CPSC 5310 01 Machine Learning
Project_Part_1
Team 4 – Priyadarshini Shanmugasundaram Murugan, Archana Janardhana Vinoda Kumari, Priyanka Vysyaraju

Our topic

Predicting Alzheimer: Alzheimer's disease is a neurodegenerative disorder causing memory loss, cognitive decline, and behavioural changes. Diagnosis involves medical history, cognitive tests, and, in some cases, brain imaging. While there is no cure, ongoing research seeks to develop effective treatments. Predictive modelling using machine learning involves creating a model to predict the likelihood of Alzheimer's based on specific input features.

Dataset

We have chosen to work with the "predict_alzheimers" dataset, which contains various demographic and clinical features related to Alzheimer's disease. The dataset includes columns such as origprot, ptid, dxbl, dx, month, m, age, ptgender, pteducat, ptethcat, ptraccat, ptmarry, cdrsb.bl, adas11.bl, adas13.bl, mmse.bl, and several others.

Dataset Address: https://github.com/pranath/predict_alzheimers

We weren't able to find any previous work done on the dataset.

Preprocessing

Data preprocessing is a crucial step in preparing data for machine learning models. It involves understanding the data, addressing missing values and outliers, encoding categorical variables, scaling numerical features, and performing feature engineering. Additionally, it is important to split the data, address imbalanced classes, and process text data, as well as to normalize the data, check for data skewness, and keep detailed documentation of preprocessing steps. The specific steps taken depend on the characteristics of the dataset and the requirements of the machine learning task. It often involves an iterative process based on model performance. We plan to preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features.

Regression for Predicting Cognitive Decline

The task is to predict cognitive decline using machine learning regression techniques. We will target variables for regression include clinical measures like "cdrsb.bl," "adas11.bl," and "mmse.bl," which assess the severity of dementia and cognitive impairment. Demographic and clinical features serve as input variables. Regression models, such as linear regression or decision trees, are trained on labelled datasets to learn the relationship between input features and cognitive decline scores. We might evaluate model performance using metrics like mean squared error, and the trained model can provide insights into the factors influencing cognitive deterioration. Our ultimate goal is to apply the model to new data for early identification and intervention in individuals at risk of cognitive decline.

Classification for individual has Alzheimer's disease

The task is to classify whether an individual has Alzheimer's disease or not using machine learning for binary classification. Our steps include collecting a relevant dataset, exploring its structure, preprocessing by handling missing values and outliers, encoding variables, and splitting data into training and testing sets. We will choose an appropriate classification algorithm, train the model, and evaluate its performance using metrics. Optionally, we would like to tune hyperparameters for optimization, interpret model decisions and validate predictions.

Clustering by demographic and clinical features

We believe, utilizing demographic and clinical features can be beneficial in identifying subgroups of patients with similar characteristics by performing clustering, such as K-means clustering. This approach can facilitate an understanding of the various phenotypes within the dataset. Once the pre-processed dataset is obtained, we will apply the selected clustering algorithm to assign each patient to a specific cluster based on the similarity of their features.

Clustering can be employed to group patients with similar characteristics, which may include age, genetic factors, or specific biomarkers. This method can be instrumental in comprehending the heterogeneity of Alzheimer's patients and may have implications for devising effective treatment strategies.

Team Number 04

Predicting Alzheimer

Team Members:

- Priyanka Vysyaraju
- Priyadarshini Shanmugasundaram Murugan
- Archana Janardhana Vinoda Kumari

Categorical Numerical

origprot	ptid	dxbl	dx	month	m	age	ptgender	pteducat	ptethcat	ptraccat	ptmarry	cdrsb.bl	adas11.bl	adas13.bl	mmse.bl	ravlt.immediate.bl	ravlt.learning.bl	ravlt.forgetting.bl	ravlt.perc.forgetting.bl
ADNI1	137_S_0438	AD	AD	0	0	81.9	Male	11	Not Hisp/Latin	White	Married	4	15.67	29.67	25	19	1	4	100
ADNI2	116_S_4209	AD	AD	0	0	77.7	Female	19	Not Hisp/Latin	White	Married	4	20	32	23	38	4	8	100
ADNI1	109_S_0777	AD	AD	0	0	75.1	Female	7	Not Hisp/Latin	Black	Divorced	3.5	18.33	28.33	22	27	3	7	100
ADNI2	068_S_5206	AD	AD	0	0	85.2	Male	20	Not Hisp/Latin	Black	Married	2	21	32	22	23	4	5	100
ADNI1	013_S_0592	AD	AD	0	0	77.9	Male	19	Not Hisp/Latin	White	Married	3.5	23.67	35.67	23	16	2	5	100
ADNI1	041_S_1435	AD	AD	0	0	82.4	Male	12	Not Hisp/Latin	White	Married	1.5	17.67	27.67	21	28	3	7	100
ADNI2	068_S_5146	AD	AD	0	0	72.9	Female	18	Not Hisp/Latin	White	Married	2.5	13	23	25	29	1	7	100
ADNI2	070_S_4692	AD	AD	0	0	83.2	Male	18	Not Hisp/Latin	White	Married	7	18	28	23	24	1	5	100
ADNI1	067_S_0828	AD	AD	0	0	77.4	Female	16	Hisp/Latino	White	Widowed	7	25.33	35.33	21	22	2	4	100
ADNI2	018_S_5240	AD	AD	0	0	62.7	Female	20	Not Hisp/Latin	White	Married	2	18	27	20	34	5	10	100
ADNI2	099_S_4994	AD	AD	0	0	84.8	Male	14	Not Hisp/Latin	White	Married	8	37	49	21	9	1	3	100
ADNI1	109_S_1192	AD	AD	0	0	69.6	Female	14	Hisp/Latino	White	Widowed	5.5	17.33	26.33	20	26	2	6	100
ADNI1	006_S_0653	AD	AD	0	0	73.4	Female	12	Not Hisp/Latin	White	Married	6	21	30	20	30	6	5	55.55555556
ADNI2	082_S_5184	AD	AD	6	6	72.5	Female	17	Not Hisp/Latin	White	Married	4	28	41	26	14	2	4	100
ADNI1	114_S_0228	AD	AD	6	6	80.2	Female	10	Unknown	White	Widowed	8	28	36	20	19	0	4	100
ADNI2	100_S_5106	AD	AD	6	6	74.1	Male	16	Not Hisp/Latin	White	Married	4	30	42	20	18	2	4	100
ADNI1	016_S_1263	AD	AD	6	6	64.8	Female	12	Not Hisp/Latin	White	Married	3.5	21	31	26	23	1	5	100
ADNI2	130_S_5059	AD	AD	6	6	72.1	Male	18	Not Hisp/Latin	White	Married	5	33	46	21	18	0	3	100
ADNI2	009_S_5252	AD	AD	6	6	56.5	Male	14	Hisp/Latino	White	Married	8	17	26	25	21	3	5	83.33333333
ADNI2	009_S_5224	AD	AD	6	6	78	Male	20	Not Hisp/Latin	White	Married	5	18	29	25	28	1	5	100
ADNI1	032_S_1037	AD	AD	6	6	73.7	Male	19	Not Hisp/Latin	White	Married	3.5	16.67	23.67	24	27	0	5	100
ADNI2	073_S_5090	AD	AD	6	6	58.8	Male	16	Not Hisp/Latin	White	Married	2.5	23	37	22	14	2	3	100
ADNI2	131_S_5138	AD	AD	6	6	60.6	Male	19	Not Hisp/Latin	Asian	Married	2	14	21	25	39	6	5	50
ADNI1	009_S_1354	AD	AD	6	6	58.9	Female	16	Not Hisp/Latin	Black	Married	9	24.67	35.67	20	31	2	8	88.88888889

Categorical

Numerical

faq.bl	fldstreng.bl	fsversion.bl	ventricles.bl	hippocampus.bl	wholebrain.bl	entorhinal.bl	fusiform.bl	midtemp.bl	icv.bl	clabel	glabel	nlabel
13	1.5 Tesla MRI	Cross-Sectional	66806	5479	797415	3200	14503	15137	1424442	D2D	2	0
15	3 Tesla MRI	Cross-Sectional	43051	7108	1041218.336	2816	16336	18513	1540655	D2D	2	0
6	1.5 Tesla MRI	Cross-Sectional	24256	5942	1008441	3149	14357	17467	1495639	D2D	2	0
7	3 Tesla MRI	Cross-Sectional	58685	4755	926893.2592	2755	16814	16177	1447799	D2D	2	0
11	1.5 Tesla MRI	Cross-Sectional	47473	5555	917086	2628	14708	14369	1608304	D2D	2	0
5	1.5 Tesla MRI	Cross-Sectional	39279	6063	1024149	2734	18521	20620	1572214	D2D	2	0
15	3 Tesla MRI	Cross-Sectional	23917	6501	969908.4096	3026	15464	19578	1364094	D2D	2	0
18	3 Tesla MRI	Cross-Sectional	69218	5316	1122794.181	2223	18186	20353	1765062	D2D	2	0
27	1.5 Tesla MRI	Cross-Sectional	18088	4431	834853	1797	11777	16796	1302036	D2D	2	0
4	3 Tesla MRI	Cross-Sectional	14149	7694	1012762.795	3592	15906	20834	1373004	D2D	2	0
15	3 Tesla MRI	Cross-Sectional	58712	5575	976216.3654	2170	16839	14592	1570558	D2D	2	0
8	1.5 Tesla MRI	Cross-Sectional	17623	5738	929185	2352	12044	21035	1401870	D2D	2	0
26	1.5 Tesla MRI	Cross-Sectional	45381	6651	858926	3838	11071	14766	1459704	D2D	2	0
2	3 Tesla MRI	Cross-Sectional	24473	5217	909414.9668	2928	13254	15569	1358748	D2D	2	0
24	1.5 Tesla MRI	Cross-Sectional	98380	4393	916457	1467	14562	15685	1584242	D2D	2	0
13	3 Tesla MRI	Cross-Sectional	78940	5629	1096041.686	2725	15164	16912	1744529	D2D	2	0
13	1.5 Tesla MRI	Cross-Sectional	40869	4833	922174	2465	15546	17099	1381135	D2D	2	0
11	3 Tesla MRI	Cross-Sectional	57270	6118	1179554.403	3857	16863	18786	1785450	D2D	2	0
19	3 Tesla MRI	Cross-Sectional	44247	7301	1177296.303	3111	17946	18920	1672305	D2D	2	0
12	3 Tesla MRI	Cross-Sectional	62632	5411	998950.1263	1757	15615	16004	1587284	D2D	2	0
15	1.5 Tesla MRI	Cross-Sectional	66420	6821	1074541	3051	19069	22163	1767411	D2D	2	0
0	3 Tesla MRI	Cross-Sectional	27620	6386	1052593.28	2655	17593	19932	1552543	D2D	2	0
4	3 Tesla MRI	Cross-Sectional	18942	8049	1040811.877	3684	17390	21747	1472156	D2D	2	0
22	1.5 Tesla MRI	Cross-Sectional	21396	4309	813683	2201	11941	11332	1280312	D2D	2	0

Features and Labels

- Features: origprot, ptid, dxbl, dx, month, m, age, clabel, glabel, nlabel, hippocampus.bl, wholebrain.bl, entorhinal.bl,fusiform.bl, adas11.bl, adas13.bl, mmse.bl and several others.
- ptid, ptmarry, ptgender ,ptethcat , ptraccat and fsversion.bl column cannot be used for analysis.
- Classify based on whether the individual has Alzheimer's disease or not.
- Perform clustering based on demographic and clinical features.
- Conduct regression for predicting cognitive decline.