Dimensionality reduction in the human activity classification problem

Objective of research

The main objective of this research is to determine if we can use a lower dimensional dataset in classifying human activities and achieve good results. The benefits of this is that we can reduce the computational time required to fit our models and reduce the effects of the “Curse of Dimensionality”. A fewer number of features also makes our model simpler and reduces chances of overfitting.

Data description

The dataset consists of 10299 rows and 562. This means we have 562 explanatory variables. The explanatory variables consist of information which is used to determine whether the person is laying, standing, sitting, walking, walking upstairs or walking downstairs.

Data cleaning and feature engineering

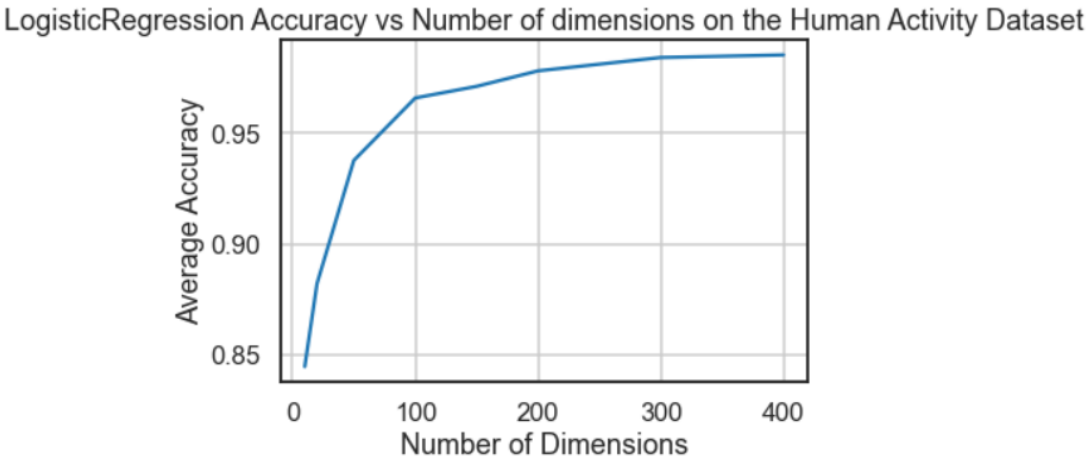
All the variables were checked to ensure that there were no missing values. There were indeed no missing values in the data.

Feature scaling was used to ensure that the variables are on a similar scale as this is important for dimensionality reduction.

Training the classification models models with reduced dimensions

Dimensionality reduction was performed on the dataset. It was reduced to 10, 20, 50, 100, 150, 200, 300 and 400 features using Principal component analysis and stored in separate datasets. 8 different logistic regression models were trained

using each of the 8 reduced datasets and the accuracy scores recorded. The figure below shows a plot of the accuracy achieved by the logistic regression model for each dataset.



Choice of final model

The final model is the one with a dataset reduced to 100 features. This model achieves an accuracy which is above 95% which is very good and which means there is no need for us to use the dataset with higher dimensions.

Summary of key findings

From the research we were able to find that we can reduce the dataset from 562 dimensions to 100 dimensions and still achieve very good accuracy in our prediction task yet using a simpler model.

Suggestion for next steps

We can try to use Kernel PCA and test if we can be able to achieve higher accuracy with our reduced datasets.