**개인 소비 패턴 분석 및 합리적 지출 클러스터링 대시보드**

20252739 김현민

**1. 주제 및 선정 동기**

**주제**: 대학생 개인 소비 패턴 분석 및 합리적 지출 클러스터링 대시보드

**선정 동기**: 최근 조사에 따르면 한국 대학생의 월평균 생활비 지출은 약 59만 2천 원에 달하며, 이는 5년 전에 비해 월평균 약 22만 6천 원이 상승한 수치이다. 이처럼 재정 관리는 대학생들에게 매우 현실적이고 중요한 문제이며, 본 프로젝트는 자신의 지출 습관 개선에 즉시 적용될 수 있을 것으로 기대된다.

**2. 프로젝트 요약 및 목표**

**목표**: 개인의 금융 거래 데이터를 Python 기반의 데이터 분석 기술(Pandas, Scikit-learn)과 웹 시각화 도구(Streamlit)를 활용하여 사용자 친화적인 대시보드를 구축하는 것을 목표로 한다.

**핵심**: K-Means 클러스터링 기법을 적용하여 지출 내역을 객관적인 소비 패턴(클러스터)으로 분류하고, 이를 시각화하여 데이터 기반의 재정 관리를 지원하는데 중점을 둔다.

**주요 특징**:

* 데이터 전처리 자동화(Regex 기반 카테고리 분류)
* Scikit-learn을 이용한 K-Means 클러스터링 및 최적 K결정
* PCA 기반의 2차원 클러스터 시각화
* Streamlit을 활용한 인터랙티브 대시보드 구현

**3. 문제 정의 및 해결 방안**

**3.1. 문제 정의**

* **잠재적 소비 성향 인식의 난제**: 단순 통계만으로는 파악하기 어려운 복합적인 소비 패턴(예: 충동적, 고정적, 사치적 지출 등)을 정량적이고 객관적으로 파악하는 방법이 없다.
* **데이터 정제 및 구조화의 기술적 복잡성**: 금융 거래 내역의 'Description' 필드는 비정형 텍스트를 포함하며, 이를 수치적 변수로 변환하고 카테고리를 부여하는 과정이 높은 기술적 장벽이 된다.
* **분석 결과의 활용 및 접근성 부족**: 비전문가인 사용자가 Python 스크립트가 아닌 웹 환경에서 직관적이고 인터랙티브하게 결과를 탐색하고 활용할 수 있는 사용자 인터페이스(UI) 구축 능력이 부족하다.

**3.2. 해결 방안**

* **데이터 정제 및 특징 공학을 통한 구조화**: Pandas를 활용하여 데이터를 정제하고, 정규표현식(Regex)을 통해 비정형의 Description 필드를 표준 Category로 변환한다. 이후 거래 금액, 빈도, 카테고리별 지출 비율 등의 패턴을 포착하는 특징 벡터를 생성하여 K-Means 입력 데이터로 활용한다.
* **비지도 학습을 통한 객관적 패턴 발견**: Scikit-learn의 K-Means 클러스터링을 적용하여 지출 내역들을 행동적 유사성에 따라 자동 군집화하고, 최적의 클러스터 개수(K)를 결정하는 Elbow Method와 Silhouette Score를 함께 사용한다.
* **Streamlit 기반의 직관적 시각화 구현**: Streamlit을 활용하여 인터랙티브 대시보드를 구축한다. 특히, PCA를 통해 2차원으로 축소된 클러스터 분포를 시각적으로 제공하고, 각 클러스터의 특성을 자동 요약하는 텍스트 기능을 추가한다.

**4. 기술 구현**

**4.1. 시스템 개요 및 데이터 흐름**

1. **데이터 획득**: 개인 금융 거래 내역 CSV 파일을 입력으로 받으며, 입력 데이터는 최소한 Date, Description, Amount, In/Out 필드를 포함해야 한다.
2. **전처리 및 특징 공학**: Pandas와NumPy를 사용하여 정규표현식으로 비정형Description을 15개 내외의 표준 Category로 분류하고, 지출 패턴을 정량화하는 특징 벡터를 생성한다.
3. **모델링 및 분석**: Scikit-learn K-Means 클러스터링 알고리즘을 특징 벡터에 적용하여 소비 패턴을 군집화하고, PCA를 적용하여 특징 벡터의 차원을 2차원으로 축소한다.
4. **시각화 및 출력**: Streamlit을 활용하여 통합 대시보드를 구현하며, Plotly와 Seaborn을 사용하여 클러스터 분포 및 카테고리별 트렌드를 인터랙티브하게 시각화한다.

**4.2. 필요한 기술 요소**

|  |  |  |
| --- | --- | --- |
| **분류** | **주요 라이브러리/도구** | **핵심 역할 및 적용 지점** |
| 데이터 처리 | Pandas, NumPy | CSV 파일 처리, 데이터 정규화, 특징 공학을 통한 패턴 변수 생성 |
| 기계 학습 | Scikit-learn (K-Means, PCA) | 비지도 학습 모델 구현, 최적 K값 결정, 차원 축소를 통한 시각화 준비 |
| 시각화 | Plotly, Seaborn | 클러스터 분포, 카테고리별 비율 등 인터랙티브 차트 생성 |
| 웹 인터페이스 | Streamlit | 분석 결과를 통합하는 대시보드 형태의 사용자 인터페이스 구축 및 배포 |

**5. 사용자 인터페이스 (UI)**

Streamlit을 기반으로 구현되며, 다음 세 가지 핵심 시각화 결과를 포함한다.

1. **클러스터 분포 시각화**: 다차원 특징을 PCA를 통해 2차원 평면에 투영하여, 지출 내역들이 색깔별로 어떻게 군집되어 있는지를 인터랙티브하게 보여준다.
2. **클러스터별 특성 요약 테이블**: 각 클러스터의 평균 지출 금액, 거래 빈도, 가장 높은 비중을 차지하는 카테고리 등을 수치 및 텍스트로 요약하여 제공한다.
3. **지출 트렌드 분석**: 전체 지출 내역에 대한 월별 또는 카테고리별 지출 추이를 막대 그래프 또는 선 그래프로 시각화한다.

**6. 구현 방법 및 단계별 계획**

**단계 1: 데이터 확보 및 구조화**

* **핵심 작업:** 개인 금융 거래 데이터를 CSV 형식으로 통합하고, Pandas를 이용하여 데이터를 불러오고 초기 결측치 및 오류를 처리한다. 정규표현식(Regex)을 사용하여 Description필드의 비정형 텍스트를 식비, 교통비, 취미 등 15개 내외의 표준 카테고리로 변환하는 분류 로직을 구축한다. 이 과정은 데이터 품질을 결정하는 기초 단계이다.

**단계 2: K-Means 모델 구축**

* **핵심 작업:** 클러스터링의 성능을 높이기 위해 금액, 빈도, 카테고리별 비율 등 핵심 패턴을 포착할 수 있는 특징 벡터를 생성한다. Scikit-learn을 사용하여 K-Means 모델을 훈련시키고, 군집 분석의 기본인 Elbow Method 및 Silhouette Score를 시각화하여 가장 의미 있는 클러스터 개수를 결정한다. 이후 PCA를 적용하여 다차원 데이터를 2차원으로 축소하여 시각화 준비를 완료한다.

**단계 3: Streamlit 대시보드 구현 및 시각화**

* **핵심 작업:** Streamlit을 기반으로 프론트엔드 인터페이스를 설계하고 구현한다. Plotly를 사용하여 PCA 기반의 2D 클러스터 산점도를 인터랙티브하게 표시한다. 각 클러스터에 대해 평균 금액, 주요 거래 시간, 최빈 카테고리 등 정량적 특성을 추출하여, 사용자가 클러스터의 의미를 즉각적으로 파악할 수 있는 요약 텍스트 또는 테이블 형태로 제공한다.

**단계 4: 통합, 테스트 및 문서화**

* **핵심 작업:** 데이터 입력부터 최종 대시보드 출력까지 전체 파이프라인의 통합 테스트를 수행하고 최적화를 진행한다. 최종 코드는 Git Repository에 업로드하고, 프로젝트 개요, 설치 및 사용법, 분석 결과를 포함하는 상세한 README 파일을 작성한다.

**7. 기대 효과**

월평균 59만 2천 원의 생활비를 관리하는 대학생들에게, 자신의 지출 습관을 객관적인 클러스터로 정의하는 도구를 제공하여, 데이터 기반의 합리적인 재정 계획 수립과 실질적인 통제력을 얻게 한다.