**Idea N1 :** Make a very highly accurate model predict the stage of Alzheimers

The data consists of MRI images. The data has four classes of images including Mild Demented, Moderate Demented, Non Demented, Very Mild Demented

**Website:** <https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images>

**-------------------------------------------------------------------------------------------------------------------------------**

**Idea N2:** Human Activity Recognition Using Smartphones Data Set

**Website:** <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

Data is from a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (walking, walking\_upstairs, walking\_downstairs, sitting, standing, laying) wearing a smartphone on the waist. Using its embedded accelerometer and gyroscope, 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz were captured.

**-------------------------------------------------------------------------------------------------------------------------------**

**Idea N3:** Distinguish presence of heart disease and severity (values 1,2,3,4) from absence (value 0).

**Website:** <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Database contains 76 attributes from four sites. Distribution of data in different sites is as follow:

Database: 0 1 2 3 4 Total

Cleveland: 164 55 36 35 13 303

Hungarian: 188 37 26 28 15 294

Switzerland: 8 48 32 30 5 123

Long Beach VA: 51 56 41 42 10 200

**-------------------------------------------------------------------------------------------------------------------------------**

# **Idea N4:** Wikipedia - Image/Caption Matching

Train the model to associate given images with article titles or complex captions, in multiple languages

**Website:** <https://www.kaggle.com/c/wikipedia-image-caption>

**-------------------------------------------------------------------------------------------------------------------------------**

# **Idea N5:** Label famous, and not-so-famous, landmarks in images

Build models that recognize the correct landmark (if any) in a dataset of challenging test images. Test set is sampled from many countries, increasing the diversity in worldwide representation. The training data comes from a cleaned version of the Google Landmarks Dataset v2 (GLDv2),

**Website:**  <https://www.kaggle.com/c/landmark-recognition-2021>

**-------------------------------------------------------------------------------------------------------------------------------**

## **Idea N6:** Predict the status of a genetic biomarker important for brain cancer treatment

**Website:**  <https://www.kaggle.com/c/rsna-miccai-brain-tumor-radiogenomic-classification>

Predict the genetic subtype of glioblastoma using MRI (magnetic resonance imaging) scans to train and test model to detect for the presence of MGMT promoter methylation. Data include 585 scan for training and 87 scan for test data. Each data set has four different scans including Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted pre-contrast (T1w), T1-weighted post-contrast (T1Gd), T2-weighted (T2).

**-------------------------------------------------------------------------------------------------------------------------------**

# **Idea N7**: Use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

The dataset includes time series of features **store\_nbr**, **family**, and **onpromotion** as well as the target **sales**. **store\_nbr** identifies the store at which the products are sold. **family** identifies the type of product sold. **sales** gives the total sales for a product family at a particular store at a given date. **Onpromotion** gives the total number of items in a product family that were being promoted at a store at a given date.

**Website:** <https://www.kaggle.com/c/store-sales-time-series-forecasting>

**-------------------------------------------------------------------------------------------------------------------------------**

# **Idea N8:** Probability of depression, by age group and sex, household population aged 12 and over, selected provinces, territories and health regions

The dataset contains 94080 series, with data for years 2003 – 2003. The data has the different features Geography, Age , Sex, Probability of depression, and Characteristics

**Website:**<https://open.canada.ca/data/en/dataset/c1d55747-2b43-4ab4-95aa-3e5b9448ed30/resource/146fdb39-4482-43d1-8055-ea67c7dd92e3>

**-------------------------------------------------------------------------------------------------------------------------------**

**Idea N9:** Use machine learning for depression states classification, MADRS score (used in psychiatry) prediction based on motor activity data , Sleep pattern analysis of depressed v.s. non-depressed participants, or develop systems capable of automatically detecting depression states based on sensor data.

The dataset contains two folders, whereas one contains the data for the controls and one for the condition group. For each patient a csv file has been provided containing the actigraph data collected over time. The columns are: timestamp (one minute intervals), date (date of measurement), activity (activity measurement from the actigraph watch). In addition, the MADRS scores provided in another csv file. It contains the following columns; number (patient identifier), days (number of days of measurements), gender (1 or 2 for female or male), age (age in age groups), afftype (1: bipolar II, 2: unipolar depressive, 3: bipolar I), melanch (1: melancholia, 2: no melancholia), inpatient (1: inpatient, 2: outpatient), edu (education grouped in years), marriage (1: married or cohabiting, 2: single), work (1: working or studying, 2: unemployed/sick leave/pension), madrs1 (MADRS score when measurement started), madrs2 (MADRS when measurement stopped).

**Website:** <https://www.kaggle.com/arashnic/the-depression-dataset>

**-------------------------------------------------------------------------------------------------------------------------------**

## **Idea N10:** Predict real crypto market data

Use machine learning to forecast short term returns in 14 popular cryptocurrencies. Dataset include millions of rows of high-frequency market data dating back to 2018.

**Website:** <https://www.kaggle.com/c/g-research-crypto-forecasting/overview/code-requirements>