CS770 Machine Learning

Assignment2: BMI Classification using Machine learning.

03/28/2024

Submitted by: Logan Schraeder (x356t577)

Abstract

This report investigates the application of machine learning algorithms, namely logistic regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), to classify Body Mass Index (BMI) using a dataset of 500 samples with four features. Results indicate that KNN achieved the highest accuracy (79%) on the full dataset, followed by SVM (70%) and logistic regression (59%). However, a significant decrease in performance was observed when the dataset was divided by sex, suggesting that the smaller training pool negatively impacted model accuracy. This highlights the importance of dataset size in the effectiveness of these classification algorithms for BMI prediction.

Introduction

This report details the application of machine learning techniques to classify Body Mass Index (BMI) using a dataset of 500 samples with four features each. Specifically, logistic regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) algorithms were employed for this classification task. The methodology included exploratory data analysis, data preprocessing, and the training of models on both the full dataset and subsets segregated by sex. The performance of each model was evaluated based on precision, recall, F1 score, and accuracy, with a particular focus on comparing the results obtained from the full dataset versus the sex-segregated datasets.

Methods

The methods used in assignment two were constrained to logistic regression, SVM, and KNN for classification tasks. As with any data product, exploratory data analysis and data preprocessing were performed at the beginning of the training pipeline. The BMI dataset was generally small - 500 total samples with four features each. The data was very evenly split along the sexes, but only showed significant correlations from BMI to weight and height. For each classification method across all and discrete sexes, the same dataset was used for training. A logistic regression, KNN, and SVM model was built for all sexes, and then specifically for male and female BMI predictions.

Fig. 1. BMI Correlation Matrix. Correlation matrix of the entire dataset. Note that the only significant correlations to BMI are height and weight.

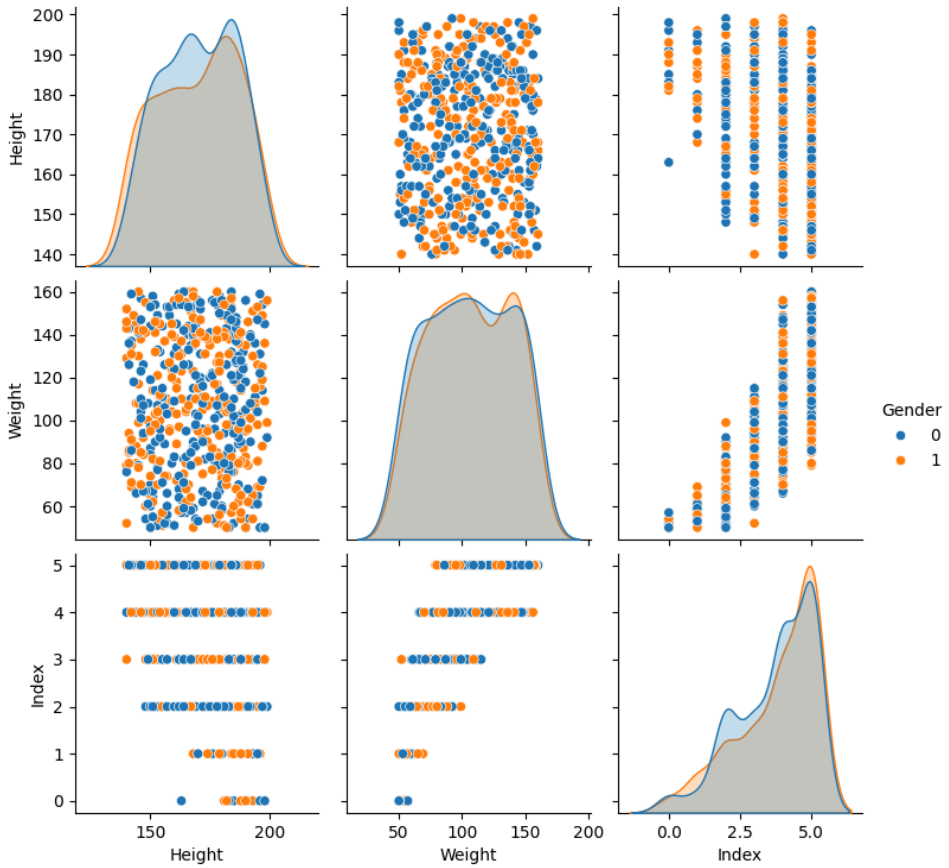

Fig. 2. Dataset Distributions. Pair plot of the dataset feature distributions. It is easily observed that height and weight have generally normal distributions across sexes while index is skewed.

Results and Discussion

When training classification on the full dataset KNN performed the best with 79% accuracy, followed by SVM (70%) and last by logistic regression (59%). This trend also generally held true across features when evaluating by F1 score. When the dataset was segregated by sex, performance fell drastically. Logistic regression and SVM both had accuracy scores of 39% and KNN only had 8% accuracy. It is hypothesized that the poor performance of the by-sex models is due to the very small training pool (about 250 samples) after splitting of the dataset. Table 1 contains the tabulated results.

Dataset	Algorithm	Precision	Recall	F1	Accuracy
Full	LogReg	0.53	0.59	0.54	0.59
Full	SVM	0.72	0.70	0.70	0.70
Full	KNN*	0.82	0.79	0.80	0.79
Male	LogReg	0.15	0.39	0.22	0.39
Male	SVM	0.15	0.39	0.22	0.39
Male	KNN*	0.01	0.08	0.01	0.08
Female	LogReg	0.37	0.58	0.44	0.58
Female	SVM	0.03	0.18	0.05	0.18
Female	KNN*	0.03	0.18	0.05	0.18

Table 1. Model Accuracy and Sensitivities. In the full dataset, KNN and SVM reliably outperformed logistic regression. However in the smaller datasets, logistic regression was more accurate (although still less than ⅔ of the time). *k = 5

Conclusions

In conclusion, while KNN initially showed the highest accuracy in BMI classification on the full dataset, the performance of all algorithms drastically decreased when the dataset was segregated by sex. This suggests that the size of the training pool significantly impacts model performance. Further research with a larger dataset and potentially more complex models may be necessary to improve the accuracy of by-sex BMI predictions.