

A project report on

Feature Selection and Extraction of Parameters for Income Prediction Using Machine Learning Techniques

Submitted to



MVP Samaj's

**K.R.T. Arts, B.H. Commerce and A.M. Science College,
Nashik**

Affiliated to Savitribai Phule Pune University, Pune

Department of Statistics



- Submitted by

**Anap Pranit Ravindra
Khairnar Gayatri Deepak
Wagh Omkar Tejendra**

Under the Guidance of

Dr. Nutan V. Khangar
Assistant Professor,
Department of Statistics,
KTHM College, Nashik

May, 2023

CERTIFICATE

This is to certify that the project entitled “**Feature Selection and Extraction of Parameters for Income Prediction Using Machine Learning Techniques**” is being submitted by Anap Pranit Ravindra, Khairnar Gayatri Deepak, Wagh Omkar Tejendra as partial fulfilment for the degree of the Master of Science (Statistics).

This is a record of considerable work carried out by them under my supervision and guidance.

Place: Nashik

Date:

Project Guide

Dr. Nutan Khangar

Head

Department of Statistics,
K.T.H.M. College, Nashik

Examiner

Declaration By Students

We M.Sc. - II (Statistics) students hereby declare that the Project entitled “**Feature Selection and Extraction of Parameters for Income Prediction Using Machine Learning Techniques**” is our original work and has not been submitted for any degree, diploma or other similar titles elsewhere. This has been undertaken for the purpose of partial fulfilment of Post Graduate in Statistics during the year 2022-2023 at MVP Samaj’s K.R.T. Arts, B.H. Commerce and A.M. Science College, Nashik.

Date: 18 / 5 / 2023

Place: Nashik

Anap Pranit Ravindra

Khairnar Gayatri Deepak

Wagh Omkar Tejendra

Acknowledgement

We have satisfaction upon completion of the project work entitled “*Feature Selection and Extraction of Parameters for Income Prediction Using Machine Learning Techniques*” at Department of Statistics in “**M.V.P. Samaj’s K.R.T. Arts, B.H. Commerce, A.M. Science College, Nashik**” during the academic year of 2022-2023.

We take this opportunity to express our deep sense of gratitude to Dr. Nutan V. Khangar for her valuable guidance and help provided to us during the course of completion of project.

Also, we are thankful to all faculties of the Department of Statistics for their valuable guidance and great co-operation.

My special thanks goes to my friends and my family, without their moral support this project could not be completed.

Anap Pranit Ravindra

Khairnar Gayatri Dipak

Wagh Omkar Tejendra

Abstract

Income inequality is a significant issue impacting individuals and societies. To understand its contributing factors, we analyzed the UCI adult dataset, also known as the "Census Income" dataset. With around 49,000 individuals, it provided information on age, education, occupation, and income level. Our project aimed to evaluate demographic variables to promote economic equality.

We focused on predicting whether a person earns over \$50,000 annually using seven machine learning models: Logistic regression, SVM, Random Forest, Bagging, Gradient Boosting, AdaBoosting, Decision Tree, Gaussian Naïve Bayes, and Multinomial Naïve Bayes. After performing feature selection, we identified relationship, age, education, and marital status as crucial factors in determining income class.

Among the models we applied, boosting (Adaboost) stood out with the highest AUC value and accuracy. This finding suggests that Adaboost is the most effective model for accurately fitting the. Analyzing income has significant importance in addressing economic inequality, and our project contributes to understanding the various factors influencing income levels.

Index

Sr.No.	Topics	Page No
1	Introduction	1
2	Literature Review	2
3	Objectives	3
4	Methodologies	4
5	Data Structure	12
6	Numerical Analysis	17
7	Detailing of Outputs	26
8	Overview of best fitted model	30
9	Conclusion	32
10	Future Work	33
11	Limitations	34
12	References	35