COMP 381: Machine Learning

Term Project - Proposal
Population and Household Projection

Name	*Student Id.*
Chaahna Chandiramani	300198011

Project Proposal

Topic:

Population and Household

Questions to be answered:

Which regional districts have similar population age distributions? (This question aims to identify groups of regional districts that have similar proportions of population in different age groups, such as 0-14, 15-19, 20-24, etc.)

Can we group regional districts based on the gender distribution of their population? (This question seeks to identify clusters of regional districts that have similar proportions of males and females across different age groups.)

Are there distinct clusters of regional districts based on their household projections? (This question aims to identify groups of regional districts that have similar patterns of household sizes and changes over time.)

Can we predict the gender distribution of a regional district's population based on the age distribution? (This question aims to classify regional districts into categories based on the expected proportion of males and females in different age groups.)

Can we classify regional districts into different household types based on the average number of persons per household and the age distribution? (This question seeks to categorize regional districts into groups that have similar household sizes and age structures, such as small families, large families, or elderly households.)

Is it possible to predict the regional district of a household based on its demographic characteristics, such as age distribution and average number of persons per household? (This question aims to classify households into different regional districts based on their demographic attributes.)

Can we predict the future population growth rate of a regional district based on its current median age and average age? (This question explores the relationship between age demographics and population growth.)

Is there a correlation between the median age of a regional district and its average household size? (This question aims to understand the relationship between age demographics and household composition.)

Can we estimate the future population size of a regional district based on its current household projections? (This question investigates the relationship between household projections and overall population trends.)

Approach:

- Find data sets based on the topic with are open and direct access.
- Work on each data set (table/excel sheet) handling missing values, encoding categorical variables where required and splitting it into training and testing sets.
- Find the best fit model for each question to be answered.
- Training the model.
- Validating the model performance.
- Finding patterns, trends, and connections to create analysis and make predictions.

Expected Results:

The projections will give insights into the expected changes in population composition over time. The predictions will provide information on the proportion of males and females in different age groups. The estimates will offer insights into the demographic characteristics and population trends in each district. The projections will give an indication of the expected changes in the housing market and population distribution. The predictions will offer insights into the household sizes and potential implications for housing demands and infrastructure planning.

The outcome should answer most of the questions listed, if not all. The outcome should provide valuable results based on the data. The accuracy and reliability of the results will depend on the quality of the data and the assumptions made during the analysis. However, the confidence level of the predictions should be at least 85%.

Data set selected:

Population and household projection by Regional District for 2019-2028; from BC Data Catalogue

Access to the data set on the BC Data Catalogue:

https://catalogue.data.gov.bc.ca/dataset/population-and-household-projections-2019-2028-/resource/d39bda1b-b9a9-4d2c-8c83-2eb1d8b6dbcc

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page_Download-Telecharger.cfm?Lang=E&Tab=1&Geo1=PR&Code1=59&Geo2=PR&Code2=01&SearchText=&SearchType=Begins&SearchPR=01&B1=Housing&TABID=1&type=1

Access to the data set on google sheet:

https://docs.google.com/spreadsheets/d/1fS4KwSlAIrqx1ir9hlW-OGjwFpAbuyGQ/edit?usp=sharing&ouid=110023239984027275402&rtpof=true&sd=true

Machine Learning model involvement in the project:

a. Clustering

   K-means clustering or hierarchical clustering can be used.
   Data input: Population attributes such as age distribution , gender distribution, and total population along with average number of persons per household.
   The clustering models will analyse patterns and similarities in the population and household data to create clusters or groups. The resulting clusters can provide insights into the distribution and characteristics of regional districts based on the given attributes. These insights can be used for further analysis and decision-making in population and household projection studies.

b. Classification

   One or many classification models, such as Logistic Regression, Decision Tree, Random Forest, or Naive Bayes will be selected to answer the different questions.
   Data input: Age distribution in different age groups. Demographic characteristics, such as age distribution and average number of persons per household.
   Target Variable: Gender distribution (proportion of males and females) in the regional district and regional district of the household.
   The classification models will be trained using historical data and evaluated using appropriate evaluation metrics, such as accuracy and precision. The outcomes will provide insights into the relationships between demographic characteristics and population/household attributes, enabling predictions and classifications based on the trained models.

c. Regression

   We will be using linear regression in our dataset to predict the population in the coming years. Starting with selection the features which we will be using to estimate the population would be Year, regional district, Annual Population changes and Age.

   By using the above selected features, we will then choose a categorical data column that will be further used in training the model and feed the Age, Year, Regional district to predict the target variable which will be Annual Population Changes. The outcome of this can help us gain more insights on the relationship of these variables. After we are done and satisfied with the result, we will then create another table with the same variable and use the outcome of the training model.

References:

Khavari, B., Korkovelos, A., Sahlberg, A. *et al.* Population cluster data to assess the urban-rural split and electrification in Sub-Saharan Africa. *Sci Data* **8**, 117 (2021). https://doi.org/10.1038/s41597-021-00897-9

Fleetwood, D. (2018, August 9). *Population Data: Definition, Classification, Estimation and Importance*. QuestionPro. https://www.questionpro.com/blog/population-data/#:~:text=There%20are%20two%20primary%20classifications

How to use linear regression to model population growth? (n.d.). Cross Validated. https://stats.stackexchange.com/questions/207606/how-to-use-linear-regression-to-model-population-growth

Şahinarslan, F., Tekin, A., & Cebi, F. (2021). *Application of machine learning algorithms for population forecasting.* International Journal of Data Science, 6, 257-270. https://doi.org/10.1504/IJDS.2021.10047231

Appendix:

TCPS2 Core training Certificates:

1. Member 1


PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE 2022

Certificate of Completion

This document certifies that

Chaahna Chandiramani

successfully completed the Course on Research Ethics based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022)

Certificate # 0000932584 — 26 May, 2023

2. Member 2


PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE 2022

Certificate of Completion

This document certifies that

Parneet Kaur

successfully completed the Course on Research Ethics based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022)

Certificate # 0000932034 — 23 May, 2023

3. Member 3


PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE 2022

Certificate of Completion

This document certifies that

Mehak Makkar

successfully completed the Course on Research Ethics based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022)

Certificate # 0000932044 — 29 May, 2023