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**

Topic: Identifying Next-Wave IT-BPM Hubs in the Philippines Using GPS Connectivity and Employment Data

DATASET SUMMARY

The following tables list all datasets used:

[01] Dataset Name
MachineLearningModel_Encallado.ipynb
Description
The "GPS Connectivity Analysis for IT-BPM Hub Prediction in the Philippines" project's main analytical workflow is presented in this Jupyter Notebook. It contains the complete code for the data loading process from different sources (e.g., CSV, GeoJSON, and ZIP files with shapefiles), then continues with exploratory data analysis (EDA) using techniques like histograms, scatter plots, and bar charts for connectivity metrics, leading to the feature engineering steps (e.g., calculating indices for Talent Pool, Digital Readiness, Accessibility, and Saturation), followed by machine learning modeling using K-Means clustering for city classification, and eventually yielding the outputs such as interactive maps and cluster analysis. The notebook is comprised of markdown explanations, code cells, and inline comments to maintain reproducibility throughout the process of importing, preprocessing, modeling, and result interpretation.
Source Link
[*Internal – this project*]
File Information & Size
Jupiter Source File (.ipynb) – 4.00 MB
About
This notebook serves as an extensive guide for the entire project's analytical process, justifying the choices and outputs of machine learning models such as the interpretation of clusters for the recommendations of IT-BPM hubs. It combines geospatial, connectivity, and employment data to foresee upcoming hubs, revealing the situation regarding digital divides and the presence of talent, while visualizations are backing the findings on the differences between urban and rural areas.

[02] Dataset Name
gadm41_PHL_2.json
Description
The GADM (Global Administrative Areas) dataset gives access to detailed spatial data of administrative borders all over the planet, and the latest version 4.1 marks more than 400,276 such areas in different countries. In the case of the Philippines,

the Level 2 file (gadm41_PHL_2.json) contains the complete polygon geometries for provinces, municipalities, and cities. The data comes from official government maps, crowdsourced contributions, and satellite images, which are updated regularly. It supports GeoJSON, thus is easily integrated with GIS tools. The dataset is available for non-commercial use, with the condition of attribution.
Source Link https://gadm.org/download_country.html
File Information & Size JSON Source File (.json) – 2.34 MB
About The geospatial analysis of the project would not be complete without this dataset as it is an important factor in the modelling process for mapping the connectivity tiles from Ookla data to the different municipalities in the Philippines. The dataset support choropleth visualizations, spatial joins with connectivity metrics, and distance calculations which together will help pinpoint areas of disparity in digital infrastructure and thus recommend areas for IT-BPM investments that are most lucrative.

[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 The Ookla Open Data dataset is the source of the global network performance metrics, which are presented here. The dataset contains several important parameters including average download speed (avg_d_kbps), average upload speed (avg_u_kbps), average latency (avg_lat_ms), number of tests, and unique devices. These parameters were obtained through the millions of Speedtests with GPS location accuracy. Esri Shapefile components (.shp for geometries, .dbf for attributes, .prj for projection, .shx for indexing) are included in the ZIP file. The WGS 84 (EPSG:4326) standard is applied 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 included in this dataset plays a significant role in enhancing network performance, ensuring regulatory accountability, and promoting equitable Internet access by helping the operators, governments, and institutions to identify and remediate connectivity gaps. The project delivers the basic connectivity metrics for the Philippines, city or municipality-wise aggregated, in order to calculate Digital Readiness Indices, uncovering urban-rural splits and guiding ML clustering for IT-BPM hub forecasts.

[04] Dataset Name

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

Description

The Philippine Standard Geographic Code (PSGC) is a systematic geographical classification system created by the Philippine Statistics Authority (PSA) that encompasses all areas in the Philippines, and it uses unique codes for regions, provinces, municipalities/cities, and barangays. This CSV file for the third quarter of 2025 contains thorough information like 10-digit PSGC codes, names, correspondence codes, geographic levels (e.g., region, province, municipality), old names, city classes, income classifications (according to DOF DO No. 074.2024), and 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

The dataset in question adds to the already existing population and boundary data, thus allowing municipalities to be accurately matched in geospatial analysis and be integrated with connectivity and employment metrics. It serves to substantiate population-based indices such as the Talent Pool Index by giving 2024 forecasts and assists in the normalization of features for ML modeling, which in turn helps in the recognition of high-potential IT-BPM hubs.

[05] Dataset Name

psgc_data_cleaned.csv

Description

The PSGC dataset version that has been cleaned up concentrates on municipality-level data, comprising normalized names, geographic levels, income classifications according to DOF DO No. 074.2024, and population forecasts for 2024. It is a derivative of the complete PSGC, which has eliminated duplicate entries, has standardized its formatting for the purpose of merging, and has added columns such as Name_Normalized to allow 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

This cleaned file, which is key for the data integration in the project, not only delivers regularized names and population numbers but also helps to merge them with the corresponding metrics of connectivity from Ookla and employment indicators from LFS. It supports population-weighted analyses and feature engineering, for instance, scaling indices by population size, which in turn helps to get accurate forecasts of talent availability in the possible 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 Philippine Statistics Authority (PSA) conducts the Labor Force Survey (LFS) that is, a quarterly household-based survey, to gather data on demographic and socioeconomic characteristics of the population with a focus on labor market indicators. The October 2024 Public Use File (PUF) ZIP contains the main CSV dataset with anonymized records, metadata in Excel (dictionary with value sets), questionnaire HTML, and data codebook files. It is based on a very large sample (around 45,000 households all over the country), employing a multi-stage sampling design to produce estimates of employment, unemployment (3.9% in October 2024 preliminary results), underemployment, and related metrics for 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
This file serves as the key resource for determining employment status and "No Work" figures for each location, and it is thus indispensable for combining labor indicators and connectivity data to forecast the availability of IT-BPM talent. It is also a support to feature engineering (e.g., proxies of unemployment) and ML inputs and is able to deliver evidence-backed insights about the workforce potential for hub recommendations, thereby complying with national goals related to economic planning and job creation.

[07] Dataset Name
lfs_october_2024_metadata(dictionary).xlsx lfs_october_2024_valueset_C12A.csv

Description
The provided CSV file is extracted from the LFS metadata Excel file and it only consists of the value set for variable C12A (Location of Work - Province, Municipality), where codes and names are mapped to each other (for example, 0101 for Abra - Bangued). It is a reference table used in processing of location-based employment data, which comes from the full metadata dictionary containing 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
The helper dataset, in this case, allows the extraction of "No Work Count" from LFS through location-specific aggregation. This is significant for the evaluation of the availability of talents in the IT-BPM forecasts. Moreover, it increases the accuracy of

data integration, which is necessary for the analysis of employment trends by municipality and the identification of clusters for hubs.

[08] Dataset Name

Feature-Engineered Dataset_Encallado.csv

Description

The initial merged dataset before aggregation is represented by this raw feature-engineered CSV which compiles quadkey-based data from Ookla tiles, such as connectivity metrics (avg_d_kbps, avg_u_kbps, avg_lat_ms, tests, devices) together with province/municipality names (NAME_1, NAME_2), 2024 population, distances from Metro Manila, and No_Work_Count from LFS.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 9.77 MB

About

This dataset, which has been used as the input for further modeling, allows for feature engineering activities such as distance calculations and the creation of unemployment proxies, thus providing a detailed view of the intersections between connectivity and employment. It explains the spatial variations in digital readiness and is a starting point for the development of aggregated indices that are applied in ML clustering.

[09] Dataset Name

Feature-Engineered Dataset_cleaned.csv

Description

This consolidated CSV file provides a summary of connectivity metrics, population, distances, and No_Work_Count at the municipality level, with the generation of features that are ready for modeling. It contains entries such as top/bottom performers and acts as the purified input for indices and clustering.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 105 KB

About

This document holds the engineered attributes for machine learning, and through the incorporation of geospatial, connectivity, and employment insights, it justifies the clustering. It reveals inequalities (for instance, a large number of people out of work in the countryside) and backs the suggestion of IT-BPM centers with appropriate metrics.

[10] Dataset Name

PreprocessedDataset_Encallado.geojson

Description

The GeoJSON file which has gone through preprocessing contains quadkey polygons along with the related connectivity metrics (speeds, latency, tests, devices)

that were transformed for visualization and analysis. The file also includes WKT geometries and attributes which can be used 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 fundamental for imposing connectivity on administrative divisions, which allows visualization of trends, and makes the project for pinpointing potential IT-BPM locations easier with spatial queries.

[11] Dataset Name InteractiveFoliumMap_Encallado.html
Description An HTML file that is interactive and created by Folium presents the Philippines' map with overlaid connectivity metrics (such as heatmaps for speeds/latency), administrative boundaries from GADM, and points marked for major cities, all of which allow for zooming, panning, and layer switching to conduct exploratory analysis.
Source Link [*Internal – this project*]
File Information & Size HTML Document (.html) – 13.0 MB
About The result of this processing makes it possible to interactively explore the connectivity trends, thus allowing for visual findings (like remote areas with poor network coverage) and backing the project's conclusions about hub potential owing to the 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.