2020년 K-water 대국민 빅데이터 공모전 수행 결과보고서

제 목	압전 에너지 하베스팅 버스 정류장			
공모분야	서비스 개발	O	융합데이터	

개인 정보란은 삭제하였습니다.

Ⅰ. 과제 목표

현 정부는 많은 위험도가 따르면서 환경 파괴 및 지구 온난화를 가속시키는 원자력 발전을 대체하는 방안을 모색하고 있습니다. 따라서 태양열, 풍력, 수력, 지열, 화력 등 다양한 에너지를 활용한 신재생 에너지가 각광을 받고 있습니다. 하지만 높은 발전 단가와 초기설치비용, 소형화의 기술 등 어려움을 겪고 있는 현실입니다. 이러한 상황 속에서 자연에서 쉽게 얻을 수 있는 소음, 진동, 압력 등을 활용한 에너지 하베스팅이 떠오르고 있습니다. 그 중 압전 하베스팅은 다른 에너지 하베스팅 기술에 비하여 높은 변환 효율을 갖고 있으며 에너지원을 우리 주변에서 쉽게 얻을 수 있습니다. 이에 따라 본 과제에서는 도로에서 발생되는 압력, 충격, 진동을 활용하는 방안을 모색하였습니다. 하지만 모든 도로에 사용하기에는 무리가 있다고 판단하여 차량이 날씨, 시간에 영향을 받지 않으며 일정한 교통량이 있고 압전 하베스팅 전기량 생산에 비례하는 요소인 무게를 고려하여 시내 버스를 대상으로 하였습니다. 그리고 정류장 앞에서 생산된 전기를 버스 정류장에 바로 활용하여 버스 정류장의 전력을 공급하는 자체 공급 시스템을 구축하고자 합니다. 대표적 예로 포항시를 선정한 이유는 사람들이 많이 찾는 관광지도 보유하고 있으며 대학교, 공항, 기차역, 항구, 대기업 등 다양한 시설이 있어 다른 지역에 접목시킬 수 있는 장점이 있었습니다.

Ⅱ. 주요 내용

먼저 포항에서 버스 정류장 데이터로 수집, 전처리, 분석 과정을 거쳐 포항의 727개 버스 정류장의 3달간의 버스 이용량을 조사하여 버스 정류장을 운영하는데 필요한 최소의 에너지를 얻을 수 있는 정류장을 선별하는 기준으로 사용하였습니다. 그리고 다른 지역에 접목시킬 수 있도록 다양한 시설의 여부를 조사하여 데이터를 구성하였습니다. 위 과정을 통해 얻은 데이터를 활용하여 버스가 정류장에 정차할 때 발생하는 압전 에너지를 전기 에너지로 변환시킨 후 버스 정류장의 버스도착정보안내시스템(BIS)과 정류장 내부 조명에 사용합니다.

Ⅲ. 활용데이터 및 수행내용

1) 분석데이터

(1) 목표변수

본 과제에서는 교통 빅데이터(한국교통연구원)에서 제공하는 포항시 버스 사용자 교통카드 사용 내역을 이용하였습니다. 자료는 2020-01-01부터 2020-03-31까지 총 857,018개의 자료이며 버스 탑승 시점과 하차 시점으로 이루어진 데이터입니다. 그 중 전처리 과정에서 3월 14일 ~ 3월 24일까지의 자료가 비어있어 3월의 자료(87,674개), 포항시의 버스 중 마을 버스의 경우 그 빈도와 통행량이 적다고 판단하여 모든 Feature들이 F인 정류장 자료(21,807개), 현재 사라지거나 개편되어 정류장이 운영되지 않는 정류장 자료(50,344개), 같은 날 같은 정류장의 경우 하나의 행에 표현하기 위하여 중복되는 정류장의 행(674,532개)을 삭제하였습니다. 이에 총 857,018개의 자료중 23,701개를 이용하여 ‘T(일일 정류장별 버스 정차수 >= 92)’, ‘F(일일 정류장별 버스 정차수 < 92)’로 새롭게 구분한 압전 하베스팅 정류장 설치 여부 변수를 목표 변수로 사용하였습니다.

구분	컬럼명 (영문)	컬럼명 (한글)	데이터 타입
	on_date	승차시각	datetime
	off_date	하차시각	datetime
	route_name	노선명	varchar
	descr	노선설명	varchar
	age_type	승객연령	varchar
	trans_yn	환승여부	varchar
	addfee_yn	추가운임여부	varchar
	start_bstop	승차정류장	varchar
	start_gps_x	승차정류장 GPS X	numeric
	start_gps_y	승차정류장 GPS Y	numeric
	end_bstop	하차정류장	varchar
	end_gps_x	하차정류장 GPS X	numeric
	end_gps_y	하차정류장 GPS Y	numeric

<수정 전 컬럼 정의서(코드북)>

구분	컬럼명 (영문)	컬럼명 (한글)	세부 정보	데이터 타입
	month	달	1월, 2월, 3월을 구분 (3월 데이터를 삭제 시 사용하기 위함)	numeric
	start_bstop	버스 정류장	버스 정류장 정보	varchar
	on_date	승차시각	승차 시각 정보	datetime
	public	공공기관 여부	경찰서, 파출소, 보건소, 소방서, 시청, 행정복지센터 등 공공기관	boolean
	factory	공장 여부	공장	boolean
	apartment	아파트 여부	아파트, 타운 (동으로 이루어진 주거지)	boolean
	kindergarden	유치원, 어린이집 여부	유치원, 어린이집	boolean
	ele_school	초등학교 여부	초등학교	boolean
	mid_school	중학교 여부	중학교	boolean
	high_school	고등학교 여부	고등학교	boolean
	universe	대학교 여부	대학교	boolean
	franchise	프랜차이즈 여부	스타벅스, 빈폴, 엔젤리너스, 네네치킨, 커피베이, bbq, 탐앤탐스, 설빙, 투썸플레이스, 롯데리아, 도미노피자, 서가앤쿡, 치킨더홈, 처가집양념치킨, 호식이두마리치킨, 교촌치킨, 멕시카나, 페리카나, 네네치킨, 한솥, 피자에땅, 뚜레주르, 푸라닭, 이삭토스트, 베스킨라빈스, 이디야, 자담치킨, 파스쿠치, 파리바게트, 맘스터치, 버거킹, 빽다방	boolean
	religion	종교시설 여부	종교 관련 시설	boolean
	mart	마트 여부	대형 마트	boolean
	convenience	편의점 여부	편의점, 작은 슈퍼	boolean
	park	공원 여부	공원	boolean
	hospital	병원 여부	병원, 보건소	boolean
	bank	은행 여부	은행	boolean
	attraction	관광지 여부	유명 관광지, 펜션, 민박	boolean
	Target	설치 가능 여부	하루 간 버스 이동량 >= 92 : T 하루 간 버스 이동량 < 92 : F	boolean

<수정 후 컬럼 정의서(코드북)>

(2) 독립변수

앞서 언급한 압전 하베스팅 정류장 설치 여부에 관한 정류장 주변 시설(반경 150m)들을 파악하여 주요 변수들을 선정하였습니다. 추가적으로 일부는 분석 목적에 맞게 변환하여 응용하였는데 예를 들어, 일일 정류장별 버스 정차 수를 구하기 위하여 bus_b라는 이름을 사용하여 전체 데이터 셋에서 엑셀 함수식을 이용하여 날짜가 같고 정류장 이름이 같은 경우 합산하여 보여줄 수 있도록 하였습니다.

2) 데이터 균형화(data balancing)

데이터 마이닝 기법을 이용하여 예측 모델을 구축하기 위해서는 하나의 목표 변수의 분포를 파악해야 합니다. 이를 위해서는 균형을 맞추는 데이터 균형화 작업이 필요합니다. 본 과제에서 사용된 데이터 셋의 경우 T(정류장 설치)가 885개, F(정류장 미설치)가 22816개 이었기에 비율을 맞추는 작업을 추가로 진행하였습니다. 비율을 맞추기 위하여 언더 샘플링과 오버 샘플링 방법을 사용하였습니다. 오버 샘플링의 경우 낮은 비율을 기준으로 높은 비율의 데이터에서 추출하는 방법으로 본 과제에서는 무작위 추출보다는 중복이 많을 경우 과대적합의 문제가 발생할 수 있기 때문에 최대한 중복없는 데이터 사용을 위하여 직접 885개씩 추출하여 총 10개의 오버 샘플링을 진행하였습니다. 언더 샘플링의 경우 높은 비율을 기준으로 낮은 비율의 데이터의 개수를 늘리는 방법입니다. 따라서 T값을 25번 넣고 619개를 추가로 추출하여 넣은 뒤 진행하였습니다.

3) 모델 구축

독립변수에 분류 알고리즘을 적용시켜 학습을 시킴으로써 데이터 마이닝 모델을 구축합니다. 이 때 사용되는 분류 알고리즘에는 의사결정나무, 인공신경망(ANN), 서포트 벡터 머신(SVM) 등이 있습니다. 이 중 의사결정나무 알고리즘은 모델 구축 시간이 비교적 빠르며, 정확도와 설명력이 뛰어나 많은 연구에서 사용되고 있습니다. 본 과제에서도 이러한 데이터 셋을 바탕으로 타지역에 적용 가능한 기준을 만들기 위하여 의사결정나무 알고리즘을 사용하였습니다.

일반적으로 분류 모델의 경우 학습데이터(training data), 검증데이터(test data)로 분할합니다. 학습데이터는 학습시키기 위해 사용되며 검증데이터는 학습데이터를 이용해 구축된 모델의 성능을 평가하기 위해 사용되며 학습 시에는 사용되지 않아야 합니다. 대부분 이 비율을 7:3으로 사용하기에 본 과제에서도 7:3의 비율로 진행하였습니다.

또한 본 과제에서는 데이터 마이닝 분석을 위해 오픈 소스 소프트

웨어인 Weka를 데이터 마이닝 툴로서 사용하였습니다.

Ⅳ. 결과 및 기대효과

1) 압전 하베스팅 사용 방법 및 전력 사용량

한 시간 당 전력 사용량 4W의 전구 4개와 300W의 BIS(버스정보안내 단말기) 사용을 위해 전구의 경우 하루 6시간, BIS의 경우 17시간 사용하므로 4W x 4개 x 6h + 300W x 17h = 5196W입니다. 압전 하베스팅을 이용하여 얻는 전력의 양은 콘크리트 포장 1cm 매설형 하베스터의 경우로 11톤 트럭 30km/h로 지나갈 때 1세트(발판 2개) 당 8.1W이며 하베스터 발전량 측정 장비는 picoscope 2000 series 사용하였습니다. 본 과제의 경우 버스에 적용을 시켜야 하나 버스의 무게는 11톤 ~ 13톤이므로 최저 무게인 11톤을 기준으로 설계였고 7세트를 적용시켜 한 대당 56.7W를 얻을 수 있습니다.

< 11톤 트럭의 시속에 따른 발전량>

따라서 최소 전력인 5196W를 얻기 위해서는 5196(w/하루)을 56.7 (w/대)로 나눈 값인 91.64(대/하루)가 필요합니다. 버스는 소수점이 있을 수 없으므로 올림을 진행하여 92대를 기준으로 92대를 넘으면 Target 변수를 “T”, 그렇지 않으면 “F”로 정의하였습니다.

압전 하베스터의 크기는 하우징을 포함해서 가로 59 cm, 세로 24 cm이다. 하베스터는 노출형 6개와 매설형 6개로 구분되어 12개의 하베스터가 콘크리트/아스팔트의 2개의 포장도로에 각각 설치되어 총 24개의 압전 하베스터가 설치되었다. 아래 그림과 같이 하베스터 배치 간격 조건은 0.3 m와 0.5 m로 선정되었으며, 차량의 윤거를 고려하여 1.2 m 간격으로 설치되었다.

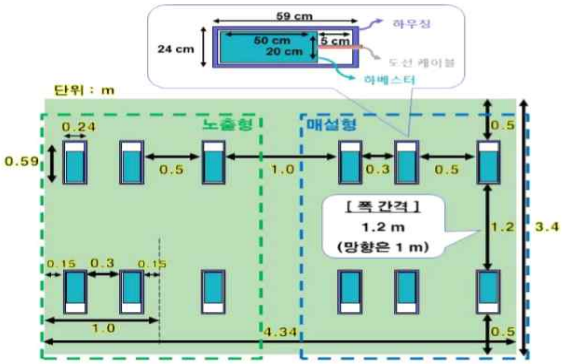


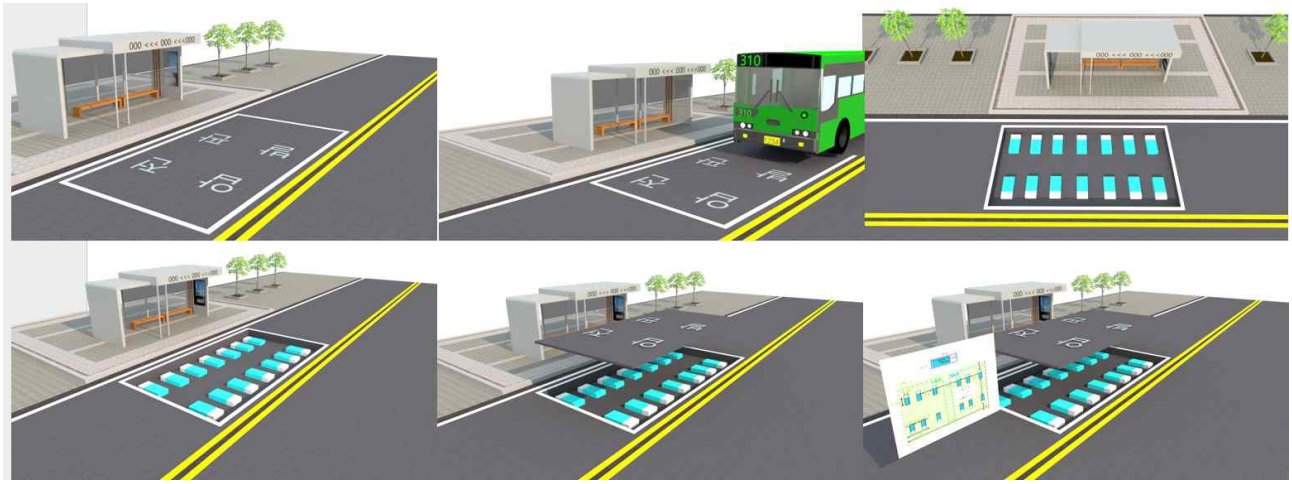
그림 6.6 각 포장도로 압전 하베스터 설치 배치도

(출처 : 한국도로공사 도로교통연구원 2017년 연구보고서)

위 사진은 참고 배치도이며 본 과제는 참고 배치도에서 한 세트가 더 추가된 7세트로 설계하여 진행하였습니다.

(출처 : 한국도로공사 도로교통연구원 2017년 연구보고서)

위 사진 2개로 1세트를 구성하게 됩니다.

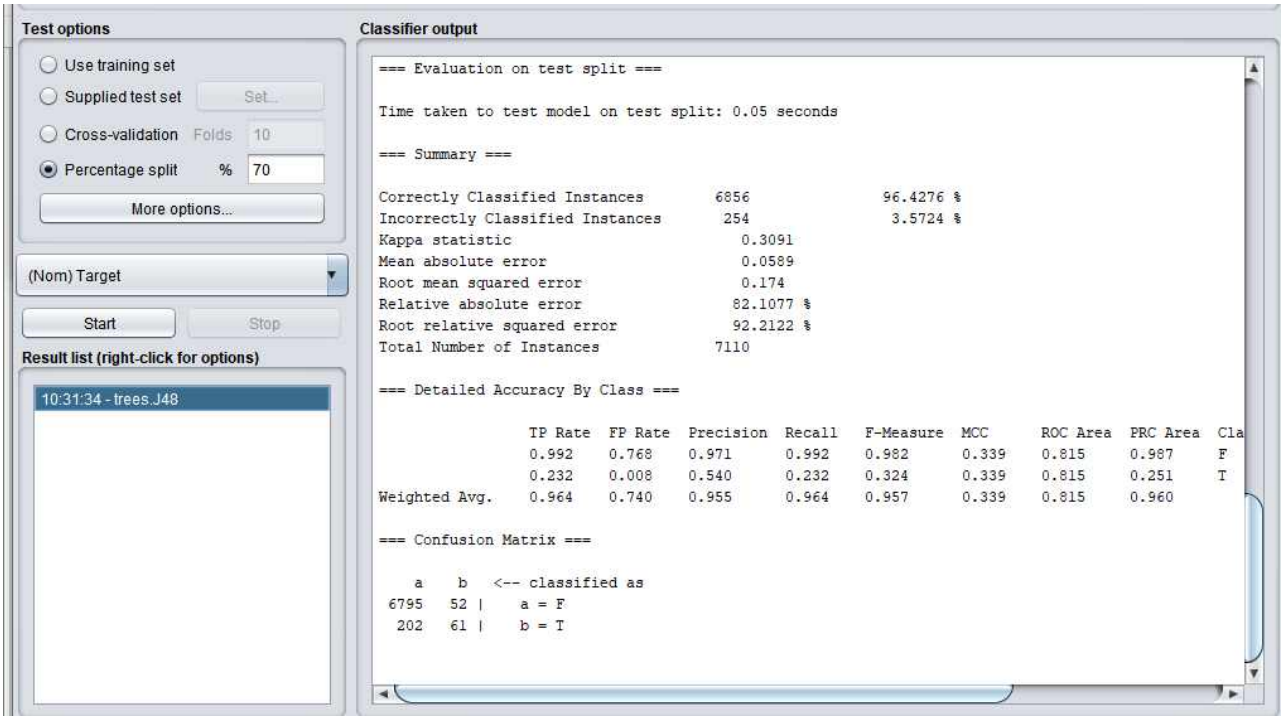


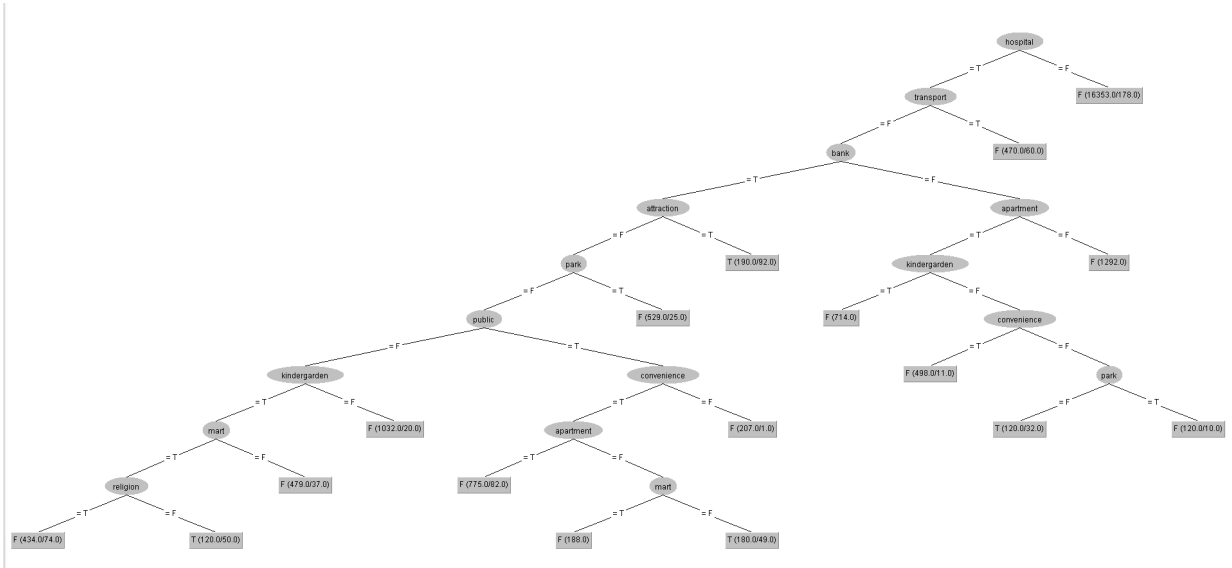
위 사진은 실제 설치하였을 경우를 SketchUp을 통하여 랜더링한 모습입니다.

2) 압전 하베스팅 정류장 설치 여부 예측 결과

의사결정나무 기법을 통해 설치 여부의 확률을 알아보기 위하여 Weka 프로그램을 이용하였습니다.

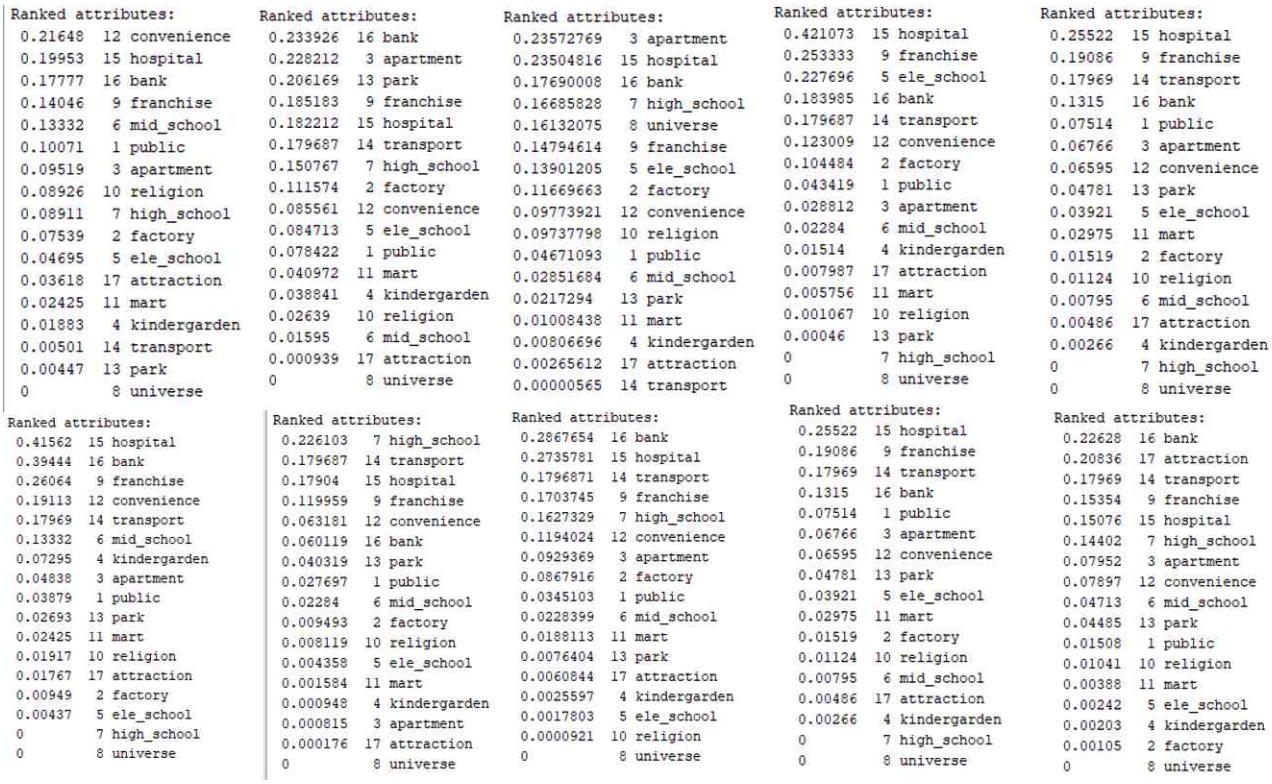
의사결정 나무 이용시 나무의 가지 수를 줄이기 위하여 최소 leaf node 수를 100으로 변경하여 진행하였습니다.





위 사진은 전처리를 거친 후 데이터 셋의 의사결정나무 알고리즘을 돌린 후 결과이며 TP Rate 값에서 알 수 있드시 F의 비율이 너무 많아 T에 대한 정보를 찾기 힘들어 언더샘플링과 오버샘플링의 과정을

진행하였습니다. 오버 샘플링의 경우 중복 데이터를 포함하지 않는 비복원 추출을 통하여 총 10개의 오버샘플링을 하였으며 각 샘플링에서의 Weka의 Select attribution-GainRatio를 이용하여 각각에서 차지하는 변수의 중요도를 파악하였으며 가장 높은 것을 17, 가장 낮은 것을 1의 점수를 부여하여 10개의 합산이 가장 높은 변수를 중요도가 높은 변수로 판단하여 샘플링 이전 모델의 결과와 비교해봤습니다.



```
Ranked attributes:
 0.21648  12 convenience
 0.19953  15 hospital
 0.17777  16 bank
 0.14046   9 franchise
 0.13332   6 mid_school
 0.10071   1 public
 0.09519   3 apartment
 0.08926  10 religion
 0.08911   7 high_school
 0.07539   2 factory
 0.04695   5 ele_school
 0.03618  17 attraction
 0.02425  11 mart
 0.01883   4 kindergarden
 0.00501  14 transport
 0.00447  13 park
 0         8 universe
```

```
Ranked attributes:
 0.233926  16 bank
 0.228212   3 apartment
 0.206169  13 park
 0.185183   9 franchise
 0.182212  15 hospital
 0.179687  14 transport
 0.150767   7 high_school
 0.111574   2 factory
 0.085561  12 convenience
 0.084713   5 ele_school
 0.078422   1 public
 0.040972  11 mart
 0.038841   4 kindergarden
 0.02639   10 religion
 0.01595    6 mid_school
 0.000939  17 attraction
 0          8 universe
```

```
Ranked attributes:
 0.23572769   3 apartment
 0.23504816  15 hospital
 0.17690008  16 bank
 0.16685828   7 high_school
 0.16132075   8 universe
 0.14794614   9 franchise
 0.13901205   5 ele_school
 0.11669663   2 factory
 0.09773921  12 convenience
 0.09737798  10 religion
 0.04671093   1 public
 0.02851684   6 mid_school
 0.0217294   13 park
 0.01008438  11 mart
 0.00806696   4 kindergarden
 0.00265612  17 attraction
 0.00000565  14 transport
```

```
Ranked attributes:
 0.421073  15 hospital
 0.253333   9 franchise
 0.227696   5 ele_school
 0.183985  16 bank
 0.179687  14 transport
 0.123009  12 convenience
 0.104484   2 factory
 0.043419   1 public
 0.028812   3 apartment
 0.02284    6 mid_school
 0.01514    4 kindergarden
 0.007987  17 attraction
 0.005756  11 mart
 0.001067  10 religion
 0.00046   13 park
 0          7 high_school
 0          8 universe
```

```
Ranked attributes:
 0.25522  15 hospital
 0.19086   9 franchise
 0.17969  14 transport
 0.1315   16 bank
 0.07514   1 public
 0.06766   3 apartment
 0.06595  12 convenience
 0.04781  13 park
 0.03921   5 ele_school
 0.02975  11 mart
 0.01519   2 factory
 0.01124  10 religion
 0.00795   6 mid_school
 0.00486  17 attraction
 0.00266   4 kindergarden
 0         7 high_school
 0         8 universe
```

```
Ranked attributes:
 0.41562  15 hospital
 0.39444  16 bank
 0.26064   9 franchise
 0.19113  12 convenience
 0.17969  14 transport
 0.13332   6 mid_school
 0.07295   4 kindergarden
 0.04838   3 apartment
 0.03879   1 public
 0.02693  13 park
 0.02425  11 mart
 0.01917  10 religion
 0.01767  17 attraction
 0.00949   2 factory
 0.00437   5 ele_school
 0         7 high_school
 0         8 universe
```

```
Ranked attributes:
 0.226103   7 high_school
 0.179687  14 transport
 0.17904   15 hospital
 0.119959   9 franchise
 0.063181  12 convenience
 0.060119  16 bank
 0.040319  13 park
 0.027697   1 public
 0.02284    6 mid_school
 0.009493   2 factory
 0.008119  10 religion
 0.004358   5 ele_school
 0.001584  11 mart
 0.000948   4 kindergarden
 0.000815   3 apartment
 0.000176  17 attraction
 0          8 universe
```

```
Ranked attributes:
 0.2867654  16 bank
 0.2735781  15 hospital
 0.1796871  14 transport
 0.1703745   9 franchise
 0.1627329   7 high_school
 0.1194024  12 convenience
 0.0929369   3 apartment
 0.0867916   2 factory
 0.0345103   1 public
 0.0228399   6 mid_school
 0.0188113  11 mart
 0.0076404  13 park
 0.0060844  17 attraction
 0.0025597   4 kindergarden
 0.0017803   5 ele_school
 0.0000921  10 religion
 0           8 universe
```

```
Ranked attributes:
 0.25522  15 hospital
 0.19086   9 franchise
 0.17969  14 transport
 0.1315   16 bank
 0.07514   1 public
 0.06766   3 apartment
 0.06595  12 convenience
 0.04781  13 park
 0.03921   5 ele_school
 0.02975  11 mart
 0.01519   2 factory
 0.01124  10 religion
 0.00795   6 mid_school
 0.00486  17 attraction
 0.00266   4 kindergarden
 0         7 high_school
 0         8 universe
```

```
Ranked attributes:
 0.22628  16 bank
 0.20836  17 attraction
 0.17969  14 transport
 0.15354   9 franchise
 0.15076  15 hospital
 0.14402   7 high_school
 0.07952   3 apartment
 0.07897  12 convenience
 0.04713   6 mid_school
 0.04485  13 park
 0.01508   1 public
 0.01041  10 religion
 0.00388  11 mart
 0.00242   5 ele_school
 0.00203   4 kindergarden
 0.00105   2 factory
 0         8 universe
```

각각의 GainRatio 값은 다음과 같으며 앞서 말한바와 같이 점수를 부여하여 등수를 메기면 다음과 같습니다.

1	hospital (157)	10	park (78)
2	bank (151)	11	factory (77)
3	franchise (144)	12	ele_school (75)
4	convenience (118)	13	mart (60)
5	transport (118)	14	religion (59)
6	apartment (112)	15	attraction (52)
7	public (97)	16	kindergarden (47)
8	high_school (84)	17	universe (22)
9	mid_school (78)		

<오버 샘플링 후 변수 중요도 등수>

```
Ranked attributes:
 0.0319736  15 hospital
 0.0265201  16 bank
 0.0187507   9 franchise
 0.0126953  12 convenience
 0.0103316   7 high_school
 0.0096387   8 universe
 0.0093207   2 factory
 0.0087003   3 apartment
 0.0058683  14 transport
 0.0049679   1 public
 0.0023306  11 mart
 0.0023025  10 religion
 0.002184    6 mid_school
 0.0012428   4 kindergarden
 0.0005024   5 ele_school
 0.0002475  13 park
 0.0000397  17 attraction
```

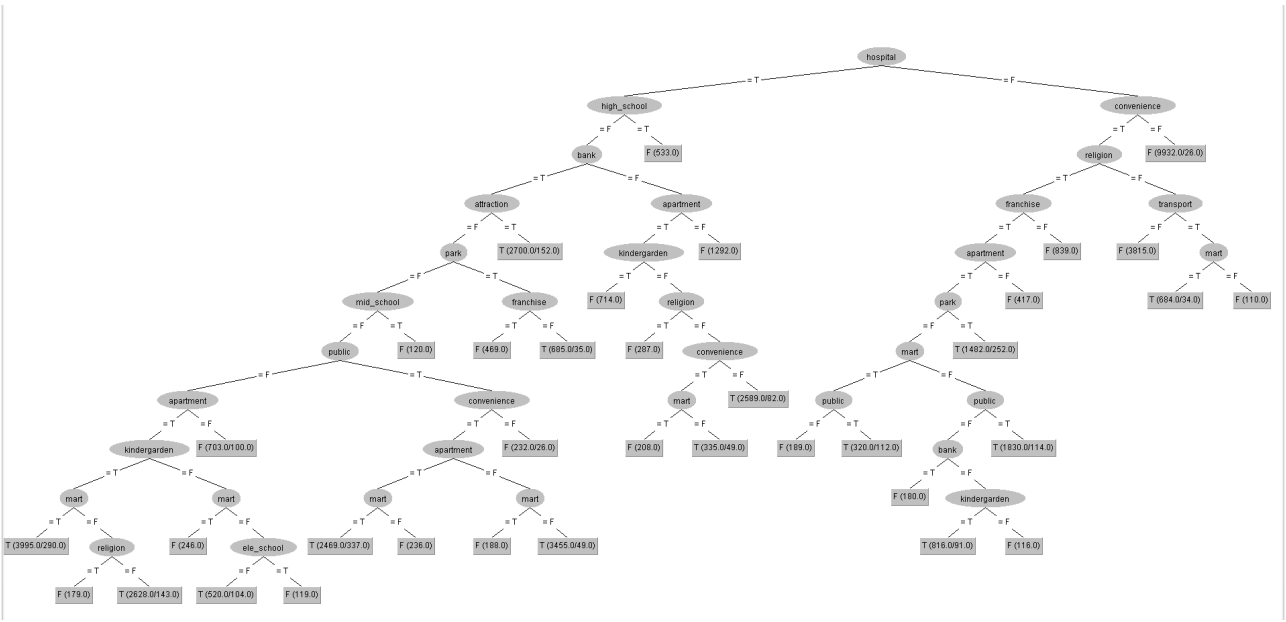
<오버 샘플링 이전 변수 중요도>

위 결과를 바탕으로 샘플링 이전의 데이터 셋과 비교해 보면 hospital, transport, bank, apartment 등 샘플링에서 중요도가 높던 변수들이 샘플링 이전의 데이터와 비슷하다는 것을 알 수 있습니다.

```
Ranked attributes:
 0.199514  15 hospital
 0.159131  16 bank
 0.158656   7 high_school
 0.149751   8 universe
 0.135329   9 franchise
 0.101795   2 factory
 0.101793  12 convenience
 0.063753   3 apartment
 0.029621   1 public
 0.024914  14 transport
 0.020918   6 mid_school
 0.015247  10 religion
 0.014437  11 mart
 0.007489   4 kindergarden
 0.003404   5 ele_school
 0.001784  13 park
 0.000222  17 attraction
```

<언더 샘플링 후 변수 중요도>

언더 샘플링의 경우도 마찬가지로 샘플링 이전의 데이터 셋과 비교해 보면 높은 중요도의 변수가 비슷하다는 것을 알 수 있었습니다. 따라서 T의 값을 가지는 경우를 명확하게 찾기 위하여 언더 샘플링을 한 데이터로 결과 분석을 하였습니다.



위 사진은 언더샘플링 이후의 의사결정 나무이며 가장 높은 곳의 변수는 hospital입니다. hospital가 있는 경우 T의 값이 더 많은 것을 알 수 있었으며 고등학교의 경우 없는 경우가 T의 값이 더 많다는 것을 알 수 있었습니다. 또한 은행의 경우 또한 있는 경우가 T의 비율을 많이 포함하고 있었으며 은행이 없더라도 아파트가 있고 유치원과 편의점이 없는 경우 2500여건 정도의 T값이 존재하였습니다. 마트의 경우에는 장을 많이 보기에 자가용을 이용하여 버스 사용량이 적은 것으로 판단됩니다. 하지만 아파트가 함께 있는 경우에는 버스를 이용하는 것으로 판단됩니다.

본 과제에서는 데이터 마이닝 기법 중 하나인 의사결정나무 알고리즘을 이용하여 압전 하베스팅 버스 정류장 설치 여부 예측 모델을 구축하였습니다.

종합적인 결과를 살펴보면 흔히 예상한 바대로 은행, 병원, 공공기관 등을 주변에 두고 있는 버스 정류장의 경우 버스 이동량이 많다는 걸 알 수 있었습니다. 하지만 마트나 특정 상황의 경우에는 자가용을 이용하는 편리성 등으로 인하여 예외의 경우도 생긴다고 판단하였습니다. 이러한 정보들을 바탕으로 타지역의 버스 정류장 또한 주변 시설들을 조사하여 적용시킨 다면 좋은 결과를 얻을 수 있을 것입니다.

3) 기대효과

전력사용에 따른 전기세 절감은 미미할 수 있지만 환경적인 측면으로 봤을 때는 작지 않은 변화라고 생각합니다. 하나의 정류장이 한달에 약 156kW를 절약한다면 포항의 경우 31개의 정류장이 설치 대상이며 4,836kW를 절감할 수 있으며 더 큰 도시 또는 버스 이용이 활발한 도시의 경우에는 더 많은 전기를 절감할 수 있을 것이라고 생각합니다.