SOUTHERN LUZON STATE UNIVERSITY
College of Engineering
COMPUTER ENGINEERING DEPARTMENT

CPE15 Cognate and Professional Course 1

Name: ENCALLADO, Carl Francis T.	**Date:** December 12, 2025
Section: GF	**Machine Learning Cognate 1 Project**

DATASET SUMMARY

The following tables list all datasets used:

[01] Dataset Name
MachineLearningModel_Encallado.ipynb
Description
This Jupyter Notebook serves as the core analytical workflow for the "GPS Connectivity Analysis for IT-BPM Hub Prediction in the Philippines" project. It includes detailed code implementations for data loading from various sources (e.g., CSV, GeoJSON, and ZIP files containing shapefiles), exploratory data analysis (EDA) with visualizations such as histograms, scatter plots, and bar charts for connectivity metrics, feature engineering steps (e.g., calculating indices like Talent Pool, Digital Readiness, Accessibility, and Saturation), machine learning modeling using K-Means clustering for city classification, and generation of outputs like interactive maps and cluster analyses. The notebook is structured with markdown explanations, code cells, and inline comments to ensure reproducibility, covering imports, preprocessing, modeling, and interpretation of results.
Source Link
[*Internal – this project*]
File Information & Size
Jupiter Source File (.ipynb) – 4.00 MB
About
This notebook acts as a comprehensive reference for all analytical steps in the project, justifying machine learning model choices and outputs such as cluster interpretations for IT-BPM hub recommendations. It integrates geospatial, connectivity, and employment data to predict emerging hubs, providing insights into digital divides and talent availability, with visualizations supporting findings on urban-rural disparities.

[02] Dataset Name
gadm41_PHL_2.json
Description
The GADM (Global Administrative Areas) dataset provides high-resolution spatial data for administrative boundaries worldwide, with version 4.1 delimiting over 400,276 administrative areas across countries. For the Philippines, the Level 2 file (gadm41_PHL_2.json) includes detailed polygon geometries for provinces, municipalities, and cities. Data is sourced from official government maps, crowdsourced contributions, and satellite imagery, updated periodically. It supports

formats like GeoJSON for easy integration with GIS tools, and is licensed for non-commercial use with attribution required.

Source Link

https://gadm.org/download_country.html

File Information & Size

JSON Source File (.json) – 2.34 MB

About

This dataset is essential for geospatial analysis in the project, providing polygon geometries for mapping connectivity tiles from Ookla data onto Philippine municipalities. It enables choropleth visualizations, spatial joins with connectivity metrics, and distance calculations, helping identify regional disparities in digital infrastructure and supporting recommendations for IT-BPM investments.

[03] Dataset Name

2020-04-01_performance_mobile_tiles.zip
2020-04-01_performance_mobile_tiles —+
2020-04-01_performance_mobile_tiles.dbf —+ |
2020-04-01_performance_mobile_tiles.prj —+ |
2020-04-01_performance_mobile_tiles.shp —+ |
2020-04-01_performance_mobile_tiles.shx —+ |

Description

This dataset from Ookla Open Data captures global network performance metrics. It includes key variables such as average download speed (avg_d_kbps), average upload speed (avg_u_kbps), average latency (avg_lat_ms), number of tests, and unique devices, derived from millions of Speedtest with GPS accuracy. Coverage spans quarterly from Q1 2019 to Q3 2025, with data filtered for cellular connections (e.g., 4G LTE, 5G NR). The ZIP file contains Esri Shapefile components (.shp for geometries, .dbf for attributes, .prj for projection, .shx for indexing), using WGS 84 (EPSG:4326) for geometries and EPSG:3857 for projection.

Source Link

https://github.com/teamookla/ookla-open-data

File Information & Size

Compressed (zipped) Folder (.zip) – 236 MB

About

The Speedtest data in this dataset supports efforts to enhance network performance, ensure regulatory accountability, and promote equitable Internet access by assisting operators, governments, and institutions in identifying and addressing connectivity gaps. In this project, it provides the foundational connectivity metrics for the Philippines, aggregated by city or municipality to compute Digital Readiness Indices, revealing urban-rural divides and informing ML clustering for IT-BPM hub predictions. Licensed under CC BY-NC-SA 4.0, it emphasizes privacy compliance through periodic reaggregation.

[04] Dataset Name

PSGC-3Q-2025-Publication-Datafile.csv

Description

The Philippine Standard Geographic Code (PSGC) is a systematic classification system developed by the Philippine Statistics Authority (PSA) for all geographic areas in the Philippines, assigning unique codes to regions, provinces, municipalities/cities, and barangays. This Q3 2025 CSV file includes comprehensive details such as 10-digit PSGC codes, names, correspondence codes, geographic levels (e.g., region, province, municipality), old names, city classes, income classifications (per DOF DO No. 074.2024), urban/rural designations.

Source Link

https://psa.gov.ph/system/files/scd/PSGC-3Q-2025-Publication-Datafile.xlsx

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 2.19 MB

About

This dataset supplements population and boundary data, enabling accurate matching of municipalities in geospatial analysis and integration with connectivity and employment metrics. It justifies population-based indices like the Talent Pool Index by providing 2024 projections and supports normalization of features for ML modeling, ultimately aiding in the identification of high-potential IT-BPM hubs.

[05] Dataset Name

psgc_data_cleaned.csv

Description

This cleaned version of the PSGC dataset focuses on municipality-level data, including normalized names, geographic levels, income classifications per DOF DO No. 074.2024, and 2024 population projections. Derived from the full PSGC, it removes redundancies, standardizes formatting for merging, and includes columns like Name_Normalized for fuzzy matching with other datasets.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 58.1 KB

About

Essential for data integration in the project, this cleaned file provides normalized names and population data to merge with connectivity metrics from Ookla and employment indicators from LFS. It justifies population-weighted analyses and feature engineering, such as scaling indices by population size, contributing to accurate predictions of talent availability in potential IT-BPM locations.

[06] Dataset Name

PHL-PSA-LFS-2024-10-PUF.zip

PHL-PSA-LFS-2024-10-PUF —+
LFS_PUF_October_2024.F2 —+ |
LFS October 2024 Questionnaire.html —+ |
LFS PUF October 2024.csv —+ |
lfs_october_2024_metadata(dictionary).xlsx —+ |
LFS_PUF_October_2024.dcf —+ |

Description

The Labor Force Survey (LFS) is a quarterly household-based survey conducted by the Philippine Statistics Authority (PSA) to gather data on the demographic and socioeconomic characteristics of the population, focusing on labor market indicators. The October 2024 Public Use File (PUF) ZIP includes the main CSV dataset with anonymized records, metadata in Excel (dictionary with value sets), questionnaire HTML, and data codebook files. It covers a large sample (approximately 45,000 households nationwide), using a multi-stage sampling design to estimate employment, unemployment (3.9% in October 2024 preliminary results), underemployment, and related metrics at national and regional levels.

Source Link

https://psada.psa.gov.ph/catalog/LFS/about

File Information & Size

Compressed (zipped) Folder (.zip) – 6.27 MB

About

As the core dataset for extracting employment status and "No Work" counts per location, this file is crucial for predicting IT-BPM talent availability by integrating labor indicators with connectivity data. It supports feature engineering (e.g., unemployment proxies) and ML inputs, providing evidence-based insights into workforce potential for hub recommendations, aligned with national goals for economic planning and job creation.

[07] Dataset Name

lfs_october_2024_metadata(dictionary).xlsx lfs_october_2024_valueset_C12A.csv

Description

This CSV extract from the LFS metadata Excel file specifically contains the value set for variable C12A (Location of Work - Province, Municipality), mapping codes to names (e.g., 0101 for Abra - Bangued). It serves as a lookup table for processing location-based employment data, derived from the full metadata dictionary which includes all variable descriptions, codes, and value labels.

Source Link

[*Extracted from lfs_october_2024_metadata(dictionary).xlsx - lfs_october_2024_valueset.csv*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 59.9 KB

About

This helper dataset facilitates the extraction of “No Work Count” from LFS by enabling location-specific aggregation, relevant for assessing talent availability in IT-BPM predictions. It enhances data integration accuracy, supporting analyses of employment trends by municipality and contributing to cluster-based hub identifications.

[08] Dataset Name

Feature-Engineered Dataset_Encallado.csv

Description

This raw feature-engineered CSV compiles quadkey-based data from Ookla tiles, including connectivity metrics (avg_d_kbps, avg_u_kbps, avg_lat_ms, tests, devices), joined with province/municipality names (NAME_1, NAME_2), 2024

population, distances from Metro Manila, and No_Work_Count from LFS. It represents the initial merged dataset before aggregation.
Source Link [*Internal – this project*]
File Information & Size Microsoft Excel Comma Separated Values File (.csv) – 9.77 MB
About As the input for further modeling, this dataset supports feature engineering tasks like distance calculations and unemployment proxies, providing a granular view of connectivity-employment intersections. It justifies spatial patterns in digital readiness and is foundational for deriving aggregated indices used in ML clustering.

[09] Dataset Name Feature-Engineered Dataset_cleaned.csv
Description This aggregated CSV summarizes connectivity metrics, population, distances, and No_Work_Count at the municipality level, with engineered features ready for modeling. It includes records like top/bottom performers and serves as the cleaned input for indices and clustering.
Source Link [*Internal – this project*]
File Information & Size Microsoft Excel Comma Separated Values File (.csv) – 105 KB
About This file encapsulates engineered features for ML, justifying clustering by integrating geospatial, connectivity, and employment insights. It highlights disparities (e.g., high No_Work_Count in rural areas) and supports recommendations for IT-BPM hubs with balanced metrics.

[10] Dataset Name PreprocessedDataset_Encallado.geojson
Description This preprocessed GeoJSON file contains quadkey polygons with associated connectivity metrics (speeds, latency, tests, devices), transformed for visualization and analysis, including WKT geometries and attributes for overlay on maps.
Source Link [*Internal – this project*]
File Information & Size GEOJSON File (.geojson) – 66.4 MB
About As the geospatial layer for mapping, this dataset is essential for overlaying connectivity on administrative boundaries, enabling visualizations of trends and supporting spatial queries in the project for identifying potential IT-BPM locations.

[11] Dataset Name InteractiveFoliumMap_Encallado.html

SOUTHERN LUZON STATE UNIVERSITY
College of Engineering
COMPUTER ENGINEERING DEPARTMENT

Description
This interactive HTML file, generated using Folium, displays a map of the Philippines with layered connectivity metrics (e.g., heatmaps for speeds/latency), administrative boundaries from GADM, and markers for key cities, allowing zooming, panning, and layer toggling for exploratory analysis.

Source Link
[*Internal – this project*]

File Information & Size
HTML Document (.html) – 13.0 MB

About
This output facilitates interactive exploration of connectivity trends, justifying visual insights into digital divides (e.g., poor coverage in remote areas) and supporting the project's findings on hub potential through geospatial context.

Note: To access all source files online, please click this link provided. All datasets, scripts, and supporting documents are available through the shared repository for easy viewing and download.