metode Data Mining

Dalam *data mining* ada banyak metode yang bisa digunakan, metode ini digunakan sesuai dengan kasus yang dihadapi. Hal itu dikarenakan setiap metode yang ada mempunyai fungsi dan tujuan yang berbeda-beda. Berikut beberapa metode yang ada di dalam *data mining*.

1. Asosiasi (*Association*)

   Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses di mana hubungan asosiasi muncul pada setiap kejadian. Salah satu contohnya *Market Basket Analysis*, yaitu salah satu metode asosiasi yang menganalisis kemungkinan pelanggan untuk membeli beberapa item secara bersamaan.

2. Klasterisasi (*Clustering*)

   Klasterisasi merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Klaster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan tidak dengan *record-record* dalam kluster lain.

3. Prediksi (*Prediction*)

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang. Metode yang dapat digunakan adalah Regresi Linear Berganda

4. Estimasi (*Estimation*)

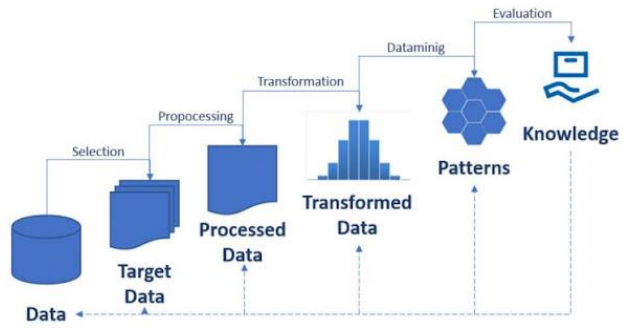
   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

5. Klasifikasi (*Classification*)

   Klasifikasi adalah fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (item) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan.

2.2.2.2. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) adalah proses yang bertujuan untuk menggali dan menganalisis sejumlah besar himpunan data dan mengekstrak informasi serta pengetahuan yang berguna.



Gambar 2. 1 *Knowledge Discovery in Database*

(Sumber:https://miqbal.staff.telkomuniversity.ac.id/belajar-data-mining/[14]**)**

Tahapan yang terdapat dalam KDD sebagai berikut.

1. *Data Selection*

   Data selection adalah tahap menetapkan variabel yang akan digunakan dalam proses *data mining*.[15]

2. *Data Preprocessing*

   Preprocessing memiliki dua tahapan yaitu tahap pertama melakukan proses *cleaning* yakni memeriksa data apakah terdapat penggandaan data atau terdapat data yang tidak konsisten, serta memperbaiki kesalahan pada data, seperti kesalahan huruf. Tahapan kedua yaitu *integration*, tahap ini dilakukan terhadap atribut yang mengidentifikasikan entitas yang unik.[15]

3. *Data Transformation*

   Tahap untuk mengubah data sesuai dengan format yang mendukung pengolahan *data mining*.[15]

4. *Data Mining*

   Tahapan utama untuk formula yang dijalankan agar memperoleh *knowledge* dari data yang diolah. Untuk penelitian ini diterapkan teknik *clustering* yaitu menggunakan metode *K-Means Clustering*.[15]

5. *Interpretation/Evaluation*

   Tahapan untuk mengidentifikasi hubungan-hubungan yang menarik di dalam *knowledge base* yang sudah diidentifikasi serta menghasilkan pola-pola khas maupun model prediksi yang dievaluasi untuk menilai kajian yang ada sudah memenuhi target yang diinginkan.[15]

6. *Knowledge Presentation*

Menampilkan pola informasi yang dihasilkan dari proses *data mining* kepada pengguna, visualisasi ini membantu mengkomunikasikan hasil *data mining* dalam bentuk yang mudah dimengerti. Tahapan ini menghasilkan pemahaman baru untuk semua orang yang nantinya bisa dijadikan acuan pengambilan keputusan. [15]

2.2.3. *K-Means Clustering*

K-means *clustering* adalah algoritma *unsupervised* yang digunakan untuk mengelompokkan objek-objek yang berbeda ke dalam klaster-klaster. klaster adalah kumpulan objek-objek data yang homogen di dalam satu klaster dan heterogen dengan objek-objek di dalam klaster yang lain. klaster objek-objek data tersebut dapat diperlakukan secara bersama-sama sebagai satu kelompok, dan dengan demikian dapat dianggap sebagai salah satu bentuk kompresi data. algoritma *k-means clustering* memiliki beberapa tahapan sebagai berikut:

1. Memberi label jumlah klaster
2. Menentukan koordinat pusat (*centroid*)
3. Tentukan jarak setiap objek ke *centroid*
4. Mengelompokkan objek berdasarkan jarak minimum

k-means clustering adalah metode partisi yang mengelompokkan 'n' observasi ke dalam 'k' klaster berdasarkan jarak minimum antara pusat klaster dan titik observasi.[16]

Proses ini dilakukan secara iteratif sehingga titik pengamatan berada pada jarak minimum dari pusat klaster. *K-Means clustering* menggunakan berbagai

fungsi jarak untuk mengukur kemiripan antar objek. fungsi jarak yang digunakan oleh algoritma ini adalah fungsi metrik jarak *Euclidean* dan fungsi metrik jarak *Manhattan*.[16]

Algoritma *k-means* sederhana dan dapat digunakan untuk berbagai macam tipe data, namun memiliki beberapa kelemahan karena algoritma *k-means* secara komputasi mahal dan membutuhkan lebih banyak waktu yang berkaitan dengan jumlah item data, jumlah klaster dan jumlah iterasi, sehingga perlu menentukan jumlah klaster terlebih dahulu. [16]

Langkah-langkah algoritma *k-means* adalah:

1. Menginputkan data
2. Menentukan jumlah klaster
3. Menentukan titik pusat klaster (*centroid*). Ambil data secara acak sebagai pusat klaster (*centroid*) awal.
4. Menghitung jarak data ke pusat klaster dengan menggunakan rumus persamaan *Euclidean Distance*:

$$D_{(i,j)} = \sqrt{\left(X_{1i} - X_{1j}\right)^2 + \left(X_{2i} - X_{2j}\right)^2 + \cdots + \left(X_{ki} - X_{kj}\right)^2} \quad \text{.....persamaan (1)}$$

$D_{(i,j)}$ pada rumus di atas adalah nilai jarak data ke-i ke pusat klaster. Untuk X_{ki} merupakan data ke-i pada atribut data k, dan X_{kj} merupakan jarak titik pusat ke-j pada atribut k.[17]

5. Menghitung pusat klaster baru dengan rumus rata-rata *centroid*

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \quad \text{.....persamaan (2)}$$

c_j berarti sebagai *centroid* baru dari klaster ke-j. $|S_j|$ adalah jumlah data poin yang ditetapkan ke klaster ke-j. S_j adalah set dari semua data poin yang ditetapkan ke klaster ke-j. Terakhir, x_i adalah data poin dalam set S_j.

6. Mengklasifikasikan data. Data diklasifikasikan berdasarkan kedekatannya dengan pusat klaster (*centroid*). Data dikelompokkan ke dalam klaster dengan nilai yang terdekat. Lakukan proses perhitungan pusat klaster baru berdasarkan rata-rata anggota yang ada pada klaster tersebut. Proses mengelompokkan data (iterasi) dilakukan sampai hasil iterasi bernilai sama dengan iterasi sebelumnya. [17]

2.3. Piranti Pendukung

Piranti pendukung adalah beberapa alat dan perangkat yang digunakan oleh penulis untuk mendapatkan hasil penelitian.

2.3.1. Python

Gambar 2. 2 Logo *Python*

(Sumber: python.org [18]**)**

Python adalah bahasa pemrograman yang memiliki keberagaman luas. Hanya diperlukan alat dan perpustakaan (*library*) yang tepat, lalu kamu bisa menjadi inovator sejati. Memulai belajar bahasa pemrograman membutuhkan nyali, kemauan, waktu, dan mungkin sejumlah minuman berenergi. Oleh karena itu, harus mulai dari menetapkan tujuan dan mempelajari apa saja kegunaan *python*. *Python* amat sangat mudah untuk dibaca. Sebagai *interpreted language* (bahasa pemrograman yang tidak perlu dikompilasi), python tidak mengubah kodenya untuk menjadi terbaca oleh komputer. Bahasa ini juga merupakan bahasa pemrograman tujuan umum tingkat tinggi. Para pengembang mendesainnya untuk menjadi bunglon dari dunia pemrograman. Selain itu, *Python* bertujuan untuk menghasilkan kode yang lebih jelas dan lebih logis tidak hanya untuk proyek skala kecil tetapi juga untuk yang lebih besar. Kamu dapat membandingkan *python* dengan kubus rubik: ia memiliki banyak sisi sehingga dapat memutar dan bermain-main dengannya. [19]

```
#Menginput Angka
angka = float(input("Tulis Sebuah Angka: "))

#Menampilkan Kondisi Angka Positif
if angka > 0:
   print("Angka Positif")

#Menampilkan Kondisi Angka Nol
elif angka == 0:
   print("Angka Nol")

#Menampilkan Kondisi Angka Negatif
else:
   print("Angka Negatif")
```

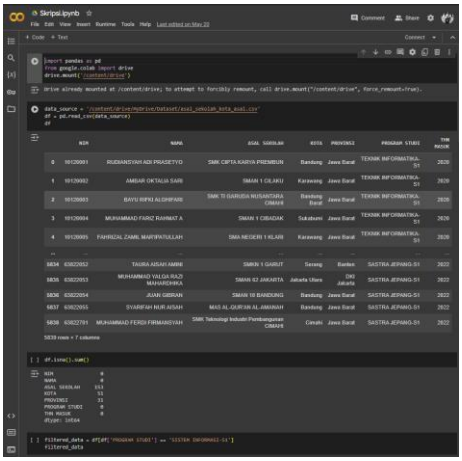
Gambar 2. 3 Contoh Sintaks Python

2.3.2. Google Colab

Gambar 2. 4 Logo *Google Colab*

(Sumber: https://commons.wikimedia.org/wiki/File:Google_Colaboratory_SVG_Logo.svg [20])

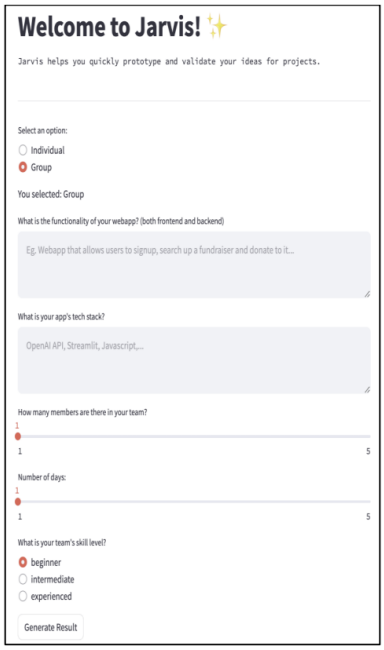
Google colab adalah sebuah web *Integrated Development Environment* (IDE) untuk *python* dan dengan demikian, semua orang dapat belajar *python* tanpa menginstal apa pun, karena *python* berjalan di browser web. Kita hanya membutuhkan browser web dan akun Google. Google Colab adalah alat yang berguna untuk memperkenalkan dan mengajarkan *python* karena beberapa alasan. Seperti pilihan untuk menginstal lingkungan *python* untuk komputasi ilmiah, yang dapat membingungkan. [21]



Gambar 2. 5 *User Interface Google Colab*

2.3.3. Streamlit

Streamlit adalah *framework open-source* dari *python* yang memungkinkan untuk membuat aplikasi web menggunakan bahasa *python* dalam mengaplikasi model dari *machine learning* atau *data science*. [22] Berfungsi untuk menyebarkan model dan visualisasi dengan mudah menggunakan bahasa *python*, yang cepat dan minimalis tetapi juga memiliki tampilan yang cukup baik serta ramah pengguna. Tersedia *widget* bawaan untuk masukan pengguna, seperti pengunggahan gambar, penggeser, masukan teks, dan elemen hypertext markup language (HTML) lain yang sudah dikenal, seperti checkboxes dan radio buttons. Setiap kali pengguna berinteraksi dengan aplikasi *Streamlit*, skrip *python* dijalankan kembali dari atas ke bawah. Hal ini merupakan sebuah konsep penting yang perlu diingat saat mempertimbangkan berbagai status aplikasi yang akan dipilih.[23]

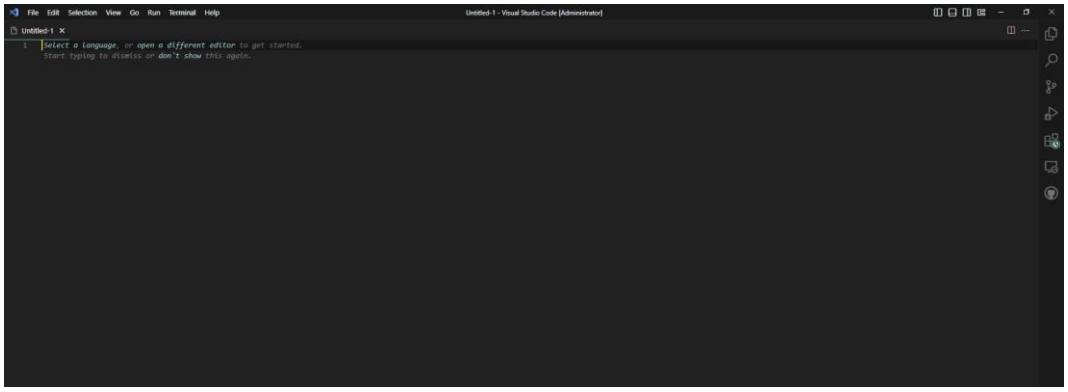


Gambar 2. 6 Contoh *Streamlit Web Based*

(Sumber: "Developing a Website to Analyze and Validate Projects Using LangChain and Streamlit" [24]**)**

2.3.4. Visual Studio Code

Visual Studio Code adalah kode editor sumber yang dikembangkan oleh Microsoft untuk Windows, Linux dan macOS. Ini termasuk dukungan untuk *debugging*, kontrol git yang tertanam dan GitHub, penyorotan sintaksis, penyelesaian kode cerdas, *snippet*, dan *refactoring* kode. Ini sangat dapat disesuaikan, memungkinkan pengguna untuk mengubah tema, pintasan *keyboard*, preferensi, dan menginstal ekstensi yang menambah fungsionalitas tambahan.[25]



Gambar 2. 7 *User Interface* Visual Studio Code