표 형식 데이터에 대한 딥러닝을 활용한 반려견 질병 예측 및 예방

팀 명 : esc

지도교수 : 송길태

부산대학교 전기컴퓨터공학부 정보컴퓨터전공

201824604 최지광

201924493 서진욱

201924496 송민재

목차

1. 연구 배경 및 목표

1.1 연구 배경

저출산 및 고령화 시대적 요인과 1인 가구의 증가로 반려동물을 키우는 가구가 급증하고 있으며, 이에 따른 시장이 급격히 성장하고 있다. 국내 반려동물 관련 물품 및 서비스 관련 시장이 2019년 3조에서 2024년 4.9조로 연평균 11%씩 증가하였고 반려동물 관련 시장이 증가함에 따라 반려동물과 관련된 물품, 돌봄 서비스, 등록 관리 등의 서비스가 중점적으로 발전했다. 하지만, 급증하는 반려동물 가구 수에 비해 반려동물과 관련된 서비스 수는 아직 많이 부족한 상황이다. 무엇보다 반려동물의 진료비 부담이 가장 큰 문제라고 할 수 있다. 동물병원 이용 경험이 있는 설문 참여자에게 진료비가 부담으로 느껴지는지에 대한 조사를 한 결과 8~90%대가 '부담된다'고 응답을 하였다. 그래서 우리는 반려견 질병을 예측할 수 있는 딥러닝 모델을 통해 질병을 예방해 굳이 동물병원에 가지 않고도 반려견이 어떤 질병을 가지고 있으며 그 질병에 대해 어떻게 대응해야 하는지 알려주는 서비스를 만들고자 한다.

1.2 연구 목표

우리는 반려견의 질병을 예측할 수 있는 모델들을 비교해 어떤 모델이 어떤 상황에서 가장 최적의 모델이 되는지를 선정하고 이를 시각화할 것이다. 반려견에 대한 기본 데이터(품종, 나이, 성별 등)와 사양관리 데이터 (생활환경, 배변 상태, 하루 식이 횟수 등) 를 입력받아 CRP, IgG, IL-6, AFP와 같은 의학 정보 데이터 및 질병 유무를 알아내고 의학 정보 데이터의 수치를 분석해 수치가 비정상적일 경우 이를 어떻게 예방하고 수치를 안정화할 수 있는지에 대해 알려줄 것이다.

2.환경 및 유의사항

2.1 개발 환경

소프트웨어 요구사항

- 프로그래밍 언어 : Python 3.8 이상
- 사용 모델 : LightGBM, FT-Transformer , TabNet
- 딥러닝 프레임워크 : Tensorflow 2.x 또는 PyTorch 2.0 이상
- 데이터 처리 : Pandas, NumPy
- 데이터 베이스 : MySQL

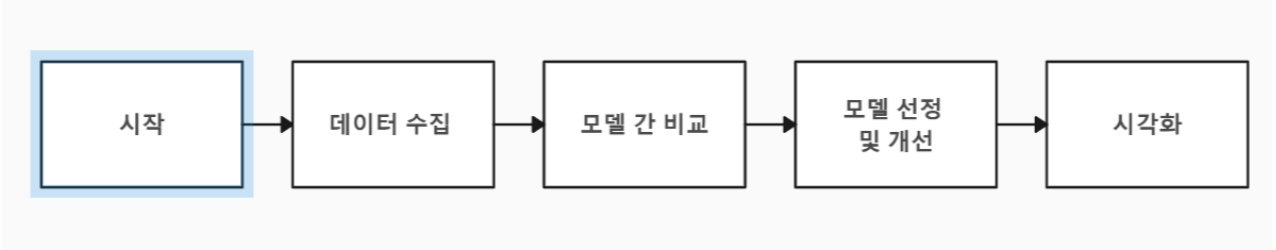
개발 도구

- 개발 환경 : Google Colab
- 버전 관리 : GitHub

2.2 유의사항

- 과적합, 해석 가능성 부족 등 표 형식 데이터에 대한 딥러닝의 문제점을 해결해야 한다.
- 데이터의 특성이 다양하고, 출력 또한 여러 형태로 나타나기 때문에 각 특성과 출력에 적합한 딥러닝 모델을 선정해야 한다.
- Accuracy, Recall, Precision, F1 Score 등 다양한 성능 지표를 통해 모델의 성능 평가를 진행한다.

3. 진행 과정



3.1 데이터 수집

반려견에 대한 다양한 데이터와 질병 데이터를 동시에 가지고 있는 표(CSV, JSON)를 찾는 것을 목표로 하여 데이터를 수집하였다. 또한 데이터의 row가 많은 대용량 데이터를 찾는 것을 목표로 하였다.

찾은 데이터는 반려견의 나이, 성별과 같은 기본적인 데이터, 체중, 크기 등과 같은 신체적 특징에 대한 데이터, 배변 상태, 하루 식이 횟수와 같은 사양관리 데이터, 그리고 질병에 대한 데이터를 JSON 형식으로 저장하고 있다. 데이터의 구성에 대한 더 자세한 내용은 참고 자료 1번의 데이터 통계, 데이터 구조 탭에서 확인할 수 있다.

3.2 모델 간 비교

표 형식 데이터에 대한 딥러닝을 위해 적합하다고 생각하는 모델들을 후보로 선택하였으며 추후 추가해 나갈 예정이다. 아래는 현재까지 선택한 모델들이다.

- LightGBM : 빠른 학습 및 예측 속도와 높은 정확도를 제공하는 모델이기에, 대규모 데이터 세트에 적합하다. 파라미터 튜닝이 용이해 다양한 상황에 쉽게 적용할 수 있다.
- FT-Transformer : 특성 간의 상호작용이 중요한 데이터에서 좋은 성능을 발휘하고, 사용 데이터 세트가 특성이 많기에 적합하다 판단했다.
- TabNet : 각 특성의 중요도를 명확히 판단할 수 있기에, 과적합을 방지하고 높은 성능을 제공한다. 모델의 결정 과정이 중요하기에 사용하기로 결정했다.

3.3 모델 선정 및 개선

평가 기준을 토대로 목표 수치를 넘긴 모델들을 선정한 후 parameter 변경, missing data 처리, 등과 같은 방식을 통해 모델의 평가 수치를 개선한 후 최종적으로 제일 좋은 성능을 보여주는 모델을 선정할 것이다.

3.4 프로그램 개발 및 시각화

웹을 통해 기본 데이터, 신체 계측 데이터, 사양 관리 데이터를 입력받은 후 이를 토대로 질병 유무, 의학 정보 데이터, 비만도 등을 예측하여 출력할 것이다.
각 의학 정보 데이터에 대해 특정 질병에 대해 위험도를 안전/예방 필요/위험으로 분류하여 사용자에게 출력할 것이다.

4. 일정 및 역할 분담

4.1 개발 일정

<table>
<tr><th colspan="2">5월</th><th colspan="4">6월</th><th colspan="4">7월</th><th colspan="4">8월</th><th colspan="4">9월</th></tr>
<tr><th>3</th><th>4</th><th>1</th><th>2</th><th>3</th><th>4</th><th>1</th><th>2</th><th>3</th><th>4</th><th>1</th><th>2</th><th>3</th><th>4</th><th>1</th><th>2</th><th>3</th><th>4</th></tr>
<tr><td colspan="2">주제 선정, 분석 , 착수 보고서 작성</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="4">데이터 전처리, 모델 분석 및 개발</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">모델 보완 및 개선, 중간 보고서 작성</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">데이터 분석 보충, 기능 점검</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">최종 보고서 작성, 발표 준비</td></tr>
</table>

4.2 역할 분담

이름	역할
최지광	딥러닝 모델 학습 및 성능 평가
서진욱	딥러닝 모델 학습 및 성능 평가
송민재	데이터 수집 및 전처리,시각화
공통	모델 개선, 보고서 작성, 발표 준비

5. 참고 자료

- 사용 데이터 : https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71520
- https://www.lgresearch.ai/blog/view?seq=387
- https://lightgbm.readthedocs.io/en/stable/