

기계학습 알고리즘을 이용한 보행만족도 예측모형 개발^{*,**}

Developing a Pedestrian Satisfaction Prediction Model Based on Machine Learning Algorithms

이제승^{***} · 이현희^{****}

Lee, Jae Seung · Lee, Hyunhee

Abstract

In order to develop pedestrian navigation service that provides optimal pedestrian routes based on pedestrian satisfaction levels, it is required to develop a prediction model that can estimate a pedestrian's satisfaction level given a certain condition. Thus, the aim of the present study is to develop a pedestrian satisfaction prediction model based on three machine learning algorithms: Logistic Regression, Random Forest, and Artificial Neural Network models. The 2009, 2012, 2013, 2014, and 2015 Pedestrian Satisfaction Survey Data in Seoul, Korea are used to train and test the machine learning models. As a result, the Random Forest model shows the best prediction performance among the three (Accuracy: 0.798, Recall: 0.906, Precision: 0.842, F1 Score: 0.873, AUC: 0.795). The performance of Artificial Neural Network is the second (Accuracy: 0.773, Recall: 0.917, Precision: 0.811, F1 Score: 0.868, AUC: 0.738) and Logistic Regression model's performance follows the second (Accuracy: 0.764, Recall: 1.000, Precision: 0.764, F1 Score: 0.868, AUC: 0.575). The precision score of the Random Forest model implies that approximately 84.2% of pedestrians may be satisfied if they walk the areas, suggested by the Random Forest model.

키워드 보행만족도, 기계학습, 로지스틱 모형, 랜덤 포레스트, 인공신경망

Keywords Pedestrian Satisfaction, Machine Learning, Logistic Regression, Random Forest, Artificial Neural Network

1. 서론

1. 연구의 배경 및 목적

현재 우리나라의 고령화 문제가 심화되면서 고령 보행자의 안전과 편리성에 대한 요구가 높아지고 있지만, 자동차 통행의 효율성 위주로 개발된 우리나라의 많은 도시에서 보행자의 이동성과 안전성에 대한 배려가 미흡한 실정이다. 차량이 고속으로 달리는

간선도로에서는 보행자의 안전한 도로 횡단 문제가 대두되고 있고, 이면도로에서는 자동차와 보행자의 혼재로 보행 편리성과 안전성이 위협받고 있다. 자동차를 중심으로 형성된 도시환경은 특히 보행능력이 부족한 고령자와 장애인 등 교통약자가 안전하고 편리하게 보행하기에 불리하다. 최근 경제가 성장함에 따라 사회적 약자들을 위한 복지에 관한 관심이 늘어나고, 그 결과 점차 교통약자들의 보행환경 개선을 위한 노력이 이루어지고 있다. 이에 따라 저상버스 도입, 지하철 엘리베이터 설치, 장애인 콜택시 확

* 이 논문은 2017년도 한국연구재단 전략과제 연구비 지원으로 연구되었음(NRF-2017R1E1A1A01074422).

** 이 논문은 2019년도 홍익대학교 학술연구진흥비에 의하여 지원되었음.

*** Assistant professor, Hongik University (jaeseung74@gmail.com)

**** Professor, Gachon University (corresponding author: hhlee@gachon.ac.kr)

보 통 교통수단에 대한 개선뿐만 아니라 볼라드 및 점자블록 정비, 보도 턱 낮춤 등 보행환경 개선 사업이 꾸준히 이루어지고 있다.

한편 교통약자를 위한 물리적 보행환경은 점차 개선되고 있으나, 교통약자들이 보행환경에 대한 정보를 얻고 공유할 수 있는 환경은 여전히 부족한 실정이다. 교통약자는 일반인보다 보행환경의 영향을 많이 받으며 보행환경이 취약한 곳에서는 위험에 노출될 확률이 높다. 따라서 교통약자가 이동에 불편함을 적게 느낄 수 있는 안전하고 편안한 보행환경을 갖춘 보행 동선 정보를 아는 것은 교통약자의 보행 활동 증진에 기여할 수 있을 것이다.

보행자 경로 추천 기능은 현재 온라인 지도와 내비게이션 서비스에서 제공되고 있지만, 최단 거리 위주로 경로를 제시하고 있어 인도 폭, 경사, 엘리베이터 등 보행자의 보행만족도에 영향을 미치는 요인들을 고려한 보행 경로를 제공하지 못하고 있다. 따라서 보행만족도에 최적화된 보행 경로 서비스가 제공된다면 일반 보행자뿐 아니라 고령자 등 교통약자의 보행 활동에 증진에 크게 기여할 것으로 기대된다.

본 연구의 목적은 보행만족도 기반의 보행 경로 서비스 개발을 위한, 기계학습 알고리즘 기반 보행만족도 예측모형 개발에 있다. 현재까지 몇몇 연구들에서 서울시 유동인구 조사에 포함된 보행만족도 자료를 이용하여 보행만족도에 영향을 미치는 요인을 분석하였다(이수기 외, 2014; Kim et al., 2014). 이 연구들은 전통적인 통계 기법을 이용하여 보행만족도에 영향을 미치는 요인들을 분석하였지만, 보행만족도 기반의 보행 경로 서비스 제공을 위해서는 특정한 보행환경에서 보행자가 느끼는 만족도를 예측하여야 서비스 이용자에게 적절한 경로를 제시할 수 있다.

따라서 본 연구에서는 예측성능 최적화에 적합한 기계학습 알고리즘을 이용하여 보행만족도 예측모형을 개발하고자 한다. 이를 위해 서울시 유동인구 조사자료를 이용하여 로지스틱, 랜덤 포레스트, 인공신경망 알고리즘 기반의 보행만족도 모형을 개발하고, 이 모형들의 예측성능을 비교하여 최적의 모형을 도출할 것이다.

더불어 보행만족도는 동일 환경에 대해서 신체 연령에 따라 다르게 평가될 수 있음을 고려하였다. 예를 들어 같은 경사도라도 젊은 보행자와 고령 보행자가 느끼는 부담감의 정도가 다를 수 있다. 따라서 교통약자인 60세 이상 고령 보행자를 다른 연령대 보행자와 분리하여 보행만족도 모형을 구축하고, 이 두 보행자 집단의 보행만족도에 영향을 미치는 요인의 중요도를 비교 분석하고자 한다.

II. 선행연구 고찰

1. 물리적 환경과 보행만족도

보행만족도는 명확하게 정의하기 쉽지 않은 모호한 개념이다.

일반적으로 주거만족도는 주거환경을 경험하면서 느끼는 긍정적인 정서적·심리적 상태를 의미한다(Amerigo and Aragones, 1997). 그러므로 보행만족도는 보행자가 보행환경을 경험하면서 느끼는 긍정적인 정서적·심리적 상태로 정의할 수 있다. 따라서 보행만족도는 보행 활동을 증진할 수 있는 환경요소와 밀접한 상관관계를 지니고 있다(Kim et al., 2014).

보행만족도에 영향을 미치는 환경요인에 관한 연구는 물리적 환경요인과 보행자 설문조사를 통해 보행만족도 수준 간의 상관관계를 밝히는 정량적 분석을 중심으로 이루어졌다. 일레로 지우석 외(2008)는 수원시 팔달구와 영통구의 보행자를 대상으로 설문 조사를 시행하여 보도 유무, 보도 노후화, 경사도 등이 보행만족도에 미치는 영향을 밝혔다. Kim et al.(2014)은 환경변수를 근린 규모 변수(건물 밀도, 교차로 밀도, 경사도 등)와 가로환경 변수(인도 넓이, 가로수, 가로등, 횡단보도 등)로 분류하고, 이 변수들이 서울시 유동인구 조사 자료의 보행만족도 수준에 미치는 영향을 다수준 순서형 로지스틱 모형을 이용하여 분석하였다. 김창국 외(2016)의 연구는 보행만족도에 보도 폭, 차로 수, 보행특화 거리 사업, 용적률, 건폐율, 건축물의 용도 등이 영향을 미치는 것을 확인하였고, 통근통학 목적의 보행만족도가 다른 목적보다 가로 환경 특성의 영향을 많이 받는 것으로 나타났다.

이수기 외(2014)는 보행환경이 보행자에 미치는 영향을 연령대별로 분석하여 물리적 환경의 영향이 연령대별로 차이가 있음을 밝혔다. 이 연구에 따르면 60대는 보행특화거리에 대한 만족도가 다른 연령대에 비해 높았다. 또한, 젊은 층의 보행만족도는 보행량이 많은 지점에서 높은 경향이 있지만 60세 이상 보행자의 만족도는 보행량이 많은 지점에서 낮아지는 경향이 있었다. 공원의 유무도 60세 이상 보행자에게 더 강한 영향을 주는 것으로 나타났다. 고령 보행자 행태와 보행만족도를 분석한 해외의 연구는 보행 안전 정도가 고령 보행자의 보행만족도와 관련이 있음을 밝혔다(Methorst and Horst, 2010).

다른 연구들은 객관적인 물리적 환경보다는 보행자의 주관적인 환경 인식에 집중하였다. 이러한 연구들은 보행환경에 대한 인지가 보행만족도에 영향을 미치며, 보행 경로 선택의 중요한 고려사항 중 하나로 분석하였다(박소현 외, 2009; 이인성, 2000). 이인성·김현옥(1998)은 보행자가 느끼는 경로에 대한 익숙도가 보행만족도에 가장 큰 영향을 미친다는 결론을 도출하였다. 박소현 외(2009)의 서울시 종로구 가회동을 중심으로 한 설문조사 분석 연구는 가로의 쾌적성이 보행만족도에 가장 큰 영향을 미친다는 결과를 도출하였다. 강남역 보행자를 대상으로 한 임하나 외(2014)의 연구에 따르면 보행자들은 주변 환경이 쾌적하고 볼거리가 많을수록 체감거리를 짧게 느끼는 것으로 나타났다.

Wang et al.(2012)은 보행로의 조화성과 개방성과 같은 주관적인 인지 특성이 객관적인 물리적 환경보다 보행만족도에 더 큰 영향을 미친다고 주장하였다. 또한, 이수기 외(2016)는 서울서베이

2013년 자료를 활용하여 행정동 단위의 근린환경 특성이 보행만족도에 미치는 영향을 분석하여 객관적으로 도출한 환경변수보다 범죄, 소음, 대기오염, 쓰레기 등 가로환경에 대한 주관적인 인식변수의 영향이 크다는 결론을 도출하였다.

성현곤 외(2011)는 구조방정식 모형을 이용하여 보행만족도에 영향을 미치는 안전성, 연속성, 쾌적성, 편리성/시인성, 생동성의 5개 잠재변수를 총 31개의 측정지표를 이용하여 추출하였다. 이 잠재요인 수준을 지역별로 비교하여 계획 가로인 강남의 보행만족도 수준이 종로에 비해 높지만, 개별 잠재요인 특성의 지역적 차이를 고려한 계획이 필요하다는 결론을 도출하였다.

2. 기계학습 모형과 도시환경

최근 들어 도시계획, 부동산, 교통 분야에서도 딥러닝 모형을 포함한 기계학습 모형이 이용되기 시작하였다. 심재현 외(2007)는 인공지능망 모형을 이용하여 산업용지 공급가격을 추정하였다. 그러나 사용된 관찰 값이 302개에 불과해, 모형이 충분히 학습하기에는 부족한 것으로 판단된다.

배성완·유정석(2018)은 전통적인 시계열 분석 기법과 기계학습 모형의 부동산 가격지수 예측성능을 비교·분석하였다. 전통적인 시계열 분석 기법으로는 단변량 시계열 분석 기법인 ARIMA(Auto Regressive Integrated Moving Average), 다변량 시계열 분석 기법인 VAR(Vector Auto Regressive) 모형, 베이지언 VAR 모형을 사용하였고, 기계학습 모형으로는 서포트 벡터 머신(SVM, Support Vector Machine), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 회귀 트리(Gradient Boosting Regression Tree), 심층신경망(Deep Neural Network), LSTM(Long Short Term Memory Network)을 사용하였다. 연구 결과, 시장 상황이 일정한 추세를 보이는 경우 시계열 모형과 기계학습 모형 모두 의미 있는 예측성능을 보여주지만, 시장이 비선형적으로 급변하는 경우에는 전통적인 시계열 모형보다 기계학습 모형이 우수한 예측성능을 보였다.

하재현·이수기(2017)는 수도권 가구통행실태조사자료를 이용하여 주거지선택과 통근통행패턴을 예측하는 다층 퍼셉트론 딥러닝 모형 구축하였다. 횡단면적 연구 결과, 개인이 주거하는 시군구 지역의 90%까지 예측성능이 확보되었지만, 과거 데이터로 학습한 모형의 이후 시점의 개인 주거지의 예측성능은 최대 25.7%였다. 또한, 예측한 주거지를 바탕으로 구축한 통근통행 기종점 자료와 실제 기종점 통행량과 비교한 결과 최대 0.76 사이의 설명력을 보였다.

이호준·이수기(2018)는 이미지 분석을 위한 합성곱 신경망(CNN, Convolutional Neural Network) 모형을 활용하여 Google Street View API에서 획득한 가로 이미지에서 생활도로에 노상 주차된 차량을 식별하는 모형을 구축하였다. 이 모형은

노상 주차된 차량에 대하여 약 82.07%의 객체 탐색 정확도를 나타내었다.

박근덕(2018)은 서울시 유동인구 조사 데이터와 딥러닝 모형을 이용하여 가로 이미지와 보행만족도 간의 상관관계를 분석하였다. Google Street View API를 통해 획득한 서울시 유동인구 속성조사 조사지점 중 2,450개 지점의 가로 이미지 데이터로 사전학습된 CNN 모형들을 전이 학습(transfer learning)하여 보행만족도 예측모형을 구축하였다. 예측성능은 최대 상관계수 0.84 수준으로 평가되었다.

3. 본 연구의 차별성

이제까지 보행 관련 연구는 보행량을 중심으로 많이 이루어졌지만, 보행만족도에 관한 연구도 상당 부분 이루어졌다. 보행만족도 연구는 대부분 보행만족도에 영향을 미치는 요인의 통계적 유의성을 검증하는 데 집중하였다. 그러나 보행만족도 기반의 보행 경로 서비스 제공을 위해서는 특정한 보행환경에서 보행자가 느끼는 만족도를 예측하는 모형이 필요하다.

이러한 예측성능을 최적화하기 위해서 최근 기계학습 기법이 도시계획 분야에서도 활용되고 있다. 특히 박근덕(2018)의 연구는 딥러닝 모형을 활용하여 2,450개 지점의 가로 이미지와 보행만족도의 상관관계를 분석하였다.

본 연구의 차별성은 기계학습 모형의 과적합을 방지하고 충분한 예측성능을 확보하기 위해 10만 건에 달하는 서울시 유동인구 조사의 보행만족도 설문자료를 모두 이용하여 예측모형을 구축하는 데 있다. 또한, 다양한 기계학습 모형을 학습시킨 후 각 모형의 예측성능을 비교하여 최적의 보행만족도 예측모형을 개발함으로써, 단일 보행만족도 모형을 사용한 선행연구와 차별성을 가진다.

III. 분석자료 및 분석 방법

1. 분석자료 및 변수 구성

본 연구의 공간적 범위는 서울특별시이며, 분석자료는 서울에서 2009년, 2012년, 2013년, 2014년, 2015년에 실시한 유동인구 조사를 통해 수집된 보행만족도 자료를 이용하였다. 유동인구 조사의 보행만족도 데이터는 서울시의 1,000개 속성조사 지점에서 보행자 설문 조사를 통해 수집되었다(그림 1).

현재 서울시 열린데이터광장에 공개된 서울시 유동인구 조사 자료에는 연도별로 2만 건의 보행만족도 설문 조사 표본이 포함되어 있으며, 5년간 총 표본 수는 총 10만 개이다. 이 중 결측치나 오류를 제거하고 총 99,589개의 표본을 사용하였다.

설문 조사 문항에는 응답자의 보행특성과 개인특성에 관한 자



Figure 1. Survey locations in Seoul

Source: Seoul Floating Population Report 2015

료가 포함되어 있다. 또한, 유동인구 조사의 속성조사 자료에는 조사지점의 토지이용, 가로특성, 가로시설, 대중교통 관련 자료가 포함되어 있다. 유동인구 조사자료에서 추출한 주요 변수의 정의와 기술통계는 <표 1>에 제시하였다.

유동인구 조사의 보행만족도 변수는 “매우 불만족”에서 “매우 만족”까지의 5점 리커트 척도로 측정되어 있다. 본 연구에서는 보행만족도 변수를 변형하여 “보통”, “만족”, “매우 만족”을 “만족”으로 하고, “매우 불만족”과 “불만족”을 “불만족”으로 하는 더미 변수를 생성하여 종속 변수로 사용하였다. 이는 본 연구의 목표인 보행자를 위한 보행 경로 추천 모형을 제공하는 데 있어, 최소한 이용자가 불만족하지 않으리라고 예측되는 경로를 추출하고자 하기 때문이다. 보행만족도 더미 변수 구축 결과 약 76.7%의 보행자의 만족도가 보통 또는 만족으로 나타났다.

보행자 단위의 설명변수로는 보행목적, 보행빈도, 성별, 연령, 직업을 포함하였다. 설문 조사지점 단위 설명변수로는 토지이용, 대중교통, 가로 유형 및 가로설계요소를 포함하였다. 더불어 설문조사가 5년에 걸쳐 이루어졌기 때문에 설문 조사 연도별 차이를 고려하기 위한 연도 더미 변수를 추가하였다. 마지막으로 관찰되지 않은 구별 특성을 통제하기 위한 더미 변수를 포함하였다.

기계학습 모형에 사용되는 변수는 모형의 성능을 향상하기 위해 최솟값 0과 최댓값을 1로 변환하여 정규화하거나, 평균 0, 표준편차 1로 표준화하여 사용하는 경우가 많다. 본 연구에서 사용된 설명변수 대부분은 0과 1의 값을 가지는 더미 변수이지만, 경사도, 보도 넓이, 차선 수 변수는 연속변수이다. 본 연구에서는 이 연속변수들(x)을 다음의 공식을 이용하여 최솟값 0과 최댓값을 1을 가지는 새로운 변수(x')로 표준화하여 사용하였다. 이렇게 연속변수를 정규화하는 다른 이유는 변수 간의 중요도를 평가하기 위하여 모든 설=명변수를 동일한 스케일(0~1)로 만드는

데 있다.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. 기계학습 기법

기계학습(machine learning)은 인공지능의 한 분야로서, 대량의 데이터를 학습한 모형을 이용하여 판단이나 예측 분석을 수행하고 지식을 추출하는 작업이다(이요섭·문필주, 2017). 기계학습 중 명시적 정답이 주어진 데이터를 학습하고 미래에 주어진 데이터에 관한 예측을 수행하는 방법을 지도 학습(supervised learning)이라고 한다. 지도 학습은 예측값이 범주형 변수인 분류(classification) 또는 예측값이 연속변수인 회귀(regression)분석에 이용된다(Muller and Guido, 2016).

지도 학습에서는 훈련 데이터로 학습한 모형이 정답을 주지 않은 새로운 데이터의 답을 정확히 예측할 경우 일반화(generalization)되었다고 한다. 그러나, 훈련 데이터에 과도하게 학습된 복잡한 모형의 경우에는 훈련 데이터에만 적합한 모형이 되고 새로운 데이터를 이용한 예측성능은 낮아질 수 있다. 이렇게 너무 복잡한 모형이 새로운 데이터에 일반화하지 못하는 현상을 과적합(overfitting)이라고 한다. 보통 데이터셋의 크기가 커질수록 과적합 없이 복잡한 모형을 만들 수 있다(Muller and Guido, 2016).

또한, 기계학습 모형들은 대부분 학습 전에 미리 지정해야 하는 하이퍼파라미터(hyperparameter)를 가지고 있다. 하이퍼파라미터는 모형의 복잡성에 영향을 미치며, 데이터 학습으로 완성된 모형의 계수를 뜻하는 모델 파라미터와 달리 연구자에 의해 미리 지정되어야 한다. 따라서, 서로 다른 하이퍼파라미터 설정으로 만들어진 여러 가지의 모형 중에 가장 예측성능이 우수한 모형을 선택할 수 있다. 이러한 과정을 모형선택(model selection)이라 한다(Muller and Guido, 2016).

지도 학습의 대표적 학습 알고리즘은 릿지 회귀(Ridge Regression), 라쏘 회귀(Lasso Regression), 로지스틱 회귀(Logistic Regression), 서포트 벡터 머신(Support Vector Machine), 의사결정나무(Decision Tree), 랜덤 포레스트(Random Forest), 인공신경망(Artificial Neural Network), 등이 있다(Muller and Guido, 2016). 본 연구에서는 지도 학습 방식의 기계학습 중에서 분류모형 중 가장 기본적인 로지스틱회귀 모형과 기초분석을 통해 유동인구 데이터 분류에 우수한 성능을 나타낸 랜덤 포레스트와 인공신경망 모형을 이용하였다.

1) 로지스틱 회귀

로지스틱 회귀는 데이터셋에 있는 특성 변수마다 하나의 가중치 계수(weight coefficient)와 편향(bias)과 같은 모델 파라미터

Table 1. Definition of variables and descriptive statistics (n = 99,589)

Variables	Definitions	Mean	S.D.	Min.	Max.	
Pedestrian satisfaction	5-point Likert scale	1. Very unsatisfied; 2. Little unsatisfied; 3. Normal; 4. Little satisfied; 5. Very satisfied	3.073	0.982	1	
	Dummy	1. Satisfied or normal; 0. Unsatisfied	0.767	-	0	
Trip purpose	Commute	1. yes; 0. no	0.231	-	0	
	Work	1. yes; 0. no	0.157	-	0	
	Transfer	1. yes; 0. no	0.065	-	0	
	Errand	1. yes; 0. no	0.463	-	0	
	Leisure	1. yes; 0. no	0.083	-	0	
Trip Frequency	First time	1. yes; 0. no	0.044	-	0	
	Rarely	1. yes; 0. no	0.142	-	0	
	Often	1. yes; 0. no	0.484	-	0	
	Everyday	1. yes; 0. no	0.330	-	0	
Personal Characteristics	Gender	Male	1. yes; 0. no	0.474	-	0
		Female	1. yes; 0. no	0.526	-	0
	Age group	15~19 years old	1. yes; 0. no	0.057	-	0
		20~29 years old	1. yes; 0. no	0.198	-	0
		30~39 years old	1. yes; 0. no	0.181	-	0
		40~49 years old	1. yes; 0. no	0.179	-	0
		50~59 years old	1. yes; 0. no	0.223	-	0
		60 years and older	1. yes; 0. no	0.161	-	0
	Job	White collar	1. yes; 0. no	0.378	-	0
		Blue collar	1. yes; 0. no	0.076	-	0
Student		1. yes; 0. no	0.150	-	0	
Housewife		1. yes; 0. no	0.211	-	0	
Land Use	Type 1 residential	1. yes; 0. no	0.055	-	0	
	Type 2 residential	1. yes; 0. no	0.286	-	0	
	Type 3 residential	1. yes; 0. no	0.337	-	0	
	Semi-residential	1. yes; 0. no	0.058	-	0	
	Semi-industrial	1. yes; 0. no	0.045	-	0	
	Green	1. yes; 0. no	0.020	-	0	
	Commercial	1. yes; 0. no	0.200	-	0	
Public transit	Bus stop	1. yes; 0. no	0.428	-	0	
	Subway station	1. yes; 0. no	0.270	-	0	
Street type	Pedestrian-dedicated	1. yes; 0. no	0.715	-	0	
	Pedestrian sidewalk	1. yes; 0. no	0.188	-	0	
	Pedestrian/vehicle	1. yes; 0. no	0.087	-	0	
	Pedestrian/bicycle	1. yes; 0. no	0.011	-	0	

다음 페이지에 계속

Variables	Definitions	Mean	S.D.	Min.	Max.
Street design factors	Slope*	%	1.992	2.026	0
	Sidewalk width*	Meter	4.367	2.748	0
	Number of lanes*	Number	3.965	2.690	1
	Bus-dedicated lane	1. yes; 0. no	0.491	-	0
	Centerline	1. yes; 0. no	0.565	-	0
	Safety fence	1. yes; 0. no	0.220	-	0
	Crossing	1. yes; 0. no	0.640	-	0
	Street light	1. yes; 0. no	0.030	-	0
	Trashcan	1. yes; 0. no	0.004	-	0
	Curb	1. yes; 0. no	0.010	-	0
	Street tree	1. yes; 0. no	0.332	-	0
	Shade	1. yes; 0. no	0.001	-	0
	Etc.	1. yes; 0. no	0.534	-	0
	Year	2009	1. yes; 0. no	0.198	-
2012		1. yes; 0. no	0.200	-	0
2013		1. yes; 0. no	0.201	-	0
2014		1. yes; 0. no	0.201	-	0
2015		1. yes; 0. no	0.201	-	0
District	Dobong-gu	1. yes; 0. no	0.023	-	0
	Dongdaemun-gu	1. yes; 0. no	0.038	-	0
	Dongjak-gu	1. yes; 0. no	0.022	-	0
	Eunpyeong-gu	1. yes; 0. no	0.034	-	0
	Gangbuk-gu	1. yes; 0. no	0.027	-	0
	Gangdong-gu	1. yes; 0. no	0.033	-	0
	Gangnam-gu	1. yes; 0. no	0.067	-	0
	Gangseo-gu	1. yes; 0. no	0.032	-	0
	Geumcheon-gu	1. yes; 0. no	0.016	-	0
	Guro-gu	1. yes; 0. no	0.042	-	0
	Gwanak-gu	1. yes; 0. no	0.032	-	0
	Gwangjin-gu	1. yes; 0. no	0.033	-	0
	Jongno-gu	1. yes; 0. no	0.063	-	0
	Jung-gu	1. yes; 0. no	0.067	-	0
	Junngang-gu	1. yes; 0. no	0.040	-	0
	Mapo-gu	1. yes; 0. no	0.039	-	0
	Nowon-gu	1. yes; 0. no	0.060	-	0
	Seocho-gu	1. yes; 0. no	0.063	-	0
	Seodaemun-gu	1. yes; 0. no	0.023	-	0
	Songpa-gu	1. yes; 0. no	0.052	-	0
Seongbuk-gu	1. yes; 0. no	0.038	-	0	
Seongdong-gu	1. yes; 0. no	0.025	-	0	
Yangcheon-gu	1. yes; 0. no	0.040	-	0	
Yeongdeungpo-gu	1. yes; 0. no	0.053	-	0	
Yongsan-gu	1. yes; 0. no	0.036	-	0	

* Normalized to the minimum value of 0 and the maximum value of 1 in the model

를 가진 기본적인 분류 모델이다. 로지스틱 회귀의 모델 파라미터는 로그 가능도 함수(log likelihood function)를 최대화하거나 로지스틱 비용을 최소화하는 방향으로 업데이트되는 과정을 거쳐 학습된다. 로지스틱 회귀의 입력값에 대해 가중치 업데이트 여부를 결정하는 함수는 다음과 같은 시그모이드(sigmoid) 함수이다(Muller and Guido, 2016).

$$f(x) = \frac{1}{1+e^{-x}}$$

로지스틱 회귀 모형은 L2 정규화 항의 가중치 람다(lambda) 같은 하이퍼파라미터를 조정하여 모형의 복잡도를 규제할 수 있다(Muller and Guido, 2016).

2) 랜덤 포레스트

랜덤 포레스트는 다수의 모형을 결합하는 앙상블 학습(ensemble learning) 모형의 하나로 다수의 무작위 의사결정나무를 결합한 모형이다(Breiman, 2001). 랜덤 포레스트는 주어진 학습데이터에서 무작위로 중복을 허용하는 부트스트랩(bootstrap) 방식으로 n개의 데이터 세트를 선택하고, 선택된 데이터 샘플에서 중복 허용 없이 d개의 변수를 선택한다. 이 d 값은 보통 주어진 데이터의 전체 변수 개수의 제곱근을 사용한다(Muller and Guido, 2016).

이렇게 선택된 데이터 샘플로 의사결정나무를 생성하는 과정을 k번 반복한다. 이렇게 생성된 k개의 의사결정나무의 결과의 평균이나 다수의 예측값을 선택하는 방법을 앙상블 기법이라고 한다. 이렇게 다수의 부트스트랩 샘플을 앙상블 기법을 활용하여 하나의 분류기로 결합하는 방식을 Bootstrap Aggregating을 줄여서 배깅(bagging)이라고 한다. 랜덤 포레스트 모형에서 중요한 하이퍼파라미터는 배깅에 사용되는 의사결정나무의 개수인 k이며 일반적으로 k 값이 커지면 모형 성능이 향상된다(Breiman, 2001).

랜덤 포레스트의 중요한 장점 중 하나는 각 독립 변수의 중요도(feature importance)를 계산할 수 있다는 점이다. 포레스트 모형의 모든 노드에서 사용된 독립 변수의 종류와 그 노드에서 얻은 정보량의 증가를 구할 수 있으므로, 각각의 독립 변수들이 얻어낸 정보량 증가의 평균을 비교하면 개별 독립 변수의 상대적 중요도를 비교할 수 있다(Muller and Guido, 2016).

3) 인공신경망

인공신경망은 생물학의 신경망에서 영감을 얻은 학습 알고리즘으로, 본 연구에서는 기본적인 다층 퍼셉트론(Multilayer Perceptron) 모형을 사용하였다. 다층 퍼셉트론 모형은 외부 입력을 받아들이는 입력층(input layer), 처리된 결과를 출력하는 출력층(output layer), 입력층과 출력층 사이에 위치하여 외부로 나타나지 않는 은닉층(hidden layer)으로 구성된다(Buduma, 2017).

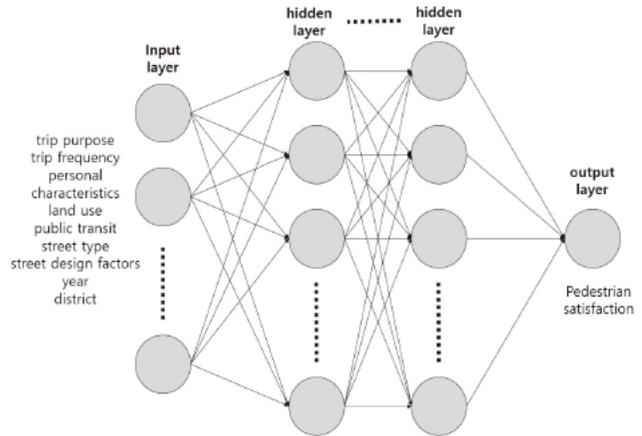


Figure 2. Structure of artificial neural network

각 층은 여러 개의 노드로 구성되며 연결 가중치로 다른 층의 노드와 연결되어 있다. 신경망 구조에서 입력층의 입력은 활성화 함수를 통과하여 은닉층을 거쳐 최종적으로 출력층으로 전달된다(그림 2). 이러한 인공신경망은 역전파 알고리즘으로 빅데이터를 학습하여 예측성능을 확보한다. 인공신경망 모형의 주요 하이퍼파라미터는 출력함수의 종류, 학습률, 은닉층 개수, 은닉층 노드의 개수 등이 있다(Buduma, 2017).

3. 모형평가 방법

최선의 예측모형을 만들기 위해서는 다양한 조건에서 모형의 성능을 평가할 수 있는 정량적인 방법이 필요하다. 본 연구에서는 보행만족도를 “만족”과 “불만족”의 두 가지 카테고리로 나누었기 때문에, 보행만족도 예측모형은 일종의 이진 분류기로 볼 수 있다. 분류기의 성능을 측정하기 위한 기본적인 도구는 혼동행렬(confusion matrix)이다. <표 2>에 제시된 혼동행렬은 분류기의 예측과 실제값을 요약해놓은 표이다(Muller and Guido, 2016).

혼동행렬에서 실제 양성인(만족한) 사람들을 기준으로 보면, True Positive는 만족한 보행자를 만족했다고 정확하게 분류하는 경우이고, False Negative는 만족한 보행자를 불만족했다고 잘못 분류하는 경우이다.

실제값이 음성인(불만족한) 사람들을 기준으로 보면, False Positive는 불만족한 보행자를 만족했다고 잘못 분류하는 경우이고, True Negative는 불만족한 보행자를 불만족했다고 정확하게

Table 2. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

분류하는 경우이다.

정확도(accuracy)는 분류기의 성능을 평가하는 가장 기본적인 방법으로, 전체 데이터 중에서 정확하게 분류된 데이터의 비율이다(Muller and Guido, 2016).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$

정확도가 가장 일반적인 분류모형 평가 방법이지만, 더욱 세분된 모형평가 지표를 이용하여 다면적으로 분류모형을 평가하는 방법이 필요하다.

보행만족도 평가모형에서 발생할 수 있는 첫 번째 오류는 만족한 보행자를 불만족했다고 잘못 분류하는 경우이다. 이러한 오류의 비율을 위양성률(false negative rate, type I error)이라고 한다(Muller and Guido, 2016).

$$False\ Negative\ Rate = \frac{FN}{TP+FN}$$

두 번째 오류는 불만족한 보행자를 만족했다고 잘못 분류하는 경우이다. 이러한 오류의 비율을 위양성률(false positive rate, type II error)이라고 한다(Muller and Guido, 2016).

$$False\ Positive\ Rate = \frac{FP}{FP+TN}$$

만족한 보행자를 만족했다고 정확하게 예측하는 비율을 민감도(sensitivity), 재현율(recall) 또는 실제 양성률(true positive rate)이라고 한다(Muller and Guido, 2016). 이 민감도가 높다는 것은 만족한 보행자를 불만족했다고 잘못 분류하는 False Negative Rate가 작다는 의미이다.

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN}$$

정밀도(Precision)는 분류모형이 만족하리라 예측한 보행자 중 실제로 만족한 보행자 비율이다(Muller and Guido, 2016). 정밀도가 높다는 것은 예측모형이 추천하는 보행로를 선택할 경우 만족하는 사람의 비율이 높다는 것을 의미한다. 따라서 보행자 경로 추천 모형의 경우 정밀도를 모형 성능 평가의 중요한 지표로 볼 수 있다.

$$Precision = \frac{TP}{TP+FP}$$

F1 Score는 민감도와 정밀도의 조화평균(harmonic mean)으로 만든 지표로서, 둘 중 하나의 값이라도 매우 작은 경우에 F1 Score도 매우 작아진다. 따라서 민감도나 정밀도 중 하나를 높이기 위해 다른 값을 희생하면서 만든 분류기는 낮은 F1 Score를 가

진다(Muller and Guido, 2016).

ROC (Receiver Operating Characteristics) Curve는 분류모형의 전반적인 성능을 평가하는 또 다른 유용한 방법이다. 분류모형은 양성으로 분류될 확률을 계산한 후 특정한 절단 값(cut-off value)을 이용하여 양성 또는 음성으로 분류한다. 예를 들어 0.5를 절단 값으로 사용하면, 양성으로 분류될 확률이 0.5 이상이면 양성으로 분류하게 되고, 0.5 이하이면 음성으로 분류하게 된다. 따라서 절단 값의 선택은 혼동행렬을 비롯한 민감도와 특이도 등의 지표에 영향을 미친다(송상욱, 2018).

ROC Curve는 각각 다른 절단 값에 대한 실제 양성률과 위양성률을 그래프로 표현한 것이다. 분류의 정확도는 ROC Curve 아래의 면적(AUC, Area Under the ROC Curve)으로 측정한다. 면적이 1이라면 완벽한 분류를 의미하고 0.5라면 무작위예가 가까운 분류를 의미한다(송상욱, 2018).

일반적으로 AUC 값에 따라 모형의 결과를 덜 정확한(0.5 < AUC ≤ 0.7), 중등도의 정확한(0.7 < AUC ≤ 0.9), 매우 정확한(0.9 < AUC < 1) 분류로 평가한다(송상욱, 2018). <그림 3>에 제시된 ROC Curve가 왼쪽 모서리에 가까울수록 실제 양성률이 높고 위양성률이 낮은 정확한 분류로 해석할 수 있다(송상욱, 2018).

4. 분석 방법 및 과정

일반적으로 기계학습 모형의 성능은 모형의 복잡도를 제어하는 하이퍼파라미터(hyperparameter) 설정에 따라 달라진다. 따라서 다른 모형과의 비교 전에 하이퍼파라미터 설정에 따른 최적의 모형을 선택할 필요가 있다. 이러한 모형 최적화를 위해서 본 연구에서는 k겹 교차검증(k-fold cross validation) 방법을 적용하였다. 이 방법은 훈련 데이터를 무작위로 k 등분하고 등분된 훈련 데이터 중 k-1개를 훈련 데이터로 사용하고 나머지 1개의 데이터를 이용하여 모형의 성능을 검증한다. 이 과정을 등분된 숫자만큼 k번 반복하여 도출한 평가결과를 평균하여 최종 평균 정

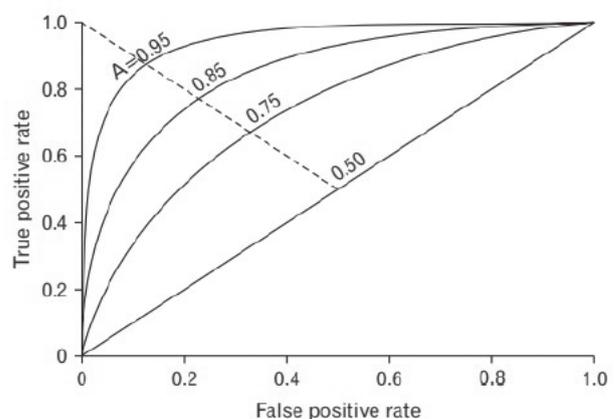


Figure 3. ROC curve

Source: Figure 1. in Song (2018)

확도를 산출한다(Muller and Guido, 2016).

일반적으로 기계학습에서는 데이터를 훈련 데이터셋, 검증 데이터셋, 테스트 데이터셋으로 나누어, 다양한 하이퍼파라미터 조건에서 훈련 데이터셋으로 훈련한 모형을, 검증 데이터셋을 이용하여 최적화하고, 테스트 데이터셋으로 평가한다. 이렇게 데이터를 한 번 나누는 것보다 k겹 교차검증 방식을 사용하면 테스트 데이터셋에 각 샘플이 정확하게 한 번씩 들어가기 때문에 모든 샘플에 대해 모델이 잘 일반화되며 더 정확한 모형을 만들어 낸다(Muller and Guido, 2016).

본 연구에서는 총 99,589개의 관찰 값을 가진 데이터 중 20%인 19,918개의 관찰 값을 무작위로 추출하여 최종 모형 성능 평가를 위한 테스트 데이터로 구축하였다. 나머지 80%로는 관찰 값을 79,671개 포함한 훈련 데이터를 생성하여 5겹 교차검증에 적용하였다. 기계학습 모형별로 하이퍼파라미터를 변화시키면서 5겹 교차검증으로 산출된 평균 정확도가 가장 높은 모형을 방법별로 최적 하이퍼파라미터 설정값을 가진 모형으로 결정하였다.

최종적으로 결정된 모형의 성능 평가는 테스트 데이터를 이용하여 정확도, 재현율, 정밀도, F1 Score, AUC를 계산하는 방식으로 수행하였다. 이 결과를 통해 기계학습 모형별 최종 모형을 선택하였다. <그림 4>는 분석 과정을 도식화하여 제시하였다.

본 연구의 모든 분석은 기계학습 분석에 가장 널리 이용되는 프로그래밍 언어인 파이썬(Python)의 기계학습 라이브러리인 사이킷런(Scikit-learn)을 이용하였다(Muller and Guido, 2016).

IV. 분석결과

1. 기계학습 모형 구축

1) 로지스틱 회귀

로지스틱 회귀 모형은 L2 정규화 항의 가중치 람다(lambda)를 조정하여 모형의 복잡도를 규제함으로써 모형의 일반화 성능을

최적화할 수 있다. 사이킷런에서는 람다의 역수로 정의되는 매개 변수 C를 이용하여 규제의 강도를 조절한다(Muller and Guido, 2016). C값이 높아지면 규제가 감소하여 최대한 훈련 데이터에 모형이 맞춰지며, C값이 낮아지면 규제가 강화되어 모형의 복잡도가 감소한다.

본 연구에서는 C값을 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1,000으로 설정하여 k겹 교차검증을 통해 평균 정확도를 계산하였다. 그 결과 C값이 0.0001일 때 평균 정확도가 0.768로 가장 높았으므로 이 모형을 최적의 로지스틱 회귀 모형으로 선택하였다.

2) 랜덤 포레스트

다수의 모형을 결합하는 랜덤 포레스트 모형의 중요한 하이퍼파라미터는 배경에 사용되는 의사결정나무의 개수이며 일반적으로 개수가 많아지면 모형 성능이 향상된다.

본 연구에서는 의사결정나무의 개수를 100개에서부터 800개까지 50개씩 증가시키면서 k겹 교차검증으로 평균 정확도를 계산하였다. 평균 정확도는 의사결정 나무의 개수가 250개일 때 0.796에 도달했고, 이후 개수의 크기가 늘어나도 더 향상하지 않았다. 따라서 의사결정나무 800개를 결합한 모형을 최적 랜덤 포레스트 모형으로 선택하였다.

3) 인공신경망

본 연구에서 사용한 인공신경망 모형은 입력층, 은닉층, 출력층으로 이루어진 단순한 다층 퍼셉트론이다. 이 모형의 주요 하이퍼파라미터는 출력함수의 종류, 학습률, 은닉층 개수, 은닉층 노드의 개수 등이 있으며, 본 연구에서는 은닉층의 개수와 노드의 개수를 변화시켜 모형을 최적화하였다. 은닉층은 1개에서 3개까지 증가시켰으며, 각각의 경우 은닉층 노드의 개수를 100개에서 300개까지 증가시키며, k겹 교차검증으로 평균 정확도를 계산하였다. 실험 결과 100개의 노드를 포함한 2개의 은닉층을 가진 모형의 정확도가 0.773으로 가장 높았으며, 은닉층의 개수나 노드의 수를 더 늘리면 오히려 정확도가 감소하였다. 따라서, 100개의 노드를 포함한 2개의 은닉층을 가진 다층 퍼셉트론을 최적 모형으로 선택하였다.

2. 기계학습 모형 성능 비교 평가

전체 데이터의 80%인 훈련 데이터를 이용하여 k겹 교차검증을 통해 최적화된 로지스틱 회귀, 랜덤 포레스트, 인공신경망 모형의 예측성능 평가를 전체 데이터의 20%인 테스트 데이터를 이용하여 수행하였다(표 3).

테스트 데이터에서의 성능 평가 결과 랜덤 포레스트 모형의 정확도가 0.798로 세 모형 중에서 가장 높게 나왔고, 인공신경망 모

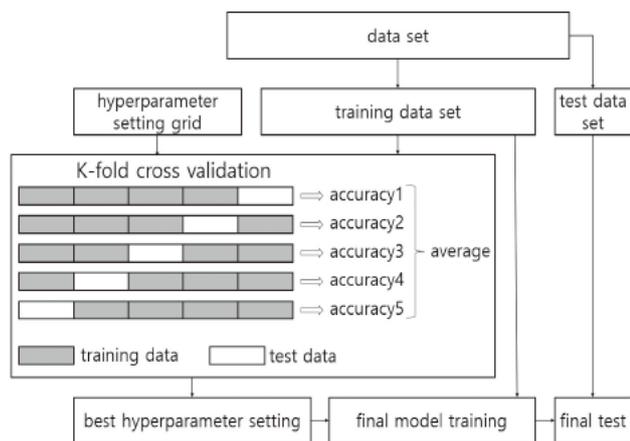
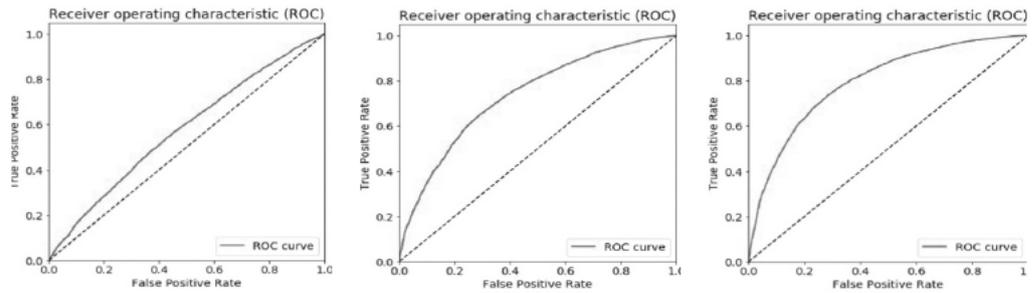


Figure 4. Machine learning procedure

Table 3. Model performance evaluation result

Model	Logistic Regression	Artificial Neural Network	Random Forest
Accuracy	0.764	0.773	0.798
Recall	1.000	0.917	0.906
Precision	0.764	0.811	0.842
F1 Score	0.868	0.868	0.873
AUC	0.575	0.738	0.795

ROC Curve



Hyper parameter setting	L2 penalty C: 0.0001	hidden layer node: 100-100	number of trees(k): 800
-------------------------	----------------------	----------------------------	-------------------------

형의 정확도는 0.773, 로지스틱 회귀 모형의 정확도는 0.764로 가장 낮았다.

실제로 만족한 보행자 중에서 모형이 정확하게 예측한 비율을 나타내는 재현율은 로지스틱 회귀 모형이 1.000으로 가장 높았다. 그러나 모형이 만족하리라 예측한 보행자 중에서 실제로 만족한 비율을 나타내는 정밀도는 랜덤 포레스트가 0.842로 가장 우수했다. 모형의 용도에 따라 재현율이 더 중요할 수도 있고, 정밀도가 더 중요할 수도 있다. 보행만족도 예측모형은 보행자가 만족할 수 있는 경로와 지역을 제시하는 것이 목표이기 때문에 정밀도가 정확한 것이 중요하다. 또한, 모형의 전반적인 성능을 측정하는 F1 Score와 AUC도 랜덤 포레스트 모형이 가장 높았다. 따라서 본 연구에서는 정확도, 정밀도, F1 Score, AUC가 가장 높은 랜덤 포레스트 모형의 성능이 가장 우수하다고 판단하였다.

3. 변수 중요도

랜덤 포레스트는 다른 회귀 모형과 비교하면 각 변수의 해석이 쉽지 않다. 그러나 랜덤 포레스트 모형의 변수 중요도는 주어진 변수가 모형의 노드 불순도(node impurity)를 얼마나 개선하는 지로 측정할 수 있다.

본 연구에서는 개별 변수들이 60세 이상 고령 보행자와 일반 보행자의 보행만족도에 영향을 미치는 정도가 다르다는 가정하에 60세 이상과 미만 보행자별로 랜덤 포레스트 모형을 구축하여 변수 중요도를 측정하였다(그림 5).

전반적으로 경사도와 보도폭이 가장 중요한 변수로 측정되었다. 60세 이상과 미만 연령대의 중요도를 비교해보면 가로설계요소의 중요도는 고령 보행자들에게 더 중요하게 나타났고, 개인특성, 보행목적, 보행빈도 변수는 60세 미만의 보행자에게 더 중요하게 나타났다.

V. 요약 및 결론

본 연구는 보행만족도 기반의 보행 경로 서비스 개발을 위한 기초연구로서, 서울시 유동인구 조사 속성자료 이용하여 로지스틱, 랜덤 포레스트, 인공신경망 알고리즘 기반의 보행만족도 예측모형을 개발하고 각 모형의 예측성능을 평가하였다. 최근 몇몇 연구들이 딥러닝 기법을 이용하여 보행만족도를 평가하고자 하는 시도를 하였으나 충분한 수의 관찰 값을 확보하지 못하는 한계가 있었다. 그에 비해 본 연구는 서울시에서 5년간 수집한 약 10만 건의 보행만족도 설문 조사자료를 이용하여 과적합을 방지하고 일반화 성능을 확보할 수 있는 예측모형을 개발했다는 데 그 의의가 있다.

본 연구의 주요 결과는 다음과 같다. 서울시 유동인구 조사의 보행만족도 설문 조사 표본 10만 개 중 결측치나 오류를 제거한 총 99,589개의 표본을 무작위로 8:2의 비율로 훈련과 테스트 데이터로 할당하였다. 이 훈련 데이터를 활용한 5겹 교차검증 방식으로 최적의 로지스틱, 랜덤 포레스트, 인공신경망 모형을 구축하였다. 마지막으로 테스트 데이터를 이용하여 최적화된 세 개의 모형의 예측성능을 비교하였다. 그 결과 재현율을 제외한 정확

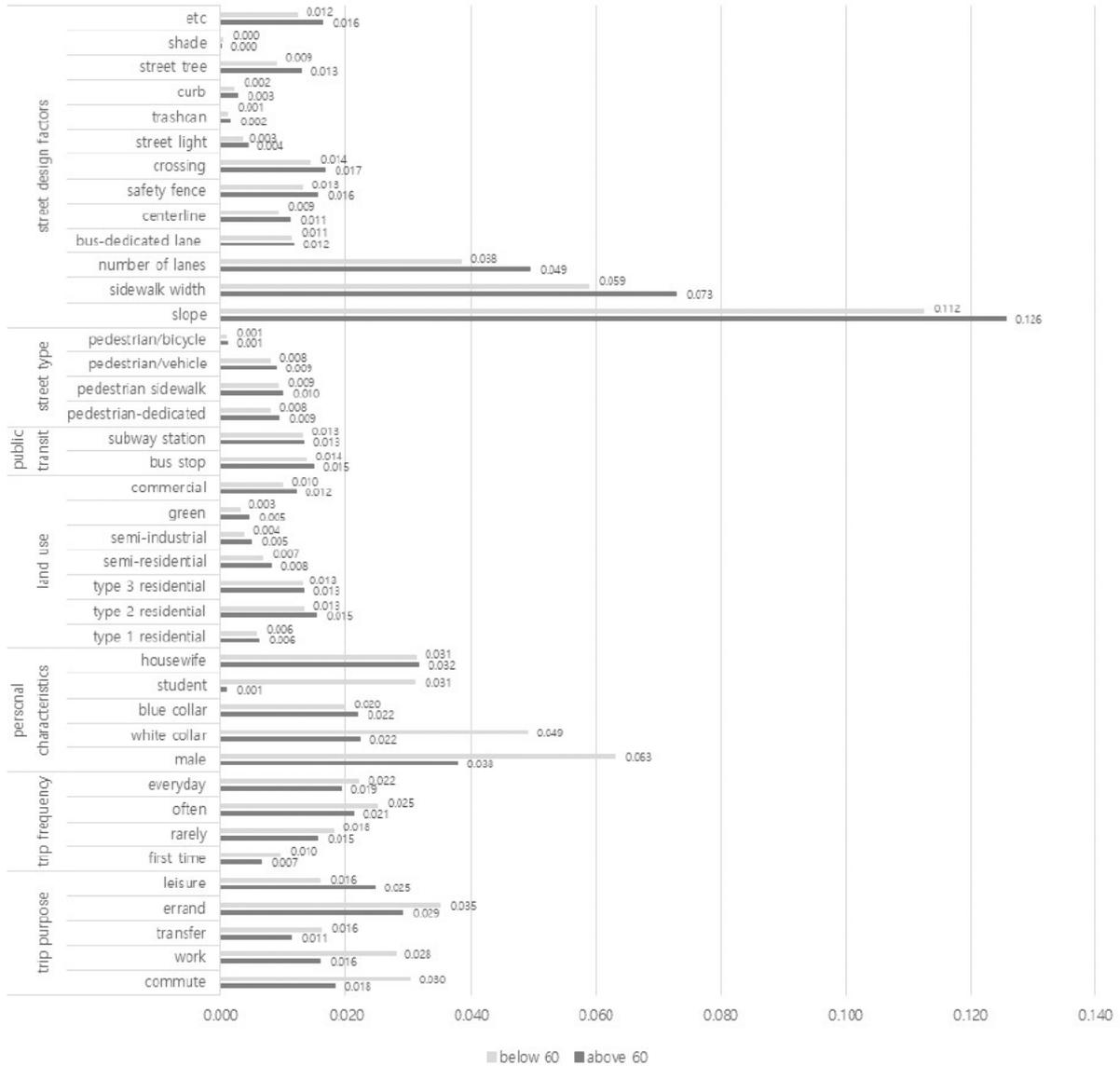


Figure 5. Feature importance of Random Forest

도, 정밀도, F1 Score, AUC에서 랜덤 포레스트 모형이 가장 우수하였다.

랜덤 포레스트 모형의 정확도는 0.798로 나타났으며, 이는 테스트 데이터에 포함된 19,918명의 보행자의 보행만족도 예측 결과 중 79.8%가 정확하였음을 의미한다. 랜덤 포레스트 모형의 재현율 0.906은 테스트 데이터에서 실제 만족에 속하는 보행자 중 90.6%를 모형이 정확히 예측했다는 의미이다. 정밀도는 0.842로 나타났으며, 이는 모형이 만족하리라 예측한 보행자 중 84.2%가 실제로 만족했다는 것을 의미한다.

정밀도가 우수하다는 의미는 개인과 지점의 특성을 고려하여 모형이 만족도가 높을 것으로 예측한 지점을 추천한다면, 그 지점에서 실제로 보행자가 만족할 확률이 높다는 것이다. 본 연구는 보행자의 만족도가 높은 보행 경로를 추천하기 위한 기초연구이기 때문에 이 정밀도가 중요한 지표로 판단된다. 그러나 84.2% 수준의 정밀도는 로지스틱 모형 등 다른 모형보다는 우수하지만,

서비스 제공 측면에서는 개선이 필요하다. 앞으로 보행만족도에 영향을 미칠 수 있는 변수를 추가하거나, 데이터 누적이 이루어진다면 예측성능이 향상될 것으로 기대한다.

또한, 연령에 따라 같은 조건에서 느끼는 보행만족도가 다를 수 있으므로 교통약자인 60세 이상 고령 보행자를 다른 연령대 보행자와 분리하여 보행만족도 랜덤 포레스트 모형을 구축하였다. 이 두 보행자 집단의 보행만족도에 영향을 미치는 요인의 중요도를 측정된 결과를 정리하면 다음과 같다.

가로환경설계요소 중 경사도가 가장 중요하게 측정되었고, 보도폭이 두 번째, 차선 수가 세 번째로 중요하게 나타났다. 이 결과는 보행만족도 관련 선행연구 결과와 일치하고 있다.

60세 이상 고령자와 60세 미만 보행자를 비교하면, 고령 보행자에게는 상대적으로 경사도, 보도폭, 차선 수 등 물리적 가로환경 변수가 중요하다고 분석되었다. 이는 고령 보행자는 경사가 심하거나 보도 폭이 좁은 등의 가로환경 제한에 더욱 불편함을 느

킨다는 것을 의미한다. 따라서 보행 경로 서비스는 이러한 고령자의 특성을 고려한 예측모형을 활용하여 개발되어야 함을 시사하고 있다.

본 연구는 랜덤 포레스트 알고리즘을 이용하여 보행자의 특성, 보행 활동의 특성, 가로 환경의 특성을 고려하여 보행자가 특정 지점에서 느끼는 보행만족도를 예측하는 모형을 개발하였다. 보행만족도가 높은 지점에서는 편리하고 효율적인 보행 활동이 이루어지고 있음을 의미한다. 따라서 보행자에게 보행만족도가 높은 경로를 제시한다면, 더욱 편리하고 효율적인 보행 활동을 증진할 수 있을 것이다.

이 연구의 결과는 앞으로 보행만족도 기반 보행 경로 서비스의 경로 최적화를 위해 활용될 것이다. 즉 개인별 특성을 고려하여 보행만족도가 가장 높게 나오도록 예측된 경로를 제시하는 서비스의 알고리즘으로 기능할 것이다. 특히 연령대별로 가로환경요인별 가중치를 최적화함으로써 고령 보행자를 위해 특화된 보행 경로 서비스가 가능하다.

그러나 본 연구의 한계는 다음과 같다. 연구에 사용된 데이터는 지점 단위로 설문 조사가 이루어졌다. 그러나 보행 경로 서비스를 위해서는 보행로 링크 단위의 분석이 이루어져야 한다. 이를 위해서는 개별 링크 단위로 본 연구에서 사용한 속성을 포함한 데이터와 링크와 링크 또는 목적지(POI: point of interest)가 만나는 노드의 속성을 포함한 데이터가 구축되어야 한다. 그러나 현재까지는 서울시 전역에 보행자 노드와 링크 데이터가 구축되어 있지 않다.

이를 고려한 향후 연구 방향을 정리하면 다음과 같다. 첫째, 서울시 내의 대상지를 선정하여 정밀 보행자 노드·링크 데이터를 구축한다. 이 데이터에는 보행에 영향을 미칠 수 있는 모든 속성 값을 부여한다. 또한, 이 지역의 POI 데이터베이스와 노드·링크 데이터를 결합한다.

둘째, 본 연구에서 개발한 보행만족도 예측 랜덤 포레스트 모형과 보행자 노드·링크 데이터베이스를 이용하여 링크별 예상 보행만족 확률을 추정하여 보행친화지수를 개발한다.

보행친화지수는 Walkscore¹⁾와 유사하게 지역별 보행성을 제시하는 지표로 이용될 수 있을 것이다. 그러나 Walkscore는 POI를 기반으로 개발되었기 때문에 보행자의 만족도를 반영하지는 못한다. 따라서 보행만족도 기반의 보행친화지수는 실제 현장에서 보행자가 느끼는 만족도를 반영하는 지표로서 의의가 있을 것이다.

마지막으로 보행친화지수를 최대화하는 보행 경로 추천 알고리즘을 개발할 것이다. 현재의 보행 경로 추천 알고리즘은 보행 거리 또는 보행 시간을 최소화하는 경로를 제공한다. 그러나 이러한 경로는 보행만족도와는 상관관계가 약할 수 있으며, 특히 고령자 등 보행 약자에게 적합한 경로를 추천하지 못할 가능성이 크다. 따라서, 본 연구 결과를 기반으로 개발할 보행친화지수와

보행만족도 기반의 경로 추천 알고리즘은 더욱 편리하고 안전한 보행 활동을 위한 정보를 제공하여, 장기적으로는 보행 활동 증진과 보행환경 개선에 기여할 것으로 기대한다.

주1. <https://www.walkscore.com/>

인용문헌 References

- 김창국·임하나·최창규, 2016. “보행 목적별 보행자 만족도에 영향을 미치는 근린 건조환경 구성요소 특성 분석”, 「국토계획」, 51(4): 145-159.
- Kim, C.G., Im, H.N., and Choi, C.G., 2016. “Built Environment, Walking Trip for Different Purposes, and Pedestrian Satisfaction”, *Journal of Korea Planning Association*, 51(4):145-159.
- 박근덕, 2018. “딥러닝과 구글 Street View API를 활용한 가로 이미지의 보행만족도 분석”, 한양대학교 대학원 석사학위논문.
- Park, K.D., 2018. “Analysis of Walking Satisfaction of Street Image Using Deep Learning and Google Street View API”, Master's Degree Dissertation, Hanyang University.
- 박소현·최이명·서한림·김준형, 2009. “주거지 보행환경 인지가 생활권 보행만족도에 미치는 영향에 관한 연구”, 「대한건축학회 논문집 계획계」, 25(8): 253-261.
- Park, S.H., Choi, Y.M., Seo, H.L., and Kim, J.H., 2009. “Perception of Pedestrian Environment and Satisfaction of Neighborhood Walking: An Impact Study Based on Four Residential Communities in Seoul, Korea”, *Journal of the Architectural Institute of Korea Planning & Design*, 25(8): 253-261.
- 배성완·유정석, 2018. “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지 예측”, 「주택연구」, 26(1): 107-133.
- Bae, S.W. and Yu, J.S., 2018. “Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model”, *Housing Studies Review*, 26(1): 107-133.
- 성현곤·김태호·강지원, 2011. “구조방정식을 활용한 보행환경 계획요소의 이용만족도 평가에 관한 연구: 종로 및 강남 일대를 대상으로”, 「국토계획」, 46(5): 275-288.
- Sung, H.G., Kim, T.H., and Kang, J.W., 2011. “A Study on Evaluation of User Satisfaction for Walking Environment Planning Elements through Structural Equation Modeling: The Case of Jongno and Kangnam Areas”, *Journal of Korea Planning Association*, 46(5): 275-288.
- 송상욱, 2018. “진단 도구의 정확성 평가”, 「대한가정의학회지」, 8(1): 1-2.
- Song, S.W., 2018. “Assess the Accuracy of Diagnostic Tools”, *Korean Journal of Family Medicine*, 8(1): 1-2.
- 심재현·조찬호·이성호, 2007. “인공신경망을 이용한 산업용지의 공급가격 산정”, 「국토계획」, 46(1): 223-232.
- Shim, J.H., Cho, C.H., and Lee, S.H., 2007. “The Industrial Land Price Appraisal Based on Artificial Neural Network”, *Journal of Korea Planning Association*, 46(1): 223-232.
- 이요섭·문필주, 2017. “딥 러닝 프레임워크의 비교 및 분석”, 「한국전자통신학회 논문지」, 12(1): 115-122.
- Lee, Y.S. and Moon, P.J., 2017. “A Comparison and Analysis of Deep

- Learning Framework”, *The Journal of the Korea Institute of Electronic Communication Sciences*, 2(1): 115-122.
9. 이수기·이윤성·이창관, 2014. “보행자 연령대별 보행만족도에 영향을 미치는 가로환경의 특성분석”, 「국토계획」, 49(8): 91-105.
Lee, S.G., Lee, Y.S., and Lee, C.K., 2014. “An Analysis of Street Environment Affecting Pedestrian Walking Satisfaction for Different Age Groups”, *Journal of Korea Planning Association*, 49(8): 91-105.
 10. 이수기·고준호·이기훈, 2016. “근린환경특성이 보행만족도에 미치는 영향 분석: 서울서베이 2013년 자료를 중심으로”, 「국토계획」, 51(1): 169-187.
Lee, S.G., Ko, J.H., and Lee, G.H., 2016. “An Analysis of Neighborhood Environment Affecting Walking Satisfaction: Focused on the ‘Seoul Survey’ 2013”, *Journal of Korea Planning Association*, 51(1): 169-187.
 11. 이인성·김현욱, 1998. “도시주거지 보행경로 선택행태에 관한 연구: GIS를 이용한 보행환경 만족도의 분석”, 「국토계획」, 33(5): 117-129.
Lee, I.S. and Kim, H.O., 1998. “Pedestrian Path-choice Behavior in Urban Residential Area: Analysis of Environmental Satisfaction Using GIS”, *Journal of Korea Planning Association*, 33(5): 117-129.
 12. 이인성, 2000. “도시주거지 보행경로 선택모형의 개발-중요도, 만족도 및 환경절충모형의 비교”, 「도시설계」, 1(1): 63-78.
Lee, I.S., 2000. “Development of Pedestrian Path-choice Model in Urban Residential Area: Comparison of Importance, Satisfaction, and Environmental Tradeoff Models”, *Journal of The Urban Design Institute of Korea*, 1(1): 63-78.
 13. 이호준·이수기, 2018. “Google Street View API와 딥러닝 모형을 활용한 생활도로 노상주차 식별 방법 연구”, 2018 한국지형공간정보학회 춘계학술대회, 진주시: 경남과학기술대학교, 79-88.
Lee, H.J. and Lee, S.G., 2018. “Study on On-street Parking Detection on Community Road using Google Street View API and Deep Learning”, Paper presented at the 2018 Spring Congress of Korean Society for Geospatial Information Science, Jinju: Gyeongnam National University of Science and Technology, 79-88.
 14. 임하나·김태현·최창규, 2014. “보행 실제거리와 인지 거리의 차이에 영향을 미치는 보행환경 특성에 관한 연구”, 「국토계획」, 49(7): 97-115.
Im, H.N., Kim, T.H., and Choi, C.G., 2014. “What Variables Make the Perceived Walking Distance Shorter than Real Physical Distance?”, *Journal of the Korea Planning Association*, 49(7): 97-115.
 15. 지우석·구연숙·좌승희, 2008. “보행환경 만족도 연구”, 경기도: 경기개발연구원.
Ji, W.S., Koo, Y.S., and Jwa, S.H., 2008. “A Study on Satisfaction for Pedestrian Environment”, Gyeonggi: Gyeonggi Research Institute.
 16. 하재현·이수기, 2017. “다층 퍼셉트론 딥러닝 기법을 활용한 주거지선택 및 통근통행패턴 예측에 관한 연구: 도권 가구통행실태 조사자료의 이동성 수준을 중심으로”, 대한국토·도시계획학회 2017 추계학술대회, 원주시: 상지대학교.
Ha, J.H. and Lee, S.G., 2017. “A Study on the Prediction of Residential Location Choice and Commuting Pattern using the Multi-layered Perceptron Deep Learning Technique: Focused on the Mobility Level of the Seoul Metropolitan Household Travel Survey”, Paper presented at the 2006 Fall Congress of Korea Planning Association, Wonju: Sangji University.
 17. Amerigo, M. and Aragones, J.I., 1997. “A Theoretical and Methodological Approach to the Study of Residential Satisfaction”, *Journal of Environmental Psychology*, 17: 47-57.
 18. Buduma, N., 2017. *Fundamentals of Deep Learning: Designing Next-generation Machine Intelligence Algorithms*, Sebastopol, California: O’Reilly Media.
 19. Brieman, L., 2001. “Random Forests”, *Machine Learning*, 45(1): 5-32.
 20. Kim, S.H., Park, S.J., and Lee, J.S. 2014. “Meso- or Micro-scale? Environmental Factors Influencing Pedestrian Satisfaction”, *Transportation Research Part D*, 30: 10-20.
 21. Methorst, R. and Horst, R., 2010. “Pedestrians’ Performance and Satisfaction”, Paper presented at the Walk21 Conference, Hague, Netherlands.
 22. Muller, A. and Guido, S., 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, Sebastopol, California: O’Reilly Media.
 23. Wang, W., Li, P., Wang, W., and Namgung, M., 2012. “Exploring Determinants of Pedestrians’ Satisfaction with Sidewalk Environments: Case Study in Korea”, *Journal of Urban Planning and Development*, 138(2): 166-172.

Date Received 2019-01-28
 Reviewed(1st) 2019-03-22
 Date Revised 2019-05-03
 Reviewed(2nd) 2019-05-23
 Date Accepted 2019-05-23
 Final Received 2019-06-04