**Report**

**데이터 마이닝 프로세스 이해하기**

**빅데이터 학과**

**20175327 이연희**

**1) 데이터 마이닝 프로세스의 3가지 방법론을 설명하세요.**

* **KDD (Knowledge Discovery in Database)**

KDD 프로세스는 5단계로 이루어져있다.

* + Selection(데이터set 선택)단계

이 단계에서는 비즈니스이해/프로젝트목표설정, 데이터베이스로부터 필요한 데이터 선택, 데이터 모델링에 필요한 데이터 set 생성 등의 처리를 한다.

* + Preprocessing(데이터 전 처리)단계

: 이 단계에서는 잡음(Noise), 이상치(Outlier), 누락 값(Missing Value)을 식별하여 정제 및 처리를 수행한다.

* + Transformation(데이터변환)단계

: 이 단계에서는 데이터 분석에 필요한 변수생성/선택/차원축소, 모델링을 위한 학습 데이터, 테스트 데이터 분리를 수행한다.

* + Data Mining(데이터 마이닝)단계

: 이 단계에서는 학습 데이터를 이용하여 분석 목적에 맞는 데이터 마이닝 기법을 이용하여 모델링을 수행한다.

* + Interpretation/Evaluation(결과 해석 및 평가)단계

: 데이터 마이닝 결과를 해석하고 평가, 분석 목적과의 일치성 확인 등을 수행한다.

* **CRISP-DM (Cross Industry Standard Process for Data Mining)**

CRISP-DM(Cross Industry Standard Process for Data Mining)은 데이터 마이닝(Data Mining) 분야에서 일반적으로 알려진 프로세스다. 1996년 SPRIT 개발되어 2016 IBM에서 개선 확장했으며 비즈니스 데이터 분석을 위한 산업계 표준이다.



1. 비즈니스 이해(Business Understanding)
   1. Determine Business Objectives

: 비즈니스 목표 결정 - 프로젝트 계획 수립/비즈니스 성공기준 정의

* 1. Assess Situation

: 현재상황평가 - 자원목록(인력, 데이터, 컴퓨터HW/SW)/요구사항, 가정, 제약조건/위험성, 우연성/용어집/비용과 효과

* 1. Determine Data Mining Goals

: 기술관점 목표결정 - 데이터 마이닝 성공기준

* 1. Produce Project Plan

: 프로젝트 계획 수립 - 프로젝트 세부계획수립/도구 및 기법에 대한 초기평가

1. 데이터 이해(Data Understanding)
   1. Collect Initial Data

: 초기 데이터 수집 - 보고서

* 1. Describe Data

: 데이터 설명 - 수집데이터 특성 확인/데이터 특성 보고서

* 1. Explore Data

: 데이터 탐색 - 주요속성파악, 속성 사이의 관계, 간단한 통계분석/데이터탐색보고서

* 1. Verify Data Quality

: 데이터 품질 확인 - 데이터 검사/검증-오류, 누락 데이터 보고서

1. 데이터 준비(Data Preparation)
   1. Select Data

: 데이터 선택 - 데이터 마이닝 목표 연관성, 데이터 유형 제한 등에 대한 데이터 포함/제외 결정 이유 포함

* 1. Clean Data

: 데이터 정제 - 누락 값 처리(제외, 기본값, 추정)로 품질 확보

* 1. Construct Data

: 요구 데이터 구조화 - 기본 속성으로 파생된 속성/새로운 레코드 생성

* 1. Integrate Data

: 데이터 통합 - 여러 데이터베이스나 테이블을 통합, 테이블 병합/여러 레코드나 테이블로 통합

* 1. Format Data

: 데이터 형식 - 모델에 적합한 형식으로 수정

1. 모델링(Modeling)
   1. Select Modeling Techniques

: 데이터 모델링 기법 선택 – 모델링 기술/모델 가정

* 1. Generate Test Design

: 테스트 디자인 생성 – 품질/유효성 검사를 위한 계획(학습, 테스트, 유효성 검사 데이터 집합 분리)

* 1. Build Model

: 모델 생성 – 모델 파라미터 조정/생성모델 설명

* 1. Assess Model

: 모델 평가 – 모델의 정확성, 일반성 평가/파라미터 재조정

1. 평가(Evaluation)
   1. Generate Test Design

: 결과 평가 – 데이터 마이닝 결과 평가/모델의 비즈니스 목표 충족 여부 평가/모델 승인

* 1. Review Process

: 프로세스 검토 – 누락된 중요 요소나 작업에 대한 검토/품질보증관련 확인

* 1. Determine Next Steps

: 다음 단계 결정 – 평가 결과나 프로세스 검토 결과에 따라 다음 진행 방법 결정, 가능한 조치 목록/진행방법 결정

1. 개발(Deployment)
   1. Plan Deployment

: 배포 계획 – 평가 결과로 배포 전략 결정/계획 문서화

* 1. Plan Monitoring and Maintenance

: 모니터링 및 유지 관리 계획

* 1. Produce Final Report

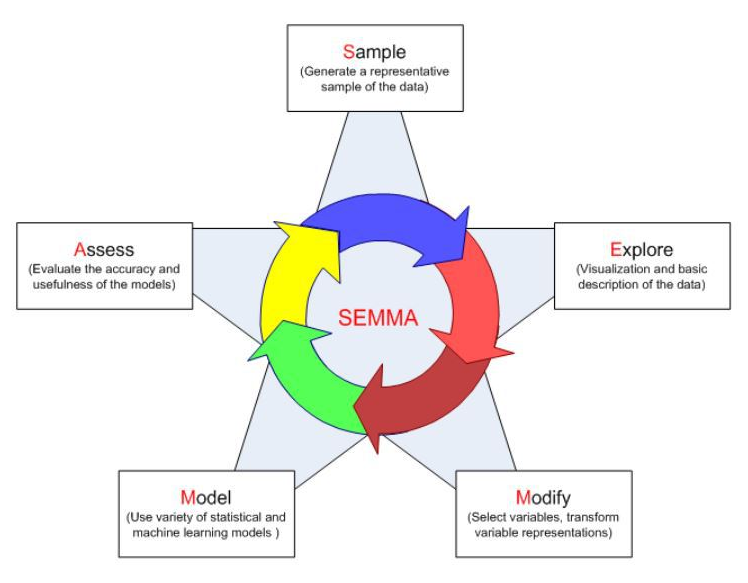
: 최종 보고서 작성 및 발표

* 1. Review Project

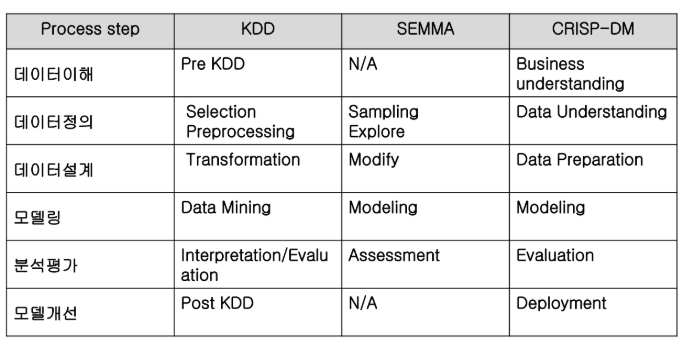
: 최종 보고서 검토

•**SEMMA (Sampling Explore Modify Model Assess)**

SEMMA: Sample, Explore, Modify, Model, Assess /SAS사에서 개발 "SAS Enterprise Miner"의 핵심 작업에 사용된다.



* Sampling: 데이터베이스로부터 데이터를 추출(extracting data)하고 샘플링(sampling) 입력 데이터를 생성, 병합, 분리 처리
  + 결과: 통계적 추출, 조건 추출
* Explore: 데이터의 기초통계, 시각화로 데이터의 오류 등 탐색, 데이터 클러스터, 연관성 탐색
  + 결과: 그래프, 기초통계, 변수유의성 및 상관분석, clustering
* Modify: 모델링을 위한 데이터 생성, 데이터 선택/제거/교체/차원감소/변환
  + 결과: 수량화, 표준화, 그룹화
* Modeling: 데이터 마이닝 기술을 적용하여 데이터 모델링, 모델구축 /패턴발견 /모델링과 알고리즘 적용
  + 결과: Neural Network, Decision Tree, Logistic Regression, 기타 통계 모델링 기법
* Assessment 모델의 유용성, 신뢰성 평가 및 검증
  + 결과: 모델비교, 의사결정, 점수화



**2) 데이터 마이닝 프로세스의 3가지 방법론의 각 단계별 유사성을 설명하세요.**

1. 결과에 대한 구체화 및 가시화

데이터 마이닝의 목적이 명확하고 최종적인 결과에 대해 구체적일 수록 데이터 마이닝의 효과가 크고 시행착오를 줄일 수 있을 것이다. 데이터 마이닝에서 가장 중요한 사항일 것이다.

1. 조직의 협조 및 방향성공유

데이터분석시 DBA, 모델, 협업담당자, 레거시시스템담당자, 의사결정권자, 빅데이터 분석가 등 다양한 인력의 긴밀한 협력 및 커뮤니케이션이 필요하다. 아울러 최종적으로 나올 그림과 방향성의 공유가 필요하다.

1. 서비스의 영향도 고려

데이터를 수집하거나 정제하거나 데이터와 관련된 프로세스에서 기존의 서비스에 영향을 최소화해야 할 것이다.

1. 데이터의 품질관점

데이터의 정제, 변환 시 데이터 품질의 고려가 필요하다. 데이터 포맷, Noise의 제거, 이상치의 모니터링, 결측 치에 대한 조치 등을 체크할 수 있는 기술적인 관점이 필요할 것이다.

1. 기술적 관점의 유연성

데이터 분석 모델을 적용하거나 데이터 변환시에 다양한 기술들이 적용될 수 있다.

**3) 강의자료 15쪽~20쪽은 주피터노트북(Jupyter Notebook) 개발환경에서 개발되고 실행되는 마크다운(markdown)과 파이썬 코드(python code)입니다.**

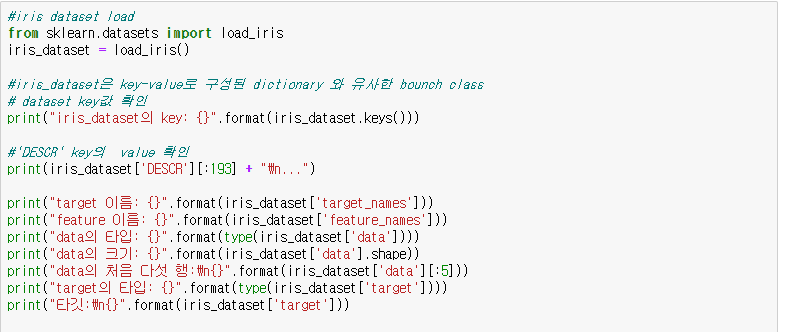
(1)주피터노트북(Jupyter Notebook) 개발환경을 사용해본 경험이 있는가?

네, 아나콘다 가상환경을 통해 주피터 노트북 개발환경을 사용해봤습니다.

(2)마크다운(markdown)이 무엇인가?

마크다운(markdown)은 일반 텍스트 문서의 양식을 편집하는 문법이다. README 파일이나 온라인 문서, 혹은 일반 텍스트 편집기로 문서 양식을 편집할 때 쓰인다. 마크다운을 이용해 작성된 문서는 쉽게 HTML 등 다른 문서형태로 변환이 가능하다. 또한 특수기호와 문자를 이용한 매우 간단한 구조의 문법을 사용하여 웹에서도 보다 빠르게 컨텐츠를 작성하고 보다 직관적으로 인식할 수 있다.

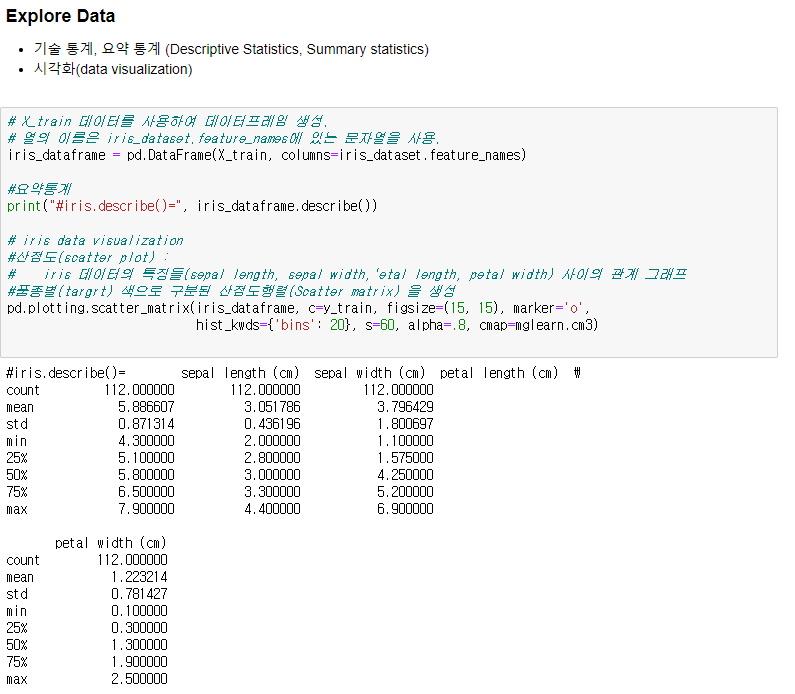
(3)파이썬 코드(python code)에서 이해가 된 부분이 있으면 설명하세요.



가져온 데이터를 분석하고 이해하는 과정이다.



1. 가져온 데이터 중에서 훈련에 쓰일 훈련용 데이터와 마지막 테스트에 쓰일 테스트용 데이터 셋을 분리한다. 해당 데이터는 112개의 훈련용 데이터와 38개의 테스트 데이터로 나누어졌다.



1. Pandas 라이브러리를 통해 pd.DataFrame으로 데이터 프레임을 생성해 붓꽃 데이터의 분석 결과를 열람한다.

다음으로는 시각화(data visualization)를 통해 산점도 그래프를 뽑아서 데이터 분포를 확인한다.



(3) k-최근접 이웃 알고리즘을 통해 훈련용 데이터인 Train Set을 모델링한다,

(4) 마지막으로 모델 예측 및 평가를 수행한다. 테스트 세트에 대한 예측 값과 이에 대한 예측 정확도를 확인한다.