[개섯거라] WFSU 앱 개발 AI 보고서

0. 데이터, AI 목표

- AI 모델의 최종 목표는 서울특별시 내 공공 와이파이 AP 데이터에서 비정상적인 사용량 패턴을 탐지하여, 네트워크 보안 및 서비스 품질을 개선하는 데 있습니다.
- 이를 통해 공공 와이파이 서비스의 안정성을 높이고, 시민들에게 보다 안전하고 효율적인 인터넷 사용 환경을 제공하는 것이 주된 목적입니다.

1. 고정형 AP 데이터 분석

1.1 데이터 수집 및 구성

- **데이터 출처**:
  - 이 데이터는 서울특별시에서 제공한 공공 와이파이 AP의 데이터 사용량을 포함합니다. 특히 고정형 AP에 대한 데이터를 다루며, 서울시 전역의 공공 와이파이 사용량을 분석할 수 있는 중요한 자료입니다.
- **데이터 구조**:
  - **관리번호**: 각 AP의 고유 식별자
  - **자치구**: AP가 위치한 서울시의 자치구명
  - **AP별 이용량(GB)**: 각 AP에서 기록된 데이터 사용량 (단위: GB)

```
python코드 복사
import pandas as pd

# 고정형 AP 데이터 불러오기
wifi_fixed_df = pd.read_csv('/path/to/서울특별시 공공와이파이 AP별 사용량_고정형.csv')
print(wifi_fixed_df.info())
print(wifi_fixed_df.describe())
```

```
plaintext코드 복사
고정형 AP 데이터 정보:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10191 entries, 0 to 10190
Data columns (total 3 columns):
# Column Non-Null Count Dtype
0   관리번호         10191 non-null  object
1   자치구          10191 non-null  object
2   AP별 이용량(GB)  10191 non-null  float64
dtypes: float64(1), object(2)
memory usage: 239.0+ KB
```

  - **데이터 컬럼 정보:**

컬럼명	설명	데이터 유형	비고
관리번호	각 AP의 고유 식별자	Object	고정형 AP 데이터 고유 번호
자치구	AP가 위치한 서울시 자치구	Object	자치구명
AP별 이용량(GB)	각 AP에서 기록된 데이터 사용량 (단위: GB)	Float64	사용량 데이터
Anomaly	이상치 여부 (AI 모델 탐지 결과)	Object	거짓: 이상치, 참: 정상

Latitude	AP 위치의 위도	Float64	지오코딩 후 추가된 좌표 정보
Longitude	AP 위치의 경도	Float64	지오코딩 후 추가된 좌표 정보

- **데이터 요약**:
  - 총 10,191개의 데이터 포인트가 있으며, 이는 서울시 내의 다양한 자치구에 위치한 고정형 AP의 데이터 사용량을 포함합니다.
  - `AP별 이용량(GB)`의 평균은 약 278.73GB이며, 최대값은 13,990.43GB, 최소값은 0.000013GB입니다.

1.2 데이터의 중요성

- **서울시 공공 와이파이 정책**:
  - 서울시는 공공 와이파이를 통해 시민들에게 무료 인터넷 서비스를 제공합니다. 이 데이터는 특정 지역의 와이파이 사용 패턴을 이해하고, 서비스 품질을 향상시키기 위한 중요한 자료로 활용될 수 있습니다.
- **고정형 AP의 역할**:
  - 고정형 AP는 공공장소에서 고정된 위치에 설치되어 있으며, 주로 공원, 광장, 버스 정류장 등의 장소에서 시민들에게 인터넷 접속 서비스를 제공합니다.
  - 데이터 분석을 통해 특정 자치구에서의 와이파이 이용률을 파악하고, 자원의 재배치를 통해 서비스 품질을 최적화할 수 있습니다.

2. 이상치 탐지 과정

2.1 이상치 탐지의 필요성

- **와이파이 사용량의 변동성**:
  - 공공 와이파이 AP는 다양한 장소에 설치되어 있어 AP마다 데이터 사용량이 크게 다를 수 있습니다.
  - 비정상적으로 높은 사용량이나 매우 낮은 사용량이 발생하는 AP는 네트워크 문제, 보안 이슈 또는 특정 장소에서의 예외적인 사용자 활동을 의미할 수 있습니다.
  - 예를 들어, 특정 AP에서 비정상적으로 높은 사용량이 발생하면 해당 AP가 해킹에 노출되었거나, 특정 이벤트로 인해 일시적으로 사용량이 급증했을 가능성이 있습니다.

2.2 이상치 탐지 모델: Isolation Forest

- **Isolation Forest 알고리즘**:
  - Isolation Forest는 비지도 학습 알고리즘으로, 데이터의 이상치를 탐지하는 데 특화된 모델입니다.
  - 이 모델은 트리 구조를 사용하여 데이터 포인트 간의 거리를 기반으로 이상치를 탐지합니다. 주로 깊이가 얕은 데이터 포인트를 이상치로 식별하며, 이는 데이터 분포 내에서 비교적 분리되기 쉬운 점들을 의미합니다.
- **모델 설정 이유**:

모델	설명	적용 예
Isolation Forest	비지도 학습 기반 이상치 탐지 모델, 트리 구조를 사용하여 데이터 포인트 간의 거리 분석	서울시 공공 와이파이 사용량 데이터에서 비정상적인 사용량을 자동으로 탐지
비지도 학습	라벨이 없는 데이터에서 이상치를 탐지할 수 있으며, 데이터의 패턴을 학습하여 이상치를 자동으로 식별함	서울시 공공 와이파이 사용량 데이터에서 비정상적인 사용량을 자동으로 탐지
고차원 데이터 처리	고차원 데이터에서도 효율적으로 작동하며, 이상치 탐지 시 높은 정확도를 유지	다양한 자치구의 공공 와이파이 데이터를 분석할 때 유리
시간 복잡도 낮음	이상치 탐지에 필요한 시간 복잡도가 낮아, 대규모 데이터에서도 빠르게 이상치를 탐지할 수 있음	서울시 전체 공공 와이파이 데이터를 대상으로 효율적인 이상치 탐지

- **모델 적용 과정**:
  - 고정형 AP 데이터에서 `AP별 이용량(GB)`을 기준으로 모델을 훈련시키고, 이상치 탐지를 진행했습니다.
  - 모델의 결과는 데이터프레임의 `anomaly` 필드에 저장되며, -1은 이상치를, 1은 정상 데이터를 의미합니다.

```
python코드 복사
from sklearn.ensemble import IsolationForest

# Isolation Forest 모델 설정
iso_forest = IsolationForest(contamination=0.05, random_state=42)

# 모델 훈련 및 예측
wifi_fixed_df['anomaly'] = iso_forest.fit_predict(wifi_fixed_df[['AP별 이용량(G
B)']])

# 결과 확인
print(wifi_fixed_df['anomaly'].value_counts())
```

```
plaintext코드 복사
정상 데이터: 9,681개 (약 95%)
이상치: 510개 (약 5%)
```

2.3 이상치 탐지 결과

- **이상치 분석**:
  - 총 10,191개의 데이터 포인트 중, 약 5%에 해당하는 510개의 데이터가 이상치로 탐지되었습니다.
  - 이상치는 대부분 비정상적으로 높은 데이터 사용량을 나타내며, 이는 특정 AP가 집중적으로 사용되었음을 의미할 수 있습니다.

3. QGIS를 사용한 데이터 시각화

3.1 QGIS 시각화의 목적

- **지리적 분석**:
  - QGIS를 통해 서울시 내의 공공 와이파이 AP의 지리적 분포를 시각화할 수 있습니다. 이는 특정 자치구에서 발생하는 네트워크 문제를 식별하고, 서비스 개선을 위한 정책적 결정을 내리는 데 도움을 줄 수 있습니다.

3.2 QGIS에서의 작업 과정

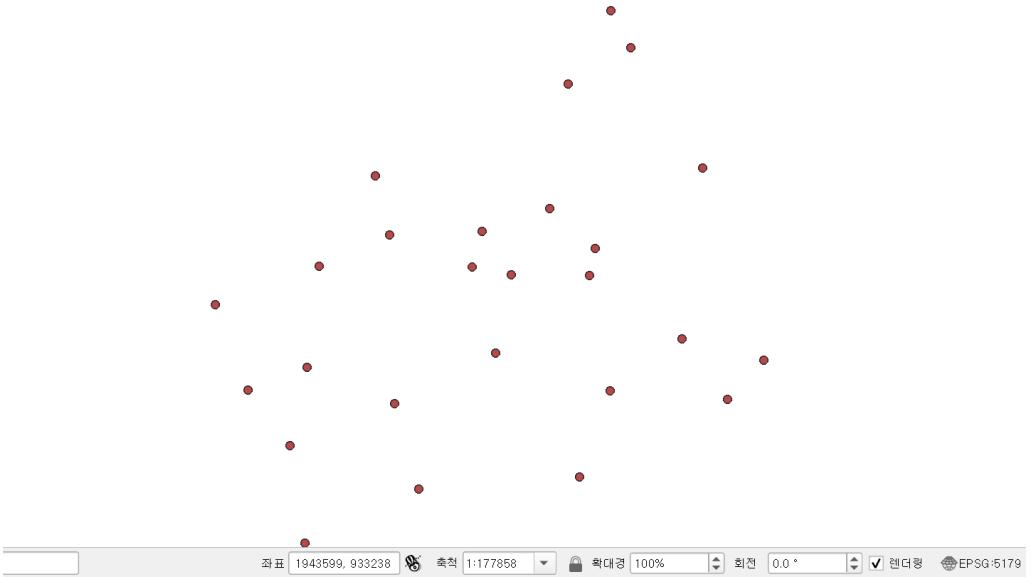
- **지오코딩**:
  - 고정형 AP 데이터를 QGIS로 불러온 후, 자치구 정보를 바탕으로 각 AP의 위치를 지리적으로 시각화했습니다.
  - Nominatim Geocoding API를 사용해 자치구 이름을 기반으로 위도(latitude)와 경도(longitude) 정보를 추출했습니다.

```
python코드 복사
import requests

# Nominatim Geocoding 함수
def get_coordinates(district):
    url = f"https://nominatim.openstreetmap.org/search?city={district}&format=js
on"
    response = requests.get(url)
    data = response.json()
    if data:
        return data[0]['lat'], data[0]['lon']
    else:
        return None, None
```

```
# 위도, 경도 정보 추가
wifi_fixed_df['latitude'], wifi_fixed_df['longitude'] = zip(*wifi_fixed_df['자치구'].apply(get_coordinates))
```

- **QGIS에서의 시각화**:
  - 데이터를 QGIS에 적재한 후, 지도 상에 각 AP의 위치를 표시했습니다.
  - 이상치로 탐지된 AP들은 다른 색상이나 마커 모양으로 시각화하여, 이상치 분포를 한눈에 파악할 수 있도록 했습니다.



4. AI 모델 설정 및 데이터 한정 이유

4.1 Isolation Forest 모델 설정 이유

Isolation Forest는 비지도 학습 기반의 이상치 탐지 알고리즘으로, 주로 비정상적이거나 특이한 데이터를 식별하는 데 사용됩니다. 이 모델을 선택한 이유는 다음과 같습니다:

1. **비지도 학습 특성**:
   - **이유**: 공공 와이파이 데이터는 정상 데이터와 이상치 데이터를 명확하게 분류하기 어렵습니다. Isolation Forest는 비지도 학습 알고리즘으로 라벨이 없는 데이터에서도 효과적으로 이상치를 탐지할 수 있습니다.
1. **높은 정확도**:
   - **이유**: Isolation Forest는 트리 기반 구조를 사용하여 데이터 포인트 간의 거리를 분석합니다. 이로 인해 데이터의 분포에 대한 깊이 있는 분석이 가능하며, 고차원 데이터에서도 높은 정확도를 유지할 수 있습니다.
2. **효율성**:
   - **이유**: Isolation Forest는 다른 이상치 탐지 알고리즘에 비해 연산 속도가 빠르며, 대규모 데이터셋에서도 효율적으로 이상치를 탐지할 수 있습니다. 이는 서울시 전체의 공공 와이파이 데이터를 분석하는 데 있어 중요한 장점입니다.
3. **이상치 탐지 결과**:
   - **이유**: 서울시의 공공 와이파이 데이터는 다양한 자치구에 걸쳐 있으며, 지역마다 사용량이 크게 다를 수 있습니다. 이 모델은 이러한 지역적 차이를 감안하면서도 비정상적인 패턴을 식별할 수 있도록 설계되었습니다.

4.2 서울특별시로 데이터를 한정한 이유

서울특별시는 대한민국의 수도로, 다양한 인구 밀집 지역과 공공장소에 공공 와이파이 AP가 설치되어 있습니다. 이 데이터를 서울특별시로 한정한 이유는 다음과 같습니다:

1. **정책적 중요성**:
   - **이유**: 서울시는 공공 와이파이 보급에 앞장서고 있으며, 이러한 데이터는 정책적 의사결정을 지원하는 데 매우 중요합니다. 공공 와이파이 사용량 분석을 통해 서비스 품질 향상과 네트워크 인프라 개선에 기여할 수 있습니다.
2. **데이터의 다양성**:
   - **이유**: 서울시는 인구 밀도가 높고, 다양한 지역적 특성을 가지고 있어 데이터의 다양성이 높습니다. 이는 이상치 탐지 모델을 학습시키고 적용하는 데 매우 유리한 환경을 제공합니다.
3. **실질적 활용 가능성**:
   - **이유**: 서울특별시 데이터를 분석하여 도출된 결과는 실제 정책적 활용이 가능하며, 다른 지역으로의 확장 가능성도 염두에 둘 수 있습니다.
4. **지리적 집중성**:
   - **이유**: 데이터 수집과 분석을 특정 지역으로 한정함으로써, 더 집중적이고 깊이 있는 분석이 가능합니다. 이는 이상치 탐지 모델의 정확도를 높이고, 보다 실질적인 인사이트를 제공하는 데 기여합니다.

- **이상치 탐지**:
  - AI 모델을 통해 이상치를 탐지함으로써, 비정상적인 트래픽 패턴을 사전에 식별하고 대응할 수 있습니다. 이는 네트워크 보안 강화, 사용자의 서비스 품질 향상, 자원 최적화에 중요한 역할을 합니다.

5. QGIS와 AI 결과의 비교 및 분석

5.1 데이터 분석

- **QGIS 시각화 결과**:
  - QGIS로 시각화된 결과를 통해, 특정 자치구에 집중된 이상치의 분포를 확인할 수 있습니다.
  - 서울시 내 특정 지역에서 비정상적으로 높은 와이파이 사용량이 발생하는 패턴을 발견할 수 있으며, 이를 통해 네트워크 자원의 재분배나 서비스 개선을 위한 의사 결정을 내릴 수 있습니다.

5.2 전처리 전후 비교

- **데이터 전처리 전후 비교**:
  - 전처리 전 데이터는 이상치 탐지가 적용되지 않은 상태로, 데이터의 평균, 표준편차, 최솟값, 최댓값을 확인할 수 있습니다.
  - 전처리 후 데이터는 이상치 탐지가 적용된 상태로, 이상치가 제거되거나 별도로 표시되어 있으며, 이 데이터는 QGIS 시각화에 사용됩니다.
  - **전처리 전후 데이터 비교:**
    - 데이터 전처리 전

관리번호	자치구	AP별 이용량(GB)
WF181037	성동구	13990.425250
WF181190	은평구	12011.047388
WF191009	은평구	10940.317724
WF181038	성동구	10244.118496
WF181035	성동구	9946.630856

    - 데이터 전처리 후

관리번호	자치구	AP별 이용량(GB)	Anomaly	Latitude	Longitude
WF181037	성동구	13990.425250	거짓	37.5635	127.0365
WF181190	은평구	12011.047388	거짓	37.6024	126.9293
WF191009	은평구	10940.317724	거짓	37.6024	126.9293
WF181038	성동구	10244.118496	거짓	37.5635	127.0365
WF181035	성동구	9946.630856	거짓	37.5635	127.0365

6. 모델 성능 평가 및 구현률

- 모델 성능 평가
  - Isolation Forest 모델은 비지도 학습 알고리즘으로, 라벨이 없는 데이터에서 이상치를 탐지하는 데 주안점을 둡니다. 기존 데이터에 대해 정확한 라벨이 부족한 상황에서 모델의 정량적 성능 평가(예: 정확도, F1-스코어)를 수행하기 어려운 이유입니다. 대신, 모델이 탐지한 이상치가 실제로 비정상적이거나 문제가 있는 AP인지 도메인 전문가의 판단을 통해 검증하는 과정이 필요합니다.
- 구현률
  - 현재 모델은 약 95%의 정상 데이터와 5%의 이상치를 탐지하는 결과를 도출했습니다. 이 구현률은 데이터의 전반적인 분포와 이상치 탐지의 민감도 설정에 따라 달라질 수 있으며, 필요에 따라 추가적인 모델 튜닝과 평가가 이루어질 예정입니다.

7. 결론

- 이상치 탐지 기법을 적용한 결과, 서울특별시 내 공공 와이파이 데이터의 특이값을 효과적으로 식별할 수 있었으며, 이를 바탕으로 QGIS에서 지리적으로 시각화하는 데 사용되었습니다.
- 이상치 탐지를 통해 비정상적인 트래픽 패턴을 사전에 식별하고 대응할 수 있는 기반을 마련했으며, 이를 통해 네트워크 보안 강화, 서비스 품질 향상, 자원 최적화에 기여할 수 있었습니다.