아이디어 계획서

충청남도 데이터 분석 아이디어 공모전 기획서

아이디어 요약서

1 아이디어 명칭

실시간 IPTV 시청 데이터 분석을 통한 사회적 취약계층 고립예방

2 분석 목적

현대 사회는 고령화와 1인 가구 증가로 인해 인구 구조가 급격히 변화하고 있다. 통계청 자료에 따르면, 2024년 기준으로 65세 이상 고령 인구가 전체의 약 18%를 차지하고, 1인 가구는 전체 가구의 3분의 1을 넘어섰다. 이러한 변화 속에서 경제적, 신체적, 정서적으로 취약한 계층은 점점 더 사회와의 접촉이 줄어들며 고립의 위험에 처하고 있다.

특히 사회적 고립은 단순히 대인관계의 감소에 그치지 않는다. 연구 결과에 따르면, 사회적으로 고립된 개인은 그렇지 않은 사람들에 비해 우울증 발병 위험이 높고, 심혈관 질환과 치매에 걸릴 가능성도 더 크다. 고립은 불규칙한 생활 습관, 영양 불균형, 질병 발견 지연 등을 통해 전반적인 건강에 부정적인 영향을 미친다. 여기에 경제적 어려움까지 겹치면, 의료 서비스 이용이 제한되고 기본적인 생활비를 감당하지 못하는 악순환에 빠진다.

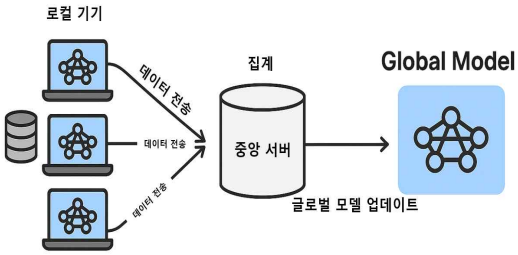
이러한 상황에서 가장 심각하게 우려되는 문제는 고독사다. 보건복지부의 고독사 실태조사에 따르면, 고독사 사망자 수는 매년 증가하고 있으며, 특히 50~60대 중장년층과 70대 이상 고령층에서 발생률이 높다. 고독사의 원인은 1인 가구 증가뿐만 아니라 지역사회와의 연결성 약화, 주변인과의 단절, 건강 악화 등이 복합적으로 작용한다.

따라서 사회적 취약계층의 고립을 예방하고 지역사회와의 연결망을 강화하는 것은 개인의 삶의 질 향상뿐 아니라 사회 전체의 지속가능성을 위해 반드시 해결해야 할 시급한 과제이다.

3 핵심 내용

현재 사회는 스마트 기기, 사물인터넷, 온디바이스 등 디지털 기술의 발전으로 개별 기기의 활용 영역이 확대되면서, 분산형 데이터의 종류와 규모가 급격히 증가하고 있다. 스마트 기기, IPTV, 의료분야도 이러한 사례에 해당하며, 데이터가 분산되어 생성될 경우 분석을 위해 데이터를 중앙으로 재수집해야 하는 별도의 과정이 필요하다. 기존 중앙 집중식 분석 모델에서 분산형 데이터를 처리하려면 데이터 수집 및 전처리에 전체 분석 시간의 80%를 소모하고, 이와 동시에 막대한 비용이 발생한다. 본 내용에서는 이러한 문제를 해결하기 위해 연합학습 기반의 IPTV AI 실시간 시청 데이터 분석 모델을 제안한다.

1). 연합학습은 각 클라이언트에서 로컬 데이터를 활용해 모델을 학습한 후, 모델의 가중치만 서버와 공유하여 글로벌 모델을 만드는 분산 기계학습 방식이며, 분산형 데이터 처리에 탁월한 머신러닝 기술이다. 과정은 학습 알고리즘 분배 및 학습, 업데이트 전송, 집계 순으로 진행하게 되며 학습 알고리즘 분배 및 학습 단계에서는 초기 학습 알고리즘을 각 기기에 분배하여 각각의 기기에서 사용자의 데이터에 대해 학습을 진행하게 된다. 학습을 마친 뒤 각 기기에서 모델 가중치, 업데이트에 대한 정보를 주기에 따라 중앙서버로 보낸다. 중앙서버에서 데이터를 집계하여 새로운 글로벌 모델을 생성한다. 이 과정은 최종모델을 완성할 때까지 반복 수행한다. 연합학습은 개인정보 보호 문제와 데이터가 독립적이고 동일하게 분포되어 있지 않는 데이터 비균질성의 문제를 해결하며, 많은 양의 분산된 데이터를 수집함에 있어 효과적으로 처리할 수 있다.



(그림 1) 연합학습 과정

2). 로컬 기기에는 강화학습을 적용하여 학습, 분류, 예측, 보상, 업데이트의 다섯가지 단계를 진행하게 된다. 강화학습은 데이터 중 반복되는 패턴 확인과 특정 신호를 찾고 체계화한다. 생성된 데이터에서 행동할 수 있는 경우의 수를 바탕으로 점수를 부여하고, 합산된 점수를 통해 스스로 학습하여 특정한 상황에 따라 행동하게 된다. 수행한 행동에 따라 보상단계에서 보상을 받고, 더 많은 보상 및 최상의 행동을 학습하기 위해 업데이트 과정을 거친다. 이러한 과정을 반복하여 거치면서 각 기기는 여러 시행착오를 통해 최적의 정책을 분석하고 학습하게 된다.

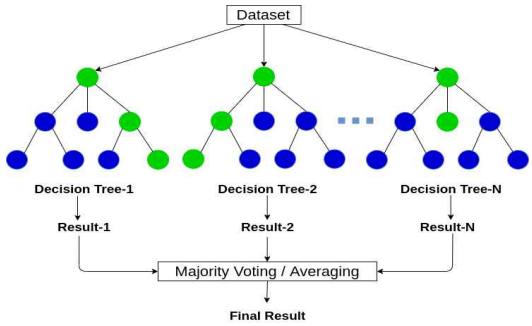


(그림 2) 강화학습 학습 과정

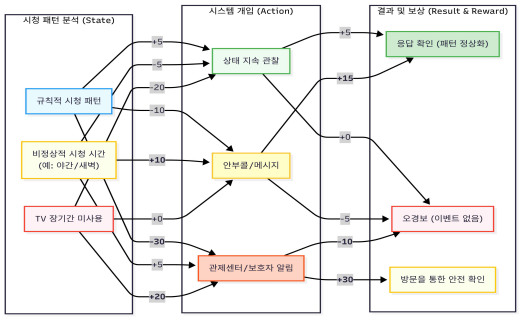
이 과정을 반복함으로써 IPTV 실시간 시청 데이터에서 의미 있는 패턴과 예측 모델을 효과적으로 도출할 수 있다. 연합학습 기반 AI 실시간 시청 데이터 분석 모델은 데이터 수집 과정의 비효율성을 제거하고, 시간과 비용의 제약을 최소화하며, 특히 정보 유출과 같은 보안 문제를 완화할 수 있다. 이 분석모델은 개인정보 보호에 민감한 가구에 뛰어난 효과를 발휘할 수 있고 사회적 취약계층이 사회적 접촉이 감소하면서 발생하는 정보 소외 및 고립, 지역사회 연결망 구성 문제에 대한 충분한 가능성을 보여준다.

4 분석 방법

본 분석에서 각 가정의 IPTV 단말기는 독립적으로 활동하여 실시간 시청 데이터를 분석하여 시청 패턴을 학습한다. 분류 예측 프로세스에서는 랜덤포레스트(Random Forest) 알고리즘을 이용하여 분석한다. 데이터를 기준으로 여러 방식의 질문을 진행하여 사용자의 행동에 대한 발생 가능한 환경의 시나리오를 구축하여 형태를 구성한다. 각각의 시나리오에 따른 예측 결과를 종합해 가장 많은 선택을 받은 결과를 최종 결론으로 도출하면서 최상의 예측 정확도를 보이는 앙상블 모델을 완성하게 된다.



(그림 3). Random Forest



(그림 4). 단말기 알고리즘의 예시

중앙 서버에서는 다수의 IPTV 단말기를 관리하고 실시간 모니터링, 통제하는 기능이 담겨있는 시스템을 구축한다. 각 기기별로 전송된 데이터는 주기별로 집계하고 글로벌 모델을 업데이트 하고, 새로운 글로벌 모델로 통합되며, 시스템의 모델 성능 향상을 주도한다. 또한 단말기에서 학습된 모델에 대한 성능지표를 실시간으로 모니터링하여 이상징후를 감지하거나 오작동에 대한 모니터링이 가능한 기능을 제공한다.

5 필요 데이터

이 분석에 대한 핵심데이터는 가정별 IPTV 실시간 시청기록이다. IPTV 고유 ID, 채널정보, 시청시간의 데이터를 수집하여 이에 맞게 특정 프로그램을 어떤 시간대에 시청하는지, 시청하는 시간대를 바탕으로 사용자의 시청 행동 패턴을 파악할 수 있는 IPTV 사용지표가 만들어진다. 이 지표를 바탕으로 강화학습을 진행한다. 수집된 데이터는 주기적으로 강화학습을 위한 데이터셋으로 변환되어 개인정보가 유출되지 않게 단말기 내에서 진행하고, 글로벌 모델 학습에 필요한 가중치(학습 패턴, 예측 규칙)를 중앙서버로 전송한다.

6 활용방안(계획)

본 분석에서는 IPTV 시청 프로그램 데이터에 대해 연합학습 및 강화학습을 적용하여 사회적 취약계층 고립예방 방법에 대해 제안하였다. 제안된 방법은 보안적 측면의 이점은 물론 기능적 측면에서의 강점을 활용하여 사회적 취약계층 복지의 취약점 해소에 효과적이며, 이외에도 가정의 특성에 대응하는 개인화 추천 시스템, 이용한 컨텐츠에 대한 분석에도 적용이 가능하다. 이 분석은 모든 데이터를 수집하는 비용과 시간을 큰 폭으로 줄이는 효과를 가지고 있고, 향후 디지털 기술의 발전과 함께 급증하는 분산 데이터의 활용의 문제를 해결하는 대안으로 더욱 주목받을 것이다. 더 나아가 모든 산업분야에서 혁신적인 솔루션으로 발전하여 여러 사회 문제에 더욱 핵심적인 역할을 수행할 것이라 판단된다.