MACHINE LEARNING (COMP7703)

Assignment - Semester 1, 2025.

1. Introduction

There are many different machine learning algorithms, and many metrics to evaluate them. In this assignment, we will focus on the K-Nearest Neighbour (K-NN) algorithm, Random Forest, and Deep Neural Networks, and the accuracy of each model. We will also looks at the different metrics to evaluate the performance of each model, and use these metrics to compare the suitability of each model for the given data.

2. Aims

- To understand the K-Nearest Neighbour (K-NN) algorithm.
- To understand the accuracy of the K-NN classifier.
- To understand the effect of data imbalance on the accuracy of the K-NN classifier.
- To understand how to use cross-validation to improve the accuracy of the K-NN classifier.

3. Data

The data used in this assignment is the `iris` dataset, which is a well-known dataset in the machine learning community. The dataset contains 150 data points, with 4 features and 3 classes. The features are the sepal length, sepal width, petal length, and petal width, and the classes are the species of iris flower: setosa, versicolor, and virginica. The dataset is available in the `sklearn` library. **3.1 Data split**
The data is split into three height categories: high, medium, and low. The data is split into 60% training data, 20% testing data, and 20% validation data. The training data is used to train the each model, the testing data is used to test the accuracy of each model, and the validation data is used to compare the performance of the models.

3.2 Data imbalance

After splitting the data, there is an imbalance in the amount of training data assigned to each category in height. There are 17364 (29.73%) data points in the high, 40800 (42.03%) data points in the medium, and 27450 (28.23%) in the low category. This is a significant imbalance, which is addressed by the use of Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic data points for the low category. The SMOTE algorithm generates synthetic data points by interpolating between existing data points in the low category. This is done by selecting a random data point from the low category, and then selecting a random data point from the k nearest neighbours of that data point. The synthetic data point is then generated by taking a weighted average of the two data points.

After applying SMOTE, the training data is balanced, with 24552 data points in each category.

4. Principle Component Analysis

To investigate the effect of dimensionality reduction on the accuracy of the models, we will use Principle Component Analysis (PCA) to reduce the dimensionality of the data. PCA is a linear transformation that transforms the data into a new coordinate system, where the first coordinate is the direction

of maximum variance, the second coordinate is the direction of maximum variance orthogonal to the first coordinate, and so on.

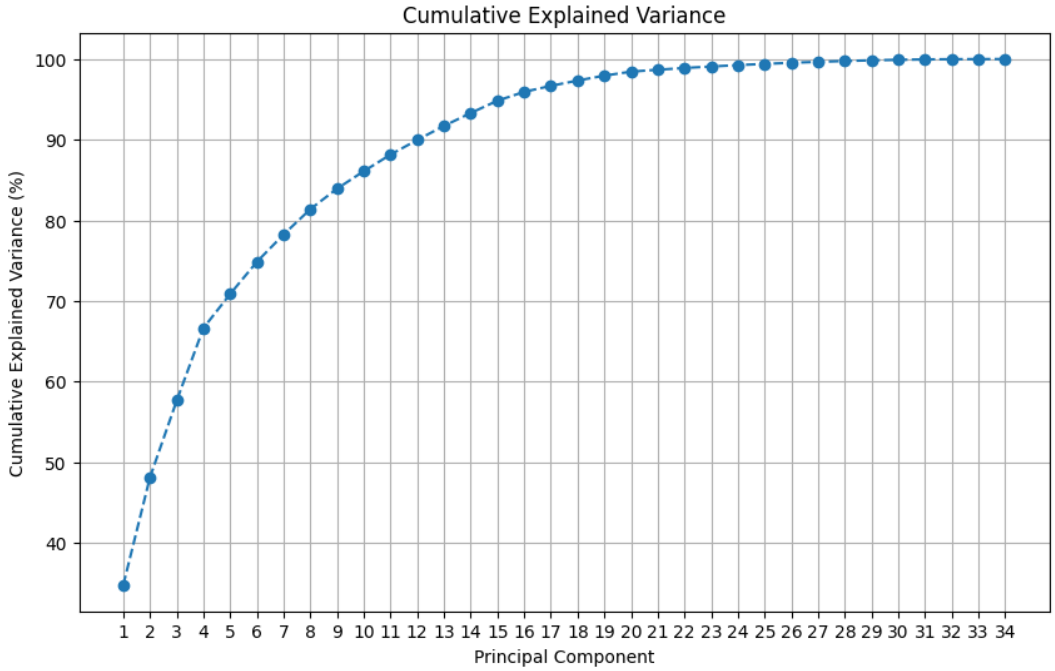

Figure 1: Cumulative variance explained by each principal component.

We see that approximately 95% of the variance is explained by the top 16 principal components. This means that we can reduce the dimensionality of the data from 33 dimensions to 16 dimensions, while still retaining most of the information in the data. We will use dimensionality reduction in our models to in hopes of imporving accuracy and/or reducing computational expense.

5. K-NN classifier

5.1 Accuracy of K-NN classifier

5.2 Cross-validation

6. Random Forest

7. Deep Neural Networks