

[Home](#) > [Project ideas](#) > Kaggle dataset projects

◆ PROJECT IDEAS · MACHINE LEARNING · DEEP LEARNING

Machine learning projects with *Kaggle datasets*

These 25 machine learning and deep learning projects use public Kaggle datasets, from brain tumour MRI and skin cancer to fake news, stock prices and crop recommendation. Each card gives the **problem, the output, the dataset link and its licence**. All suit B.Tech as they stand; for M.Tech, each says what to change so it is not a repeat.

 Last updated 2026-10-19

Every dataset below was checked on Kaggle, and the licence shown is the one Kaggle listed on 19 October 2026. Licences can change, so read the dataset page again before you submit.

Which machine learning projects with Kaggle datasets are on this page?

The table lists all 25 with the main dataset, the licence Kaggle shows and the change that makes each one fit an M.Tech. Every project suits B.Tech as it stands. Tap a project name to open its card.

25 Kaggle dataset projects, with the dataset, its licence and the M.Tech change

PROJECT	KAGGLE DATASET	LICENCE ON KAGGLE	B.TECH SIZE	FOR M.TECH, CHANGE
Brain tumour MRI	Brain Tumor MRI Dataset	CC BY 4.0	Major	Train on one source, test on another
Skin cancer	HAM10000	CC BY-NC-SA 4.0	Major	Vision Transformer, outside ISIC test

PROJECT	KAGGLE DATASET	LICENCE ON KAGGLE	B.TECH SIZE	FOR M.TECH, CHANGE
<u>Breast cancer</u>	Wisconsin (Diagnostic) or Breast Histopathology Images	CC BY-NC-SA 4.0 or CCO	Mini or major	Patient-wise split, stain normalisation
<u>Pneumonia X-ray</u>	Chest X-Ray Images (Pneumonia)	In the description	Major	X-rays from a second source
<u>Diabetic retinopathy</u>	Diabetic Retinopathy 224x224 (from APTOS 2019)	CCO	Major	Full-size images, ensemble, second dataset
<u>Alzheimer's MRI</u>	OASIS Alzheimer's Detection	Apache 2.0	Major	Patient split vs slice split, 3D volumes
<u>Parkinson's</u>	Parkinson's Disease Data Set or Parkinson's Drawings	Open Database or CC BY-NC-ND 4.0	Mini or major	Person-wise folds, both drawing types
<u>Heart disease</u>	Heart Disease Dataset	Unknown	Mini	Test on another hospital, calibration
<u>Diabetes</u>	Diabetes Health Indicators Dataset	CCO	Mini	Imbalance, fairness, calibration
<u>Plant disease</u>	PlantVillage Dataset	CC BY-NC-SA 4.0	Major	Field photos, domain adaptation
<u>Sign language</u>	Sign Language MNIST or ASL Alphabet	CCO or GPL 2	Major	Hand landmarks, words from video
<u>Number plate</u>	Car License Plate Detection	CCO	Major	Indian plates, night, small board
<u>Face mask</u>	Face Mask Detection	CCO	Mini or major	Change the problem itself
<u>Driver drowsiness</u>	Drowsiness Detection Dataset (MRL)	CCO	Major	Night, infrared, Raspberry Pi
<u>Traffic sign</u>	GTSRB	CCO	Major	Robustness, Indian signs
<u>Emotion recognition</u>	FER-2013	Open Database	Major	Noisy labels, second dataset
<u>Waste classification</u>	Garbage Classification	Data files © Original Authors	Major	Indian waste photos, detection
<u>Fake news</u>	Fake and real news dataset	CC BY-NC-SA 4.0	Major	Cross-dataset test, source clues removed
<u>Sentiment analysis</u>	IMDB 50K Movie Reviews or Sentiment140	In the description	Mini or major	Hinglish, aspects, reviews to tweets
<u>Spam SMS</u>	SMS Spam Collection Dataset	Unknown	Mini	Recent Indian scam texts
<u>Stock price</u>	NIFTY-50 Stock Market Data	CCO	Major	Walk-forward test, costs, news

PROJECT	KAGGLE DATASET	LICENCE ON KAGGLE	B.TECH SIZE	FOR M.TECH, CHANGE
House price	House Prices competition or California Housing Prices	Competition rules or CC0	Mini	Indian city data, price ranges
Crop recommendation	Crop Recommendation Dataset	Apache 2.0	Mini	Real soil and yield data, profit
Credit card fraud	Credit Card Fraud Detection (ULB)	Open Database	Mini or major	Time-ordered split, retraining
Customer churn	Telco Customer Churn	Data files © Original Authors	Mini	Uplift modelling, cost threshold

“Mini” fits one model in a notebook with a simple page; “major” fits a full app with two or three models compared. Your college’s rules decide what counts as each.

What does each Kaggle dataset project build and show?

Each card states the problem and the output, the two lines your synopsis starts with. Open “Dataset and level” for the dataset link, its licence, and what to change for M.Tech.



Topic ideas, not delivered projects

Every card below is an idea to discuss with your guide. Our [8 delivered case studies](#) are M.Tech and M.E. projects, each shown with its real paper pages, screens and results.

Medical imaging projects

Medical images are the most repeated Kaggle projects of all. Split by patient or lesion, not by image, or the same person ends up in both training and test data.

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Brain tumour classification from MRI**

Problem: Reading brain MRI scans takes a radiologist's time, and a second opinion that sorts scans by tumour type can help decide which to read first.

Output: A CNN that labels an MRI slice as glioma, meningioma, pituitary tumour or no tumour, with a Grad-CAM heatmap and recall for each class on the held-out test folder.

 Dataset and level

Dataset: [Brain Tumor MRI Dataset](#) by Masoud Nickparvar; Kaggle lists the licence as CC BY 4.0. Its description says it combines the figshare, SARTAJ and Br35H sets.

For B.Tech / for M.Tech: B.Tech: suits as-is, with a pre-trained CNN and a simple upload page. M.Tech: because the set merges three sources, do not use SARTAJ as an outside test. Remove near-duplicate images first, then train on one source and test on another, and report the drop.

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Skin cancer classification on HAM10000**

Problem: Melanoma is easy to miss among harmless moles, and dermatologists are hard to reach outside cities.

Output: A model that sorts a dermatoscope image into the seven HAM10000 lesion types, with recall on melanoma reported on its own and a heatmap on each image.

 Dataset and level

Dataset: [Skin Cancer MNIST: HAM10000](#), CC BY-NC-SA 4.0 on Kaggle. Some lesions have several images, so split by the lesion_id column in the metadata file.

For B.Tech / for M.Tech: B.Tech: a fine-tuned CNN with class weights, as-is. M.Tech: compare a Vision Transformer with two CNNs on the same split, and test on an ISIC challenge set the model never saw. See our delivered [HAM10000 Vision Transformer case study](#) and the [skin disease model that runs in the browser](#).

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Breast cancer detection from cell measurements or histopathology**

Problem: A biopsy has to be judged benign or malignant, and on a slide the cancer regions must be found by eye among thousands of small areas.

Output: Either a classifier on the 30 Wisconsin cell measurements with the main features explained, or a CNN that marks 50×50 histopathology patches as IDC positive or negative.

– Dataset and level

Dataset: [Breast Cancer Wisconsin \(Diagnostic\)](#) (CC BY-NC-SA 4.0 on Kaggle) or [Breast Histopathology Images](#) (CC0: Public Domain), patches cut from 162 whole slides.

For B.Tech / for M.Tech: B.Tech: the Wisconsin table suits a mini project and the patches a major one. M.Tech: use the images, split by patient rather than by patch, and add stain normalisation or a result for each whole slide.

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Pneumonia detection from chest X-rays**

Problem: Where radiologists are few, chest X-rays wait to be read, and pneumonia cases need attention first.

Output: A classifier that labels an X-ray as normal or pneumonia, with sensitivity and specificity on the test folder and a heatmap over the lungs.

– Dataset and level

Dataset: [Chest X-Ray Images \(Pneumonia\)](#): 5,863 images in train, test and val folders. Kaggle says the licence is in the description, which links the source paper. The val folder is very small, so make your own validation split.

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: test the model on chest X-rays from a second source, and check whether the heatmap falls on the lungs or on text and markers at the edge of the film.

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Diabetic retinopathy grading from retina photos**

Problem: Diabetic retinopathy causes no symptoms at first, and screening every diabetic patient needs more eye specialists than most places have.

Output: A model that grades a retina photo from 0 (no DR) to 4 (proliferative), scored with quadratic weighted kappa, with a heatmap on each photo.

– Dataset and level

Dataset: [Diabetic Retinopathy 224x224 Gaussian Filtered](#) (CC0: Public Domain), resized from the [APTOS 2019 Blindness Detection](#) competition, whose rules cover the original images.

For B.Tech / for M.Tech: B.Tech: the resized set and a pre-trained CNN, as-is. M.Tech: train on the full-size APTOS images, compare a Vision Transformer or an ensemble, and test on a second retina dataset. See our delivered [diabetic retinopathy Vision Transformer ensemble](#).

TOPIC IDEA · DEEP LEARNING ·
MEDICAL IMAGING**Alzheimer's stage classification from brain MRI**

Problem: Early changes from Alzheimer's disease are hard to see on an MRI, and finding them early gives families time to plan care.

Output: A model that sorts MRI slices into four stages of dementia, with results reported per patient and the mix-ups between neighbouring stages shown.

– Dataset and level

Dataset: [OASIS Alzheimer's Detection](#), Apache 2.0 on Kaggle: 2D slices from the [OASIS](#) MRI scans of 461 patients. Cite OASIS as well.

For B.Tech / for M.Tech: B.Tech: as-is, but split by patient, since many slices come from each person. M.Tech: show the gap between a split by slice and a split by patient, then use 3D volumes or a vote across each patient's slices.

Health records and signals

TOPIC IDEA · MACHINE LEARNING · HEALTH DATA

Parkinson's disease detection from voice or spiral drawings

Problem: Parkinson's is diagnosed by a doctor from symptoms, and changes in voice and handwriting are among the signs a screening tool could pick up.

Output: Either a classifier on voice measurements or a CNN on spiral and wave drawings, with results reported per person, and a page that tests a new drawing.

⊖ Dataset and level

Dataset: Voice: [Parkinson's Disease Data Set](#), 195 recordings from 31 people (Kaggle lists the Open Database licence). Drawings: [Parkinson's Drawings](#) (CC BY-NC-ND 4.0).

For B.Tech / for M.Tech: B.Tech: either set, as-is. M.Tech: both are small, so use cross-validation that keeps each person in one fold, combine the spiral and wave drawings, and report a range rather than one accuracy. See our delivered [Parkinson's detection from spiral drawings](#) case study.

TOPIC IDEA · MACHINE LEARNING · HEALTH DATA

Heart disease prediction

Problem: A doctor screening many patients needs a quick way to flag who should have further heart tests.

Output: A risk score from routine clinical values such as age, blood pressure, cholesterol and ECG results, with the reasons for each prediction and a simple form.

⊖ Dataset and level

Dataset: [Heart Disease Dataset](#), from the 1988 Cleveland, Hungary, Switzerland and Long Beach records; Kaggle shows the licence as Unknown, so cite the original UCI source.

Another option is the [Heart Failure Prediction Dataset](#).

For B.Tech / for M.Tech: B.Tech: a mini project, as-is. Check for repeated rows before you split. M.Tech: train on one hospital's records and test on another's, and report calibration as well as accuracy.

**TOPIC IDEA · MACHINE LEARNING ·
HEALTH DATA****Diabetes risk prediction**

Problem: Many people with diabetes do not know they have it, and a short questionnaire could tell who should get a blood test.

Output: A classifier on survey answers such as blood pressure, BMI, activity and diet that ranks people by risk, with recall and precision at a chosen threshold.

⊖ Dataset and level

Dataset: [Diabetes Health Indicators](#)

Dataset (CC0: Public Domain): cleaned answers from the US CDC's BRFSS 2015 survey, in three files including a balanced 50–50 one.
For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: handle the imbalance in the full file and compare with the balanced file, check fairness across age and income groups, and calibrate the scores.

Vision, agriculture and safety projects

TOPIC IDEA · DEEP LEARNING · COMPUTER VISION

Plant disease detection from leaf photos

Problem: Farmers often notice crop disease late, and expert advice is far away.

Output: A model that names the crop and the disease from a leaf photo, running on a phone, tested on field photos as well as the dataset's own images.

⊖ Dataset and level

Dataset: [PlantVillage Dataset](#) (CC BY-NC-SA 4.0 on Kaggle); the [New Plant Diseases Dataset](#) is an augmented copy.

For B.Tech / for M.Tech: B.Tech: as-is, with MobileNet on a phone. M.Tech: PlantVillage leaves are photographed on a plain background, so a model trained on them can do much worse in a field. Measure that on field photos and try domain adaptation.

TOPIC IDEA · DEEP LEARNING · COMPUTER VISION

Sign language recognition from hand images

Problem: People who sign cannot easily talk with people who do not know sign language.

Output: A model that reads American Sign Language letters from a webcam in real time, with accuracy for each letter on images you record yourself.

⊖ Dataset and level

Dataset: [Sign Language MNIST](#) (CC0: Public Domain; 24 letters, without J and Z, which need motion) or [ASL Alphabet](#) (GPL 2; 29 classes, including space, delete and nothing).

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: use hand landmarks from a tool such as MediaPipe instead of raw pixels, add the moving letters or whole words from video, or build a small Indian Sign Language set.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Number plate detection and recognition**

Problem: Parking lots and campus gates log vehicles by hand, which is slow and full of mistakes.

Output: A pipeline that finds the plate in a photo with a detector, reads the characters with OCR, and logs the number with the time.

– Dataset and level

Dataset: [Car License Plate Detection](#) (CC0: Public Domain): 433 images with plate boxes in PASCAL VOC format. It has boxes only, so the reading step uses an OCR engine.
For B.Tech / for M.Tech: B.Tech: as-is, with a YOLO detector and an OCR library. M.Tech: test on Indian plates you photograph at night and at angles, report accuracy per character, and run it on a small board.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Face mask detection**

Problem: Hospitals and labs that require masks have no simple way to check at the door.

Output: A detector that marks each face in a live video as with mask, without mask or mask worn incorrectly.

– Dataset and level

Dataset: [Face Mask Detection](#) (CC0: Public Domain): 853 images in three classes, with boxes.

For B.Tech / for M.Tech: B.Tech: as-is, but examiners have seen it many times since 2020, so give it a clear setting such as a lab entrance. M.Tech: not advised unless you change the problem, for example recognising people while they wear masks.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Driver drowsiness detection**

Problem: Tired drivers nod off on long highway drives, often at night.

Output: A webcam system that tracks the eyes and sounds an alarm when they stay closed too long, with the alarm delay and false alarms measured.

 Dataset and level

Dataset: [Drowsiness Detection](#)

Dataset (CC0: Public Domain), built from MRL and Closed Eyes in the Wild images of open and closed eyes; also the [MRL Eye Dataset](#) (MIT).

For B.Tech / for M.Tech: B.Tech: as-is, with an eye classifier or the eye aspect ratio. M.Tech: test at night with an infrared camera and with spectacles, add yawning or head pose, and run it on a Raspberry Pi.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Traffic sign recognition**

Problem: Driver-assist systems must read road signs correctly in rain, glare and at speed.

Output: A classifier for the German sign classes, with accuracy for each class and a demo on dashcam video.

 Dataset and level

Dataset: [GTSRB, the German Traffic Sign Recognition Benchmark](#) (CC0: Public Domain on Kaggle), from the IJCNN 2011 competition, with more than 40 classes and more than 50,000 images.

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: test robustness to blur, rain and stickers on signs, or test on Indian signs, which differ from German ones.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Facial emotion recognition**

Problem: Online classes and help kiosks cannot tell whether a user is confused or frustrated.

Output: A model that labels a face with one of seven emotions on a live webcam, with a confusion matrix for the test set.

 Dataset and level

Dataset: [FER-2013](#) (Kaggle lists the Open Database licence): 48×48 grayscale faces, 28,709 for training and 3,589 in the public test set.

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: the classes are uneven, with few disgust images; try methods for noisy labels, test on a second expression dataset, and discuss bias and consent.

TOPIC IDEA · DEEP LEARNING ·
COMPUTER VISION**Waste classification from photos**

Problem: Mixed waste cannot be recycled, and sorting it by hand is slow and often skipped.

Output: A model that names the material in a photo (cardboard, glass, metal, paper, plastic or trash), running on a phone or a small board.

 Dataset and level

Dataset: [Garbage Classification](#) (six classes; Kaggle shows “Data files © Original Authors”) or [Garbage Classification \(12 classes\)](#).

For B.Tech / for M.Tech: B.Tech: as-is, with a pre-trained MobileNetV2. M.Tech: test on photos of Indian household waste in poor light, and move to detecting several items in one frame. Read the [waste classification guide](#).

Text and NLP projects

TOPIC IDEA · NLP · TEXT

Fake news detection

Problem: False stories spread on social media before anyone checks them.

Output: A classifier that labels an article as fake or real, shows the words behind each decision, and is tested on articles from outside the dataset.

⊖ Dataset and level

Dataset: [Fake and real news dataset](#) (CC BY-NC-SA 4.0): Fake.csv and True.csv, with title, text, subject and date.

For B.Tech / for M.Tech: B.Tech: TF-IDF with logistic regression, then a BERT model. M.Tech: a model can learn the writing style of the sources instead of the truth of a claim; test on a different fake news set, remove source clues and see what accuracy is left.

TOPIC IDEA · NLP · TEXT

Sentiment analysis of movie reviews or tweets

Problem: Companies receive far more reviews and posts than anyone can read.

Output: A model that labels text as positive or negative, compared across TF-IDF, an LSTM and a fine-tuned transformer, with a page that scores typed text.

⊖ Dataset and level

Dataset: [IMDB Dataset of 50K Movie Reviews](#), from Stanford's Large Movie Review Dataset, or [Sentiment140](#), 1.6 million tweets labelled automatically. For both, Kaggle says the licence is in the description.

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: move to Hindi or Hinglish text, or to sentiment about each aspect of a product, or test a model trained on reviews against tweets.

TOPIC IDEA · NLP · TEXT

SMS spam detection

Problem: Spam and scam texts reach every phone, and some lead to fraud.

Output: A classifier that flags spam with precision and recall reported, because a real message marked as spam costs more than a missed spam.

 Dataset and level

Dataset: [SMS Spam Collection](#)

Dataset: 5,574 English messages labelled ham or spam. Kaggle shows the licence as Unknown, so cite the original UCI collection.

For B.Tech / for M.Tech: B.Tech: a mini project, as-is. M.Tech: the messages are old and not from India; collect recent Indian scam texts about KYC, UPI and parcel delivery, test on them, and add Hinglish.

Finance, business and farm data projects

TOPIC IDEA · MACHINE LEARNING · DATA

Stock price prediction

Problem: Students want to forecast prices, but many models only learn to copy yesterday's price.

Output: A next-day forecast of price or direction for NIFTY 50 stocks, compared honestly with a 'same as yesterday' baseline.

⊖ Dataset and level

Dataset: [NIFTY-50 Stock Market Data \(2000 - 2021\)](#) (CC0: Public Domain): daily prices and volumes from NSE India, from 1 January 2000 to 30 April 2021.

For B.Tech / for M.Tech: B.Tech: as-is with an LSTM, as long as the baseline is in your results table. M.Tech: test walk-forward through time, add news sentiment, and backtest with trading costs. Expect small gains or none, and say so; it is a study, not investment advice.

TOPIC IDEA · MACHINE LEARNING · DATA

House price prediction

Problem: Buyers and sellers cannot tell whether an asking price is fair.

Output: A regression model with its error on the log price, the features that matter most, and a form that estimates a price.

⊖ Dataset and level

Dataset: The [House Prices: Advanced Regression Techniques](#) competition (Ames, Iowa data; the competition rules apply) or [California Housing Prices](#) (CC0: Public Domain).

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: build an Indian city dataset from listings you are allowed to use, add location features from maps, and give a price range rather than one number.

TOPIC IDEA · MACHINE LEARNING · DATA

Crop recommendation from soil and weather

Problem: Farmers often choose crops by habit, even when the soil and rainfall suit another crop better.

Output: A model that recommends a crop from nitrogen, phosphorus, potassium, temperature, humidity, pH and rainfall, with a simple form in English and a local language.

 Dataset and level

Dataset: [Crop Recommendation](#)

Dataset (Apache 2.0), built by combining Indian rainfall, climate and fertiliser data.

For B.Tech / for M.Tech: B.Tech: a mini project, as-is; check how much each model adds over a simple decision tree before you call one better. M.Tech: join real district soil and yield records, and recommend by expected profit rather than a single label.

TOPIC IDEA · MACHINE LEARNING · DATA

Credit card fraud detection

Problem: Fraud is a tiny share of card payments, so a model can look accurate and still miss almost all of it.

Output: A fraud score for each transaction and a threshold set from costs, judged by the area under the precision-recall curve, not accuracy.

 Dataset and level

Dataset: [Credit Card Fraud Detection](#)

from ULB (Kaggle lists the Open Database licence); most features are anonymised principal components.

For B.Tech / for M.Tech: B.Tech: as-is.

M.Tech: split by time instead of at random, and measure how results change across time windows with and without retraining. Read the [credit card fraud detection guide](#).

TOPIC IDEA · MACHINE LEARNING · DATA

Customer churn prediction

Problem: A telecom company loses customers every month and cannot afford to call all of them with an offer.

Output: A churn score for each customer, the reasons behind it and a ranked list of who to contact first.

– Dataset and level

Dataset: [Telco Customer Churn](#), IBM's sample data (Kaggle shows "Data files © Original Authors").

For B.Tech / for M.Tech: B.Tech: as-is. M.Tech: move from who will leave to who would stay because of an offer (uplift modelling), or set the threshold from retention costs. Read the [customer churn prediction guide](#).

More to choose from: our [list of 49 machine learning projects](#), and the lists of [B.Tech project ideas](#) and [M.Tech project ideas](#).

KEY POINTS

- ◆ **25 projects with Kaggle datasets**, each with a problem, an output, a checked dataset link and the licence Kaggle shows.
- ◆ **Five groups:** medical imaging (6), health records (3), vision and agriculture (8), text and NLP (3), and finance and business (5).
- ◆ **All 25 suit B.Tech as they stand.** For M.Tech, each card names one change to make and measure.
- ◆ **Split by patient, person or time**, not at random, wherever one source gives many rows or images.
- ◆ **Do not quote accuracy from a notebook** unless you can reproduce it on your own split.

Picked a Kaggle project?

Send the project, your level (B.Tech or M.Tech) and your review dates. The consultation is free.

Message before 7 pm for a reply today.

Free consultation →

WhatsApp

Are these topics too common for M.Tech?

Yes, as they stand. Every project on this page has been built many times, and thousands of public notebooks exist for the same datasets. An M.Tech examiner who sees a well-known dataset, a standard CNN and a high accuracy will ask what is new. If the answer is nothing, the project looks copied, even when it is your own work.

The topic itself is not the problem. What makes it acceptable is **one change that you measure**:

- ◆ **A harder or outside test set.** Train on Kaggle data and test on data from another hospital, city, camera or year. The drop in score is a finding.
- ◆ **A fair split.** Show how much the score falls when you split by patient, lesion, person or time instead of at random.
- ◆ **A newer model on the same split.** Compare a Vision Transformer or a fine-tuned transformer with the CNN or LSTM baseline, with the same data and settings.
- ◆ **Explainability you check.** Heatmaps or SHAP values, then a check that they point at the right thing, such as the lesion and not a ruler or a label.
- ◆ **Deployment under limits.** Run the model on a phone or a small board, and report the size, the speed and the accuracy lost.

Start from a published paper that used the same dataset, so you have numbers to compare against. Our guide on [how to select a base paper](#) explains how to choose one.

How do you pick a dataset on Kaggle?

Check four things on the dataset page before you write the synopsis.

- 1. Licence.** It is shown on the dataset page. CC0 and CC BY allow most uses; CC BY-NC means no commercial use; “Unknown” or “Data files © Original Authors” means you should find and cite the original source. Competition data follows the competition’s rules.
- 2. Size.** Large enough to split into training, validation and test sets, small enough to train on a free cloud GPU in hours. Download it the day you choose it.
- 3. Class balance.** Count the examples in each class. If one class is rare, as with fraud, melanoma or disgust faces, plan to report recall and precision for it, not overall accuracy.
- 4. A published baseline.** Search for a paper that used the same dataset and reports its split and metric. Without one, you have nothing to compare your results with.

Also read the description for where the data came from. Many Kaggle datasets are copies or mixes of older public sets, and that changes both what you cite and which data counts as an outside test.

How can we help with a Kaggle dataset project?

For a B.Tech team, we help build the project from the dataset to a working demo: the code, a comparison of two or three models, the black book, the PPT and viva preparation. For an M.Tech student, we help choose the base paper, reproduce its numbers, plan and run the extension, and write the thesis and a paper in the format your college asks for.

See how we work on [final year projects for CSE and IT](#) and [M.Tech projects](#), or read our delivered case studies on [Parkinson’s detection](#) and [diabetic retinopathy](#).

Frequently asked questions

Are Kaggle datasets good enough for a final year project?



Yes, for most B.Tech projects. A Kaggle dataset is public, already labelled and usually has notebooks you can learn from. Examiners care more about a fair test split, the right metric and a working demo than about where the data came from, so explain those clearly in the report and viva.

Can I use a Kaggle dataset project for M.Tech?



Yes, if you change something and measure the result. Reproducing a common notebook is not enough at M.Tech level. Test on a second dataset, use a newer model on the same split, add explainability, or run the model on a phone, and compare against a published baseline.

Can I use a Kaggle dataset in a published paper?



Check the licence on the dataset page first. Many are CC0 or CC BY, some are non-commercial, and some show Unknown. Cite the original source of the data as well as the Kaggle page, because many Kaggle datasets are copies of data first released elsewhere.

Which Kaggle project is easiest for a beginner?



A table of numbers is easier than images or text. Heart disease, diabetes and SMS spam train in minutes on a laptop, so you can spend your time on cleaning the data, comparing two or three models and explaining the results.

Why does my model score so high on a Kaggle dataset?



Often because the test set is not truly separate. The same patient, lesion or person may appear in both training and test data, or the set may contain repeated images. Split by patient or person, remove duplicates, and test on outside data before you trust a very high score.



Can The Ultimate Project World help with a Kaggle dataset project?

Yes. Share the project, your level and your review dates in the free consultation. We will tell you exactly which parts we can take on, such as the code, the comparison of models, a working demo, the report, the PPT and viva preparation.

◆ RELATED SERVICES

From a topic to a project you can present



Final-year projects

Code, a working demo, the black book, a PPT and viva prep.



M.Tech projects

A base paper, one measured extension, a paper and the thesis.



Report & PPT

The report, black book or dissertation, and review slides.

