# 

# Census Income

**Blog/article**

## Predicting Income Levels Using Census Data: A Machine Learning Approach

### Introduction

In the era of data-driven decision-making, understanding the socio-economic patterns in population data can offer significant insights for policy-making and business strategies. This project leverages machine learning to predict whether an individual earns over $50K annually based on data extracted from the 1994 Census bureau database. The data was meticulously prepared by Ronny Kohavi and Barry Becker and includes various demographic and employment-related attributes.

### Dataset Overview

The dataset comprises records of individuals with the following selection criteria: age greater than 16, gross income over $100, positive final weight (fnlwgt), and non-zero hours worked per week. The final weight (fnlwgt) is a key attribute derived from the Current Population Survey (CPS) files, controlled to match independent estimates of the civilian non-institutional population in the US.

### Data Preprocessing

A crucial first step in any machine learning project is data preprocessing. This involves cleaning the data, handling missing values, and converting categorical variables to numerical formats. In this dataset, we performed the following preprocessing steps:

* Stripped leading and trailing spaces from column names.
* Verified the integrity of the data, ensuring no missing values.
* Used Label Encoding to convert categorical features into numerical values for easier processing by machine learning algorithms.

### Model Selection

Three machine learning models were considered for this classification task: Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). Each model was trained on the dataset and evaluated based on accuracy, precision, recall, and F1-score.

#### Logistic Regression

The Logistic Regression model achieved an accuracy of approximately 82%, with high precision and recall for predicting individuals earning less than $50K.

#### Random Forest Classifier

The Random Forest Classifier outperformed the other models, achieving an accuracy of approximately 85%. This model provided a good balance between precision and recall for both classes.

#### Support Vector Machine

The SVM model also performed well, with an accuracy of around 84%. However, it lagged slightly behind the Random Forest in terms of overall performance.

### Cross-Validation and Hyperparameter Tuning

To ensure the robustness of our models, we performed cross-validation, which provided average accuracy scores over multiple folds:

* Logistic Regression: 78.7%
* Random Forest Classifier: 85.5%
* Support Vector Machine: 79.5%

Based on these scores, the Random Forest Classifier was selected as the best model. We then fine-tuned this model using GridSearchCV, optimizing parameters like the number of estimators, maximum depth, and minimum samples for splitting and leaf nodes. The tuned model achieved an accuracy of approximately 87% on the test set.

### Conclusion

This project demonstrates the efficacy of machine learning in predicting income levels from demographic and employment data. The Random Forest Classifier, with its robust performance, was identified as the best model for this task. By tuning the model's hyperparameters, we achieved a significant improvement in accuracy, making it a valuable tool for socio-economic analysis. The final model was saved and is ready for deployment, providing a foundation for further enhancements and applications in real-world scenarios.

### Future Work

Future enhancements could include:

* Incorporating additional features or external data sources to improve model accuracy.
* Applying advanced techniques like ensemble learning to combine the strengths of multiple models.
* Deploying the model as a web service for real-time predictions.

This project highlights the power of machine learning in deriving meaningful insights from complex datasets, paving the way for more informed decision-making in socio-economic planning and analysis.

4o