**Income Analysis of Census Database**

1 C.Tejesh, 2 K.Sriram, 3 S.Kedharnath

*Department of Computer Science and Engineering,*

*Amrita School of Computing,* *Amrita Vishwa Vidyapeetham,*  Bengaluru,Karnataka.

1 bl.en.u4aie22111@bl.students.amrita.edu , 2 bl.en.u4aie22124@bl.students.amrita.edu , 3 bl.en.u4aie22146@bl.students.amrita.edu

***Abstract*—** The income analysis of a census database plays a crucial role in understanding the factors influencing individuals' earnings. In this study, we utilize PySpark, a Python library for big data processing, along with machine learning (ML) and deep learning (DL) techniques to analyze income patterns. The dataset, obtained from the census database, consists of various features such as age, education level, marital status, and occupation, along with the corresponding income information.The analysis begins by importing the necessary libraries, including PySpark ML and DL, and creating a Spark session for data processing. The data is then loaded into a PySpark DataFrame and preprocessed to handle missing values and convert categorical variables into numerical representations. The dataset is split into training and testing sets to facilitate model evaluation.For the ML approach, a KNN and SVM algorithms are employed as the predictive model.The k-Nearest Neighbours (k-NN) and Support Vector Machine (SVM) algorithms is a supervised machine learning algorithm used for classification and regression tasks.In addition to the ML approach, deep learning techniques are also applied to the income analysis. These extracted features are then used to train a deep learning model, such as a neural network, to predict income levels.

I. INTRODUCTION

Income analysis based on census data is a critical task in understanding the socioeconomic landscape of a population. By studying the factors that contribute to individuals' earnings, we can gain insights into the dynamics of income distribution and identify patterns that can inform policies and decision-making processes. In this project, we employ PySpark, a powerful Python library for big data processing, along with machine learning (ML) and deep learning (DL) techniques, to analyze income patterns in a census database. The census database provides a wealth of information, including features such as age, education level, marital status, occupation, and more, along with corresponding income data. By leveraging the capabilities of PySpark's ML and DL libraries, we can extract valuable insights and build predictive models that can estimate income levels based on individual attributes.

• Machine learning techniques for income analysis-

Machine learning algorithms enable us to understand the relationship between different variables and predict income outcomes. We utilize PySpark's ML library to develop models such as Random Forest Classifier, which learns patterns from the provided census data and predicts income categories for new instances. Additionally, we preprocess the data to handle missing values, convert categorical variables into numerical representations, and split the dataset into training and testing setsfor model evaluation.

ML approaches for analysis:

• k-Nearest Neighbours (k-NN)

• Support Vector Machine (SVM) algorithm

* KNN algorithm for income analysis:

Income analysis is an important aspect of understanding economic patterns and identifying factors that contribute to individuals' earnings. One approach to performing income analysis is by utilizing the k-Nearest Neighbours (k-NN) algorithm. This algorithm is a popular choice for classification tasks, including predicting income levels based on various demographic and socioeconomic features. The k-NN algorithm is a non-parametric and instance-based machine learning algorithm. It operates on the principle that similar instances or data points tend to belong to the same class. In the context of income analysis, the k-NN algorithm examines the characteristics of individuals, such as age, education level, occupation, and other relevant attributes, to predict their income level. In this analysis, we will demonstrate how to apply the k-NN algorithm using PySpark for income analysis. We will load the census data, preprocess it by handling missing values and transforming categorical variables, and split it into training and testing sets. Then, we will build and train the k-NN model using the training data. Finally, we will evaluate the model's performance and analyze its predictions to gain insights into income patterns within the census database.

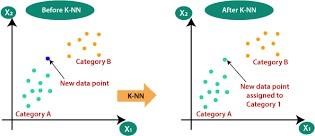


Fig1: KNN algorithm

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the above diagram

* SVM algorithm for income analysis:

Income analysis plays a significant role in understanding the economic landscape and identifying the key factors that contribute to an individual's earning potential. In this study, we employ the Support Vector Machine (SVM) algorithm, implemented using PySpark, to analyze income patterns in a census database. By leveraging the capabilities of PySpark's distributed computing framework, we can handle large-scale datasets efficiently and derive valuable insights.

The census database provides a comprehensive set of features, including demographic information, education level, occupation, and marital status, which can influence income levels. Through the application of SVM, we aim to build a predictive model that can accurately classify individuals into different income categories and understand the underlying patterns and relationships within the data.

SVM is a powerful algorithm that excels in handling both linearly separable and non-linearly separable data. By finding an optimal hyperplane that maximizes the margin between different income classes, SVM can effectively capture the decision boundaries and generalize well to unseen data. Additionally, the SVM algorithm can utilize kernel functions to transform the data into higher- dimensional spaces, allowing for non-linear classification and capturing complex income patterns.

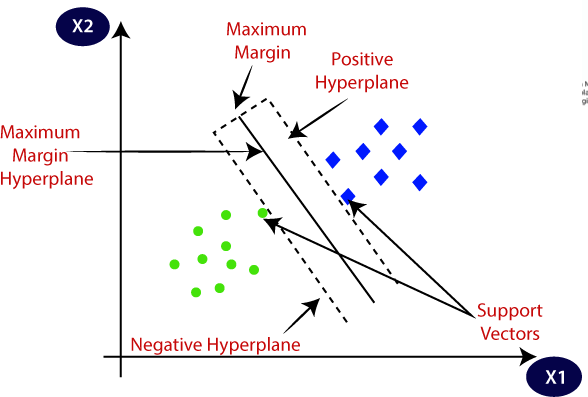


Fig2: SVM algorithm

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the above diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

* Deep learning techniques for income analysis:

Income analysis based on deep learning techniques has gained significant attention due to its ability to uncover complex patterns and relationships within large-scale datasets. In this study, we explore the application of deep learning techniques using PySpark, a Python library for big data processing, to analyze income patterns in a census database. By harnessing the power of PySpark's distributed computing framework, we can efficiently process and analyze the data, enabling deep learning models to learn intricate representations and make accurate predictions. The census database provides a rich collection of features, including demographic information, education level, occupation, and marital status, which can have a significant impact on an individual's income. Deep learning techniques, such as neural networks, offer the capability to capture non-linear relationships and extract high-level representations from these features. By leveraging PySpark's support for deep learning through its PySpark DL library, we can harness the distributed computing capabilities to train deep learning models effectively.By utilizing PySpark's distributed computing capabilities, deep learning models can efficiently handle large-scale datasets, effectively leveraging the computational resources of distributed clusters. This allows for parallelized training and inference, significantly reducing the computational burden and improving the scalability of the income analysis.

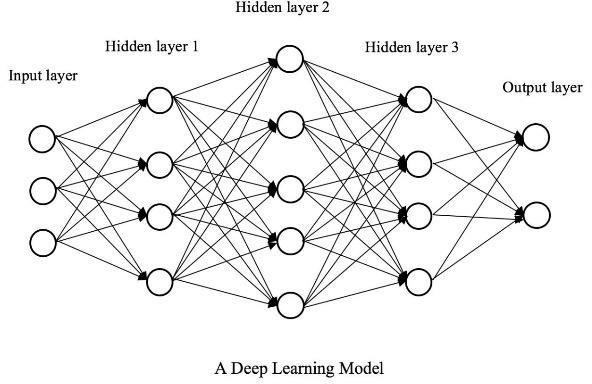


Fig 3: analysis of Deep learning

Neural Networks comprise of three layers,

1. Input layer,

2. Hidden layer,

3. Output layer.

Each layer consists of one or more nodes, as shown in the below diagram by small circles. The lines between the nodes indicate the flow of information from one node to the next. The information flows from the input to the output,

i.e. from left to right (in some cases, it may be from right to left or both ways).

* PYSpark techniques using for income analysis:

PySpark is the Python library for Apache Spark, an open-source big data processing framework. Spark provides a unified and distributed computing system that enables processing and analyzing large-scale datasets across clusters of computers.And also allows you to interact with Spark using the Python programming language, making it easier for Python developers to leverage the power of Spark for big data processing tasks. It provides an API that allows you to write Spark applications using Python syntax and take advantage of Spark's distributed computing capabilities.With PySpark, you can perform various data processing tasks such as filtering, transforming, aggregating, and analyzing large datasets in a distributed manner. It provides a wide range of built-in functions and libraries for handling structured, semi-structured, and unstructured data, including data from various sources like Hadoop Distributed File System (HDFS), Apache Hive, and more. PySpark supports a variety of data processing operations, including batch processing, real-time streaming, machine learning, and graph processing. It leverages the distributed computing model of Spark, which allows you to scale your applications to process massive amounts of data across multiple nodes in a cluster.

# II. LITERATURE SURVEY

The study likely uncovers a significant correlation between occupation and income, potentially identifying specific professions with higher earning potential. Beyond this correlation, it may delve into the nuanced factors influencing income variation, such as education, experience, and geographic location. By exploring temporal trends, the research could reveal evolving income landscapes within different occupations, aiding individuals and policymakers in anticipating changes. Moreover, an examination of the impact of technological advancements on income distribution offers insights for navigating a dynamically shifting job market. Overall, the study serves as a crucial resource for individuals making career choices and policymakers aiming to strategically bolster economic growth in specific sectors [1]. The study likely involves a comprehensive investigation into the direct and indirect effects of the COVID-19 pandemic on life expectancy. The research is expected to delve into various factors influencing how the pandemic has shaped mortality rates, taking into account not only the immediate consequences of the virus but also the broader societal and economic repercussions. As part of this analysis, the study may explore the differential impacts on various demographic groups, considering factors such as age, preexisting health conditions, and socioeconomic status[2]. SVM demonstrated high accuracy in analysing the income and The study appears to be focused on unraveling the intricate web of factors that exert a significant influence on income levels among adults. It likely employs a multifaceted approach to understanding the dynamics of income determination, with a keen interest in various socio-economic and occupational variables. Education is likely a key focal point, The research probably investigates the relationship between educational attainment and income, exploring how higher levels of education correlate with increased earning capacity[3]. The research project indeed seems to focus on the evaluation of various machine learning models, particularly in the context of optimizing features for accurate salary predictions. It suggests a systematic exploration of different algorithms to discern their effectiveness in handling the complexities of predicting income levels. In addition to the general assessment of machine learning models, there is a specific interest in the performance of Support Vector Machines (SVMs) in this predictive task [4]. Support Vector Machines (SVMs), the research might explore their application in regression tasks for house price prediction. SVMs, typically recognized for their effectiveness in classification tasks, can be adapted for regression by optimizing the prediction of continuous values. The paper could detail how SVMs are employed to capture complex relationships between various features and housing prices, particularly focusing on their ability to handle non-linear patterns in the data [5]. This study uses human mobility data from Shenzhen to estimate income distribution for sustainable urban development. Three representations of mobility, integrated into models like XGBoost and CNN, outperform traditional approaches. Graph-based representations and deep learning models minimize information loss and handle complex structures. Spatial attributes, like transport accessibility, play a key role, while activity-related indicators contribute less [6]. In Indonesia, this research proposes a framework using e-commerce data and Support Vector Machines (SVM) to estimate city-level poverty rates. The study highlights cars and motorbikes as significant predictors, with experimental results indicating SVM's effectiveness in using e-commerce data for accurate poverty rate calculations [7]. This study focuses on crime prevention in India using machine learning algorithms like Naive Bayes, Support Vector Machine, Linear Regression, Decision Tree, Bagging Regression, Stacking Regression, and Random Forest. The proposed technique achieves a high 99.9% classification accuracy on test data, surpassing baseline studies and demonstrating strong predictive power, aligning with criminological theories for accurate crime predictions [8]. This paper investigates India's population impact on Uttarakhand, emphasizing natural resources, poverty, unemployment, and state growth. Using machine learning and data visualization, it quantitatively assesses education, income inequality, and mortality relationships. Achieving 85.78% accuracy with the Gradient Boosting Classifier and Support Vector Machine models [9]. This paper focuses on optimal locations for new Electric Vehicle Charging Stations (EVCS) in Dubai, UAE, utilizing machine learning models like Knearest neighbors and Support Vector Machines (SVM). Features include population density, points of interest (POI), and security cameras. Performance evaluation indicates an 89% validation accuracy with the KNearest Neighbors model. [10].

III. METHODOLOGY

**Data Preparation:**

1. **Setting the Stage:** We kick off by importing necessary libraries like PySpark ML and DL, establishing a Spark session for distributed data handling.
2. **Data Ingestion:** The census data is then loaded into a PySpark DataFrame for efficient manipulation.
3. **Data Refinement:** Missing values are addressed through imputation or deletion, while categorical variables are transformed into numerical representations using techniques like one-hotencoding.
4. **Training & Testing:** Finally, the prepped data is split into training and testing sets (e.g., 70/30 split) to facilitate model evaluation.

**Machine Learning Exploration:**

1. **Model Selection:** K-Nearest Neighbors (k-NN) and Support Vector Machine (SVM) algorithms are chosen for their effectiveness in income prediction tasks.
2. **Hyperparameter Tuning:** We meticulously tweak hyperparameters specific to each model (e.g., k for k-NN, kernel & regularization parameters for SVM) through grid or randomized search to optimize performance.
3. **Model Training:** Both k-NN and SVM models are trained on the training set, internalizing income-influencing patterns.
4. **Model Evaluation:** Their performance is rigorously assessed on the unseen testing set using metrics like mean squared error (MSE) or R-squared for regression tasks.
5. **Champion Selection:** By comparing their performance, we identify the ML model that delivers the most accurate income predictions.

**Deep Learning Delving:**

1. **Feature Engineering:** Beyond raw data, we extract relevant features like interaction terms between categorical variables, potentially holding valuable insights into income variation.
2. **Model Architecture:** A suitable deep learning architecture, like a Multilayer Perceptron (MLP) or a Convolutional Neural Network (CNN), is chosen based on the extracted features' nature.
3. **Hyperparameter Tuning:** Hyperparameters of the chosen architecture (e.g., number of layers, neurons per layer, learning rate) are meticulously adjusted to refine its predictive power.
4. **Model Training:** The deep learning model is then trained on the training set with the extracted features, learning complex relationships that influence income.
5. **Model Evaluation:** Its performance is evaluated on the testing set using the same metrics as the ML models, allowing for a fair comparison.
6. **Overall Champion:** The best performing model, whether from ML or deep learning, is crowned the overall champion for income prediction accuracy.

# REFERENCES

1. A. U. Rehman, R. M. Saleem, Z. Shafi, M. Imran, M. Pradhan and H. M. Alzoubi,2023, "Analysis of Income on the Basis of Occupation using Data Mining," 2022 International Conference on Business

Analytics for Technology and Security (ICBATS) .

1. M. Najibullah Schwandt H, Currie J, von Wachter T, Kowarski J, Chapman D, Woolf SH.2023, Changes in the Relationship Between Income and Life Expectancy Before and During the COVID-19 Pandemic.
2. E. E. Moe, S. S. M. Win and K. L. Lai Khine 2022, "Adult Income Classification using Machine Learning Techniques," 2023 IEEE Conference on Computer Applications (ICCA) .
3. L. Pawar, A. K. Saw, A. Tomar and N. Kaur 2022, "Optimized Features Based Machine Learning Model for Adult Salary Prediction," 2022 IEEE International Conference on Data Science and Information System (ICDSIS)
4. K. Lakshmi., P. V. Narayana, P. D. Bhavani, V. V. S. Madhavacharyulu, N. Lavanya and S. Pousia,2022, "An Enhanced Regression Technique for House Price Prediction," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I)
5. Chen Zhong b,\*, Yang Yue c, Rui Cao d, Bowen Zhang Shenzhen Key Laboratory of Spatial Smart Sensing & Department of Urban Informatics, Shenzhen University, Shenzhen, 518060, Guangdong, China d Department of Land Surveying and Geo Informatics & Smart Cities Research Institute, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
6. Ana Uluwiyah,Muhammad Rheza,Annisa Zahara ,Dwi Rani Puspita https://link.springer.com/article/10.1007/s10660-020-09424-1.
7. S. Khatun, K. Banoth, A. Dilli, S. Kakarlapudi, S. V. Karrola and G. C. Babu, "Machine Learning based Advanced Crime Prediction and Analysis," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 90-96, doi: 10.1109/ICSCDS56580.2023.10104655.
8. G. D. Singh, H. Vig and A. Kumar, "A data visualization approach for predicting the income class of the population," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1042-1047, doi: 10.1109/ICECA52323.2021.9675850
9. M. C. El Rai, S. A. Hadi, H. A. Damis and A. Gawanmeh, "Prediction of Electric Vehicle Charging Stations Distribution Using Machine Learning," 2022 5th International Conference on Signal Processing and Information Security (ICSPIS), Dubai, United Arab Emirates, 2022, pp. 154-157, doi: 10.1109/ICSPIS57063.2022.10002556