IDS 575 – Machine Learning Statistics – Project Proposal

Group 15 – Anindita Mitra, Anisha Vijayan, Murtaza Agha, Yu Ting Sun

Online News Popularity

Problem Description/ Motivation

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict if the article is among the most popular ones based on sharing in social networks (coded by the variable "is_popular")

Dataset

Number of Instances	39797
Number of Attributes:	61
Data Set Characteristics:	Multivariate
Attribute Characteristics:	Integer, Real
Associated Tasks:	Classification
Area:	Business

Roles of the group member

1. Project Proposal – All members
2. Data Cleaning & Exploratory Data Analysis - Anisha Vijayan
3. Modeling
    a. KNN – Yu Ting Sun
    b. Logistic Regression - Anindita Mitra
    c. SVM – Anisha Vijayan
    d. Naive Bayes – Murtaza Agha
    e. Any additional modeling (Kmeans, PCA) – Subject to additional topics learnt throughout the course
4. Evaluation
    a. Model selection and assessment – Yu Ting Sun and Anisha Vijayan
    b. Comparing Performance across models – Anindita Mitra and Murtaza Agha
5. Presentation and Report – All members

Weekly schedule

Week 1 (Oct 10 - Oct 16): Data Cleaning and EDA.
Week 2 (Oct 17 - Oct 23): Naive Bayes and KNN. Test and result.
Week 3 (Oct 24 - Oct 30): Logistic Regression. Test and result.
Week 4 (Oct 31 - Nov 6): SVM. Test and result.
Week 5 (Nov 7 - Nov 13): Additional models. Test and result.
Week 6 (Nov 14 - Nov 20): Evaluation and Conclusions.
Week 7 (Nov 21 - Nov 27): Prepare for final presentation.
Week 8 (Nov 28 - Dec 6): Final presentation and final review for report.

References

Online News Popularity Data Set
https://www.kaggle.com/competitions/online-news-popularity/data
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity