Assignment 2

กำหนดส่ง วันอังคารที่ 9 กุมภาพันธ์ 2564

นิสิตแต่ละกลุ่มจะได้รับการกำหนดชุดข้อมูลคนละ 1 ชุด โดยดูรายละเอียดได้ในไฟล์รายชื่อ

โดยมี code ดังนี้ F = Fertility dataset, W = Wine dataset, T = Thyroid dataset, C = Contraceptive dataset, A = Autism dataset

คำสั่ง

1. Download ชุดข้อมูลจากเว็บ UCI machine learning repository
2. ทำความเข้าใจชุดข้อมูล ความหมายของแต่ละ feature
3. ระบุ target class
4. Preprocess dataset ในกรณีที่นิสิตคิดว่าจำเป็น
5. เขียนโปรแกรมภาษา Python โดยเรียกใช้ decision tree algorithm จาก scikit-learn
6. สิ่งที่นิสิตต้องส่งมอบคือ pdf file (submit เข้า google classroom) มีรายละเอียดดังนี้

   6.1 อธิบายความหมายของ feature และ target class

   6.2 อธิบายการทำ preprocess

   6.3 ไฟล์ที่บรรจุ python source code

   6.4 ผลการทดลอง ให้ระบุสัดส่วน train: test และให้แสดงค่า accuracy โดยทำการทดลองด้วยขั้นตอนวิธีเพื่อนบ้านใกล้สุด k ตัว และ Neural Network

   6.5 วิเคราะห์ผลการทดลองทั้ง 2 algorithms

รายการชุดข้อมูลสำหรับการทดลอง

Fertility Data Set

Download: Data Folder, Data Set Description

Abstract: 100 volunteers provide a semen sample analyzed according to the WHO 2010 crit
health status, and life habits

Data Set Characteristics:	Multivariate	**Number of Instances:**	100
Attribute Characteristics:	Real	**Number of Attributes:**	10
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A

Wine Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from
based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898
Attribute Characteristics:	Real	**Number of Attributes:**	12
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A

Thyroid Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 10 separate databases from Garavan Institute

Data Set Characteristics:	Multivariate, Domain-Theory	**Number of Instances:**	7200
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	21
Associated Tasks:	Classification	**Missing Values?**	N/A

Contraceptive Method Choice Data Set

Download: Data Folder, Data Set Description

Abstract: Dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence

Data Set Characteristics:	Multivariate	**Number of Instances:**	1473
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	9
Associated Tasks:	Classification	**Missing Values?**	No

Autism Screening Adult Data Set

Download: Data Folder, Data Set Description

Abstract: Autistic Spectrum Disorder Screening Data for Adult. This dataset is related to classification and predictive tasks

Data Set Characteristics:	N/A	**Number of Instances:**	704	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	21	**Date Donated**	2017-12-24
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	47664