**PROJECT REPORT**

**INCOME PREDICTION USING RANDOM FOREST**

**1.1 Introduction**

Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick. Python can be treated in a procedural way, an object-orientated way or a functional way.



Fig: Features of python

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally this study outputs intelligent software systems. The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving.

The intelligence is intangible. It is composed of

* Reasoning
* Learning
* Problem Solving
* Perception
* Linguistic Intelligence

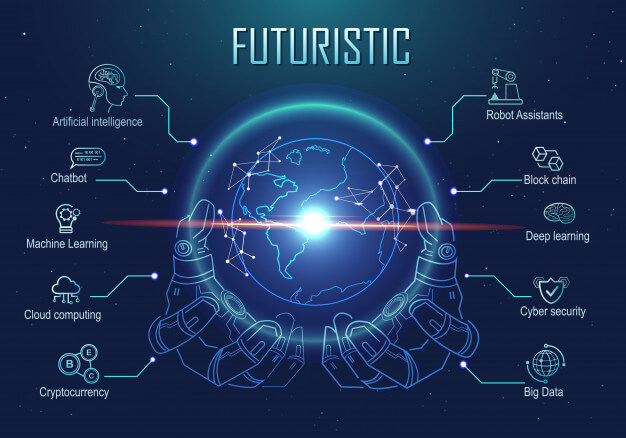


Fig: A glimpse of Artificial Intelligence.

**1.2 Objective of Research**

The main objective of income prediction is used to determine the past financial performance of the employee, predict future performance, and assess the capability of generating future. The objective considers the person working hours, experience and so on, based on these criteria's prediction is evolved.

It also includes the survival of the person whether they can spend more on their

necessities or not. The other objective of prediction is to know the status of the person and to know the ability of a person. If the person could not predict his/her salary then it includes a reduction in the rate of savings and increase in the debt.

**1.3 Problem Statement**

Few rich people appear in the lowest income group. It has been witnessed that older working people have higher properties and income than their younger colleagues. On the other hand, people who were retired from their job consume more wealth and lesser income than the youngsters. Thus the relationship between income and wealth is not strong.

To solve these predictions we must use some algorithms. Here we used Random Forest algorithm for our prediction. **Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of it’s simplicity and the fact that it can be used for both classification and regression tasks.** Random Forest is a supervised learning algorithm. The Random Forest is a combination of multiple trees. Each tree is provided with some random sample of data with replacement and gives different classifications. It takes votes from the result of all the trees and chooses the classification having maximum votes and when the dependent variable is continuous, it takes the mean from the outputs given by different trees.

**2. REVIEW OF LITERATURE**

Individual consumer income information enables the banks to validate application data, target new prospects, and segment existing customers. Income can be understood as the summation of all earnings (wages, salaries, profits, interests, payments, rents etc.) in a given period of time (generally a month). Income is a crucial demographic element that is used at a wide variety of customer touch points. Therefore, it is very important to have an income prediction for existing and potential customers. However, accurate indicators of income are difficult to collect but essential for companies that want to create high-quality revenue budgets, especially in an uncertain economic environment with changing government policies. Increased competition have made banks and financial institutions to search for new ways to minimize the denial of credit to creditworthy customers and to keep out fraudulent ones as far as possible.

In the literature there are only a few empirical studies on income prediction. This is justified by the fact that it is highly difficult to get exact information about individual income, wealth and their characteristics. Here we concentrate on individual income models and predictions but the modelling is often limited by lack of adequate data. For example, as Carrier and Sand argued that employers do not easily volunteer to give salary data. Bone and Mitchell showed that obtaining more appropriate data and good model for retirement income elements can lead to better estimation. Thereby, Using sample data, Dominitz presented income prediction by comparing datasets from 1993 and 1994. He finds that income expectations were optimistic on average. Therefore, it is very important to have an income prediction for existing and potential customers. However, accurate indicators of income are difficult to collect but essential for companies that want to create high-quality revenue budgets, especially in an uncertain economic environment with changing government policies.

**3. DATA COLLECTION**

The data which is given to the algorithm plays a major role in prediction mechanism. So, it is important to find the correct suitable dataset to our problem. After finding the dataset we must have clear view about our dependent variable that is what needs to be predicted and independent variables that is what are the factors on which the dependent variable is dependent on. Here in our project we have searched for a dataset that will be useful to predict the income of employee and below are the list of fields we have in the dataset. These are the independent variables on which the income depends on.

**Age:** This is to know the age of employee.

**Education:** This field includes employee education background based on which he has his work.

**Work class:** It comprises of the type of work that the employee is engaged in i.e., whether he is working in a private company or government or he is an entrepreneur.

**Gender:** This field is to classify the employees as female and male

**Capital-gain:** Capitalgain is additional increment in the income of the employee based on his working. This also plays a role in the income prediction.

**Capital-loss:** A capitalloss is the decrementing amount that is deduced from the income of person as a result of his actions.

**Hours-per-week:** This field is to find the number of hours an employee works in a week and based on the hours and other criterion the income of the employee is decided

**Native Country:** This field is related to the country which the employee belongs to.

**4. METHADOLOGY**

**4.1 Exploratory Data Analysis:**

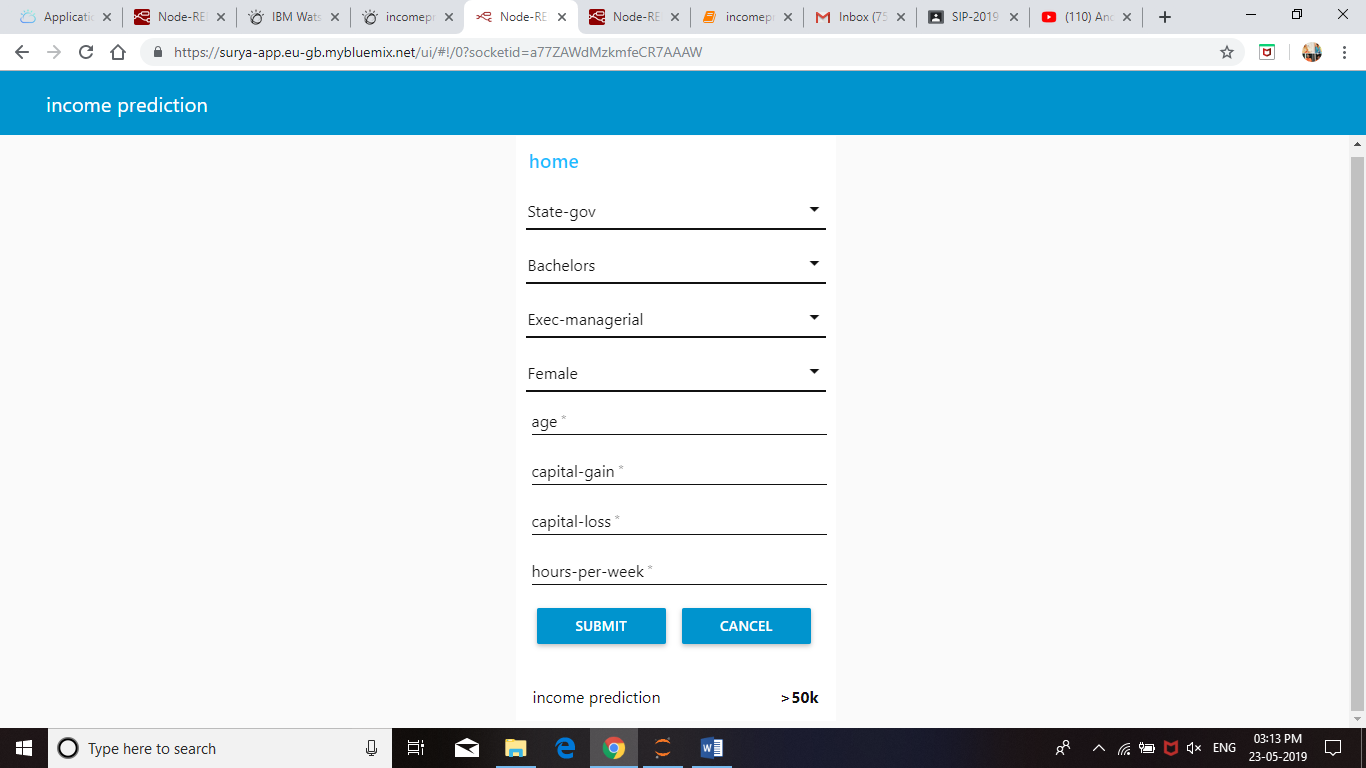


Fig 1: Predicting the income as greater than 50k based on the given parameters

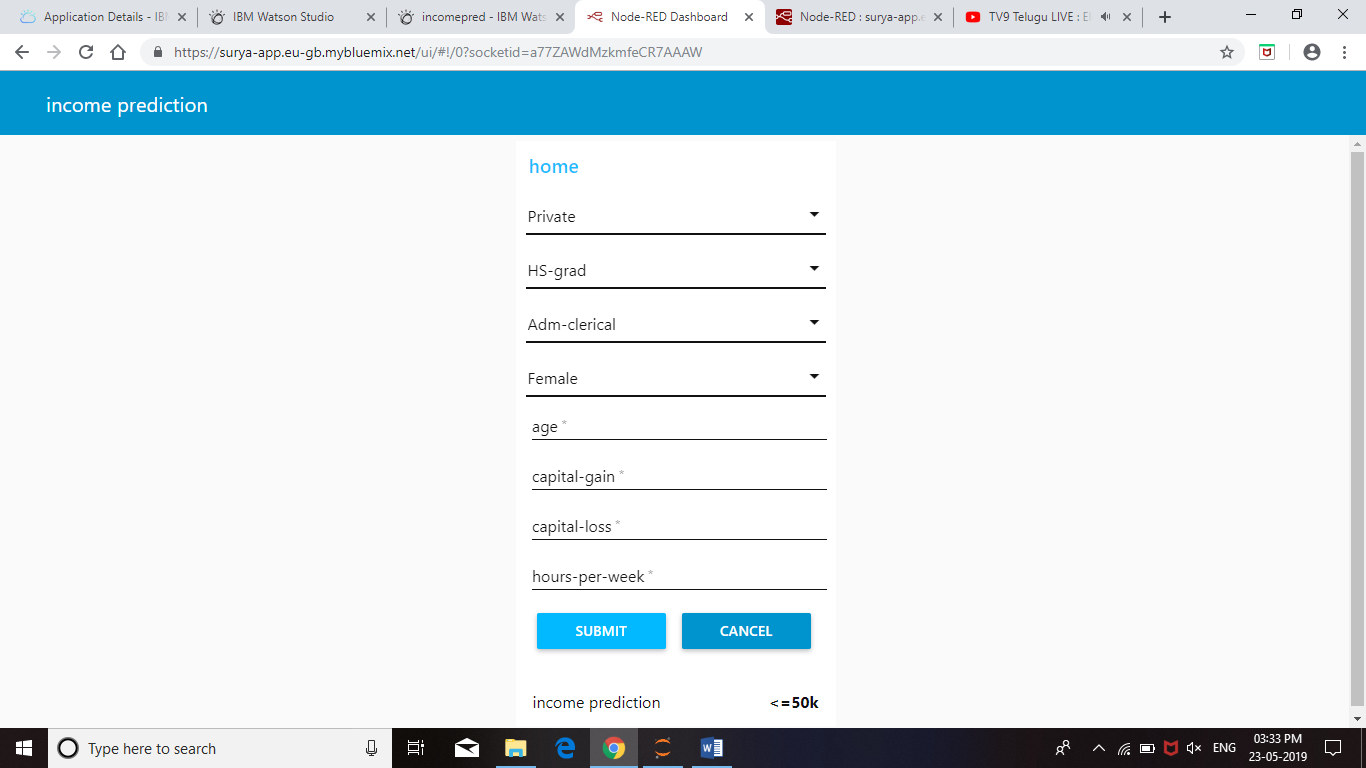


Fig 2: Predicting the income as less than or equal to 50K based on the given inputs

**4.2 Data Modelling:**

In this project in order to predict the income we have used random forest algorithm. This algorithm is a machine learning classification algorithm which classifies the input based on the majority of votes for classes. The input is labelled with the class name which gets the majority of votes.

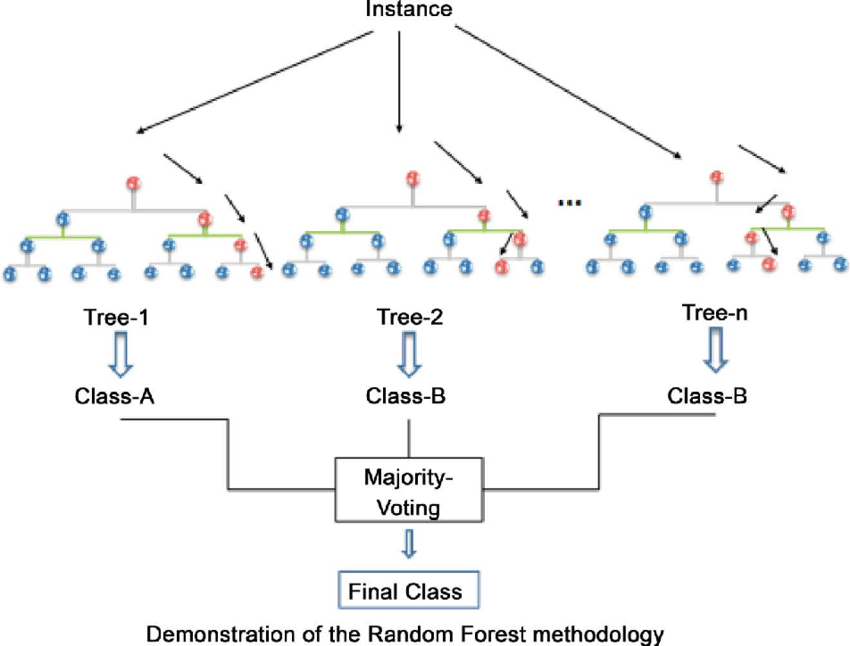


Fig 3: Random Forest algorithm

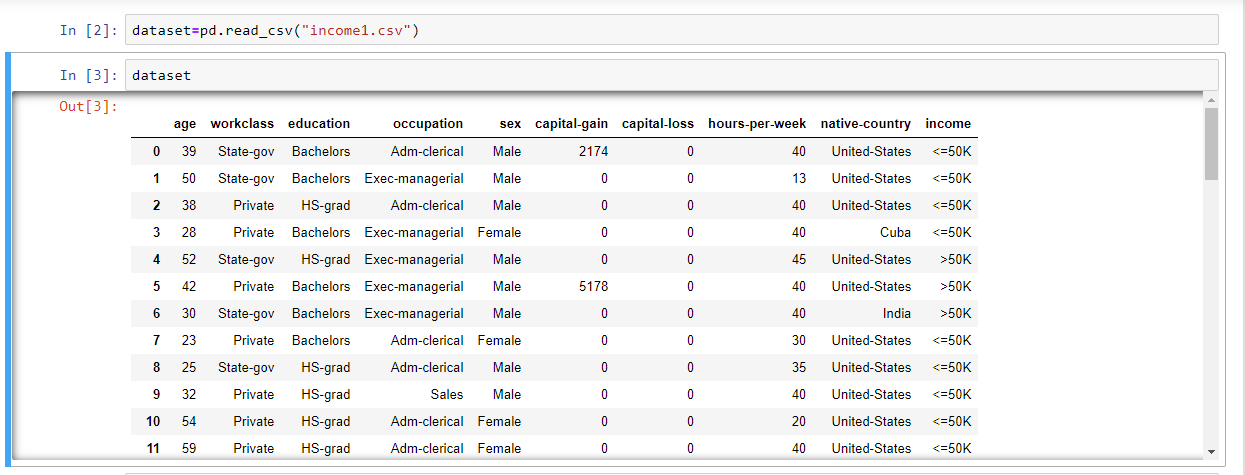


Fig 4: The dataset used in the algorithm.

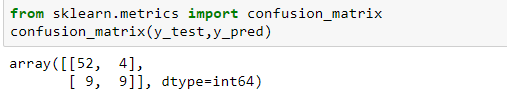


Fig 5: Confusion matrix to find the accuracy of the prediction.

**5. REFERENCES**

1. <https://www.kaggle.com/>

2. <https://www.analyticsvidhya.com/>

3. <http://r-project.com/>

4.<https://www.youtube.com/results?search_query=random+forest+algorithm+in+machine+learning>

5. <https://www.wikipedia.org/>

6. <https://archive.ics.uci.edu/ml/datasets.php>

7. <https://data.gov.in/>

8. <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

9. <https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/>

10. <https://github.com/mljs/random-forest>

**6. CONCLUSION**

In this project we constructed a model that is able to accurately predict whether an individual earns over $50k per year based on various data like work class, education, age, nationality, gender, working hours of employee per week and many more. This prediction is done by using random forest algorithm. Random forest algorithm is a classification algorithm where the decision is made based on the majority of votes. This prediction model might be useful to estimate the number of employees with high income and number of employees of low income. Also we can use this income prediction for market analysis so that based on the average income of employee we can make some changes in our business to meet the needs of the majority of the people. The model might be tuned to meet the business needs which could be to classify high-income customers more accurately.