Dementia is an irreversible condition where cognitive processes and brain structure deteriorates rapidly and shortens lifespan significantly. It is most common in the elderly but can occur earlier in life. There are treatments that can slow the process but there is yet a cure. The aim of this research is to investigate whether age, socioeconomic status, gender and education can predict dementia. Therefore, it is hypothesised age, gender, socioeconomic status, and education can distinguish individuals with dementia and non-demented individuals.

**Method**

Participants

There were 373 participants including males and females and ages ranged from 60-96 years. All participants were right-handed. 190 were classed as non-demented, 127 were demented while 37 converted during the study.

Procedure

Python 3 using Jupyter notebook was used to explore the dataset. Pandas was used for data analysis and matplotlib was used for data visualization.

Materials

Dataset was downloaded from Kaggle as a csv file.

Data analysis

M/F column was recoded to integers, so females represented 0 and males represented 1. This was done considering the machine learning algorithm later used needed all features to be non-string. The features used in the analysis was socioeconomic status, gender, education, and age. Considering the spread of values varied between numeric features (e.g. education and age), the features were scaled to normalise the data. After checking for nan’s, 19 participants had an empty value for socioeconomic status, so they were excluded which brought the amount of data points down to 354. The target to predict was group which had 3 classes: converted, demented and non-demented.

The data was divided into a train and test split: the train split had 70% of the original data while the test set had 30% of the original data. The target was stratified so the proportions of each class was the same as the full dataset. KNN using 5 neighbours as the hyperparameter was used to fit the training data and the Minkowski distance was used. K-fold cross validation and grid searching was used later to get a more accurate metric of the data and finding optimal neighbours to use.

**Results**

|  |  |
| --- | --- |
| Model | Accuracy |
| KNN with 5 neighbours, train/test split | 59% |
| KNN with 5-fold cross validation, 3 neighbours. Average of all 5 folds. | 48% |
| KNN with 7 neighbours (found optimal with grid search) | 74% |

The first Knn model using train/test split with 5 neighbours resulted in 59% classification accuracy while the second model using 5-fold cross validation with 3 neighbours had 48% classification accuracy averaged over the 5 folds. After using grid searching and finding 7 neighbours to be optimal for the knn model, the classification accuracy went up to 74%.

**Discussion**

Based on the results, the hypothesis that socioeconomic status, gender, age and education can collectively predict dementia was supported. After fine tuning the knn model, classification accuracy was 74% for distinguishing dementia from non-dementia patients. This suggests age, gender, education, and socioeconomic status all play a role in demented patients. Clinicians should use this finding as supplementary to diagnosis and treatment.