**Industrial Internship Report**

**On**

**Salary Estimation Using Machine Learning**

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

Submitted By

**Ashwini Tarun**

Roll No.  **19671A1207**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**J.B. INSTITUTE OF ENGINEERING & TECHNOLOGY**

**UGC AUTONOMOUS**

**(Accredited by NAAC,Permanently Affiliated to JNTUH)**

**Bhaskar Nagar, Moinabad Mandal, R.R. District,Telangana State,500075**

|  |  |
| --- | --- |
| C:\Users\Department-IT\Desktop\download.jpg | **J.B.Institute of Engineering & Technology**  **(UGC Autonomous)**  **(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)** |

CERTIFICATE

This is to certify that of III B. Tech (Information Technology) bearing Roll No**- 19671A1207** has completed his of internship “**Salary Estimation using Machine Learning** **“**being submitted to the Department of information Technology, J.B Institute of Engineering and Technology, in accordance with JNTUH regulations as partial fulfillment required for successful completion of Bachelor of Technology in Information Technology is a bonafide work done during the academic year 2021-22 carried out by **ASHWINI TARUN** bearing HT.NO: 19671A1207.

**Head of the Department**

**Mr. M. Ravi Kumar**

**Asst. Professor** **&HOD**

**ACKNOWLEDGEMENT**

I would like to take this opportunity to thank the **M.Ravi Kumar** Head of Department of , Information Technology, for making the industrial Training a part of the Curriculum and possible for the students.

I would like to express my gratitude to **M.K. Jeeva Rajan Pantech Prolabs India Pvt Ltd,** for allowing me with the golden opportunity of taking the internship.

And of all I am highly indebted to the principal **Dr. P.C. Krishnamachary,** for giving me permission to carry out this Internship.

I also take this opportunity to thank all staff who have directly or indirectly helped in my project. I pay my respects and love to my parents and all other family members and friends for their love and encouragement throughout career.

**Ashwini Tarun**

**CERTIFICATE**

Graphical user interface, application

Description automatically generated

**COMPANY PROFILE**

A picture containing text, tool

Description automatically generated

Pantech Prolabs India Pvt Ltd is best known for manufacturing and marketing of high-quality, state-of-the-art Engineering Lab Equipments and Electronic Hobby Kits. Its products are delivered across the globe and many satisfied customers are the best guarantee of its first-rate service. Many of our products are also extensively used by Hobbyists, Electronics, Electrical, Instrumentation, Communication and Bio Medical engineers.

With a client list spanning nearly in all industries, and colleges, Pantech Solutions’ product solutions have benefited customers of many different sizes, from non-profit organizations to companies. By acquaintance with Pantech you’ll have access to current technology, Development tools, Reference Schematics, sample source code, and step-by-step action plans for completing Key projects. You’ll also be provided full access to our research archives and knowledge base.

## Our Vision

**“To Gain Global Leadership in providing Technological Solutions through Sustained Innovation.”**

Core Values

You’ll see that our 6 core values are derived from our stewardship quality.

* **Integrity** – Honesty in how we deal with our clients, each other and with the world.
* **Candor** – Be open and upfront in all our conversations. Keep clients updated on the real situation. Deal with situations early; avoid last minute surprises.
* **Service** – Seek to empower and enable our clients. Consider ourselves successful not when we deliver our client’s final product but when the product is launched and meets success.
* **Kindness** – Go the extra mile. Speak the truth with grace. Deliver more than is expected or promised.
* **Competence**– Benchmark with the best in the business. Try new and better things. Never rest on laurels. Move out of comfort zones. Keep suggesting new things. Seek to know more.
* **Growth** – Success is a journey, not a destination. Seek to multiply/increase what we have – wealth, skills, influence, and our client’s business.

**Training Details**

* Name:- Ashwini Tarun
* Year:- Third Year
* Project:- Salary Estimation Using Machine Learning
* Organization:- Pantech Prolabs India Pvt Ltd
* Place of Training:- Online
* Supervisor:- Sanjay Kumar A P
* Period:- 03’Nov 2021 to 05’Feb 2021
* Duration:- 1 month
* Field of training:- Python, Machine Learning, scikit-learn, pandas, Numpy.

**CONTENTS**

**TITLE PAGE NO**

**Certificate II**

**Acknowledgement III**

**Certificate of Internship IV**

**Company Profile V**

**Training Details VI**

1. **Abstract 1**
2. **Problem Statement 1**
3. **Objective 1**
4. **System Requirements 2**
5. **Model Requirements 2**
   1. **Pandas 2**
   2. **Numpy 2**
   3. **Matplotlib 2**
   4. **Seaborn 3**
6. **System Architecture 3**
7. **Activity Diagram 4**
8. **Flow chart 4**
9. **Gathering the required dataset 6**
10. **Preprocessing the dataset 7**
11. **Splitting the dataset 7**
12. **EXPLORATORY DATA ANALYSIS (EDA) 8**
    1. **Mutual Information 8**
    2. **Histogram 9**
    3. **Boxplot 11**
    4. **Scatterplot 12**
    5. **Heatmap 14**
13. **Model Compression 15**
    1. **Linear Regression 15**
    2. **Support Vector Regression 20**
    3. **Decision Tree Regression 21**
    4. **Random Forest Regression 24**
14. **Graphical User Interface 27**
15. **Streamlit 29**
    1. **Command to run Streamlit 29**
    2. **Predication Page in Streamlit 29**
    3. **Data Visualizing page in Streamlit 30**
16. **Sample code 31**
17. **Conclusion 31**
18. **ABSTRACT**

This is a project is which helps common students to estimate their salary by consider some data such as their gender, 10th grade score, 12th grade score, b-tech score and so on. The tool also helps student to estimate there salary so that they can negotiate the salary during their interview and get hired with a good salary package and develop their career.

This tool used Linear, SVR, Decision Tree, Random Forest algorithm to predict the best salary. But among all the algorithms Random Forest algorithm is the one best algorithm which predicts the salary with more accuracy. And it also uses a Graphical User Interface (GUI) which was made by using streamlit that maintains the interaction with the real time users.

1. **PROBLEM STATEMENT**

The problem of a statement is to help students to estimate the salary so that they can estimate salary suited for their skills and negotiate for their salary during their interview process.

1. **OBJECTIVE**

The primary objective of this project/tool is to help students to estimate their salary by asking some data. This helps students to negotiate with their salary during the interview time.

1. **SYSTEM REQUIREMENTS**

* Modern Operating System:
* Windows 7 or 10
* Mac OS X 10.11 or higher, 64-bit
* Linux: RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu)
* x86 64-bit CPU (Intel / AMD architecture)
* 4 GB RAM
* 5 GB free disk space

1. **MODULES REQUIREMENTS**
2. **sklearn**

Scikit-learn is an indispensable part of the Python machine learning toolkit at JPMorgan. It is very widely used across all parts of the bank for classification, predictive analytics, and very many other machine learning tasks.

1. **Pandas:**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

1. **Numpy:**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

1. **Matplotlib:**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

1. **Seaborn:**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

1. **SYSTEM ARCHITECTURE**

Diagram

Description automatically generated

1. **ACTIVITY DIAGRAM**

Diagram

Description automatically generated

1. **FLOW CHART**

**Diagram

Description automatically generated**

1. **GATHERING THE REQUIRED DATA**

We have collected this dataset from the website called as Kaggle. Kaggle is a platform that has a numerous datasets which is popular and commonly used by many Data Science and ML professionals and newbies to ML and Data Science so to learn Data Science and Machine Learning.

Graphical user interface, application

Description automatically generated

To know more about Kaggle, please click [here](https://www.kaggle.com/).

1. **PREPROCESSING THE DATASET**

Preprocessing is one of the very crucial phase and much time consuming phase. Because their can be a lot of noise in data, missing data, categorical data which may throw our problem during the model building the phase. So every dataset should go preprocessing before other steps at least once so that they can become noise-free, error free.

There are also many ways to handle with such kind of missing, noisy data such as

1. Drop Categorical Variables
2. Ordinal Encoding
3. One-Hot Encoding

To know more about handling categorical data, please click [here](https://www.kaggle.com/alexisbcook/categorical-variables).

1. **SPLITTING OF DATA**

Graphical user interface, text, application, email

Description automatically generated

1. **EXPLORATORY DATA ANALYSIS (EDA)**

Exploratory Data Analysis, simply EDA is a way analyzing the data set using the graphs such has Scatterplot, Regression line, Box plot and many more.

In this project we used Mutual Information technique to know that which feature gives more information about dependent feature.

**12.1** **MUTUAL INFORMATION**

The mutual information (MI) between two random variables captures how much information entropy is obtained about one random variable by observing the other. Since that definition does not specify which is the observed random variable, we might suspect this is a symmetric quantity.

Chart

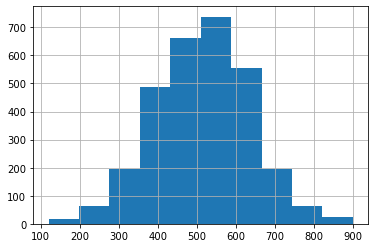
Description automatically generated

To know more about the MI please click [here](http://www.scholarpedia.org/article/Mutual_information#:~:text=Mutual%20information%20is%20one%20of,variable%20given%20knowledge%20of%20another.).

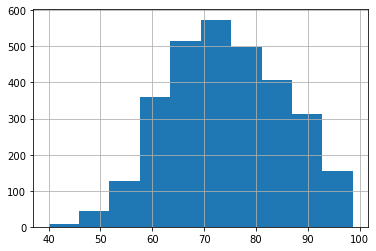
From the above graph we can conclude that [ Quant, Graduation Year, experience, 12percentage, 10percentage ] these features contribute more information to dependent value. But other features are approximately give the same level of information so therefore we can say that all most all the feature paly a vital role during the prediction.

**12.2** **HISTOGRAMS**

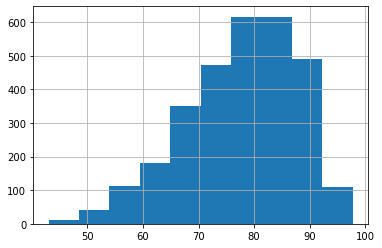
1. **Histogram of Quant Feature**

****

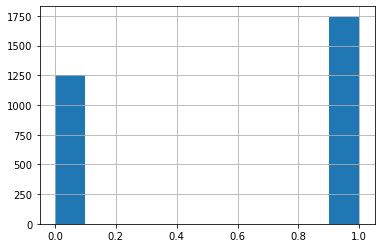
1. **Histogram of 12th grade percentage year Feature**

****

1. **Histogram of 10th grade percentage year Feature**

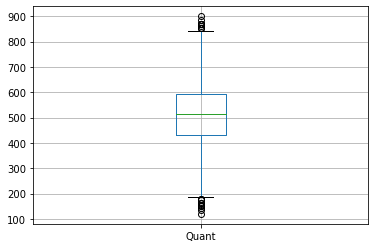
****

1. **Histogram of experience**

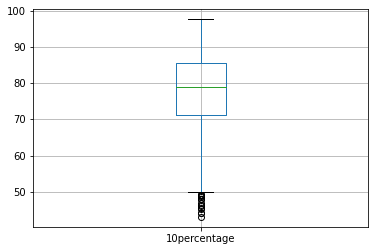
****

**12.3** **BOXPLOTS**

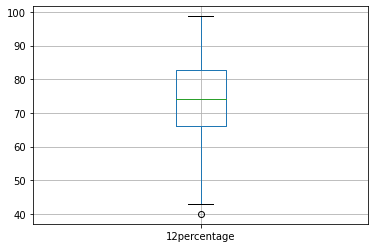
1. **Boxplot of Quant**

****

1. **Boxplot of 10th grade percentage year Feature**

****

1. **Boxplot of 12th grade percentage year Feature**

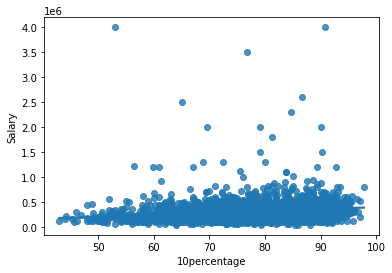
****

**12.4** **SCATTER PLOT**

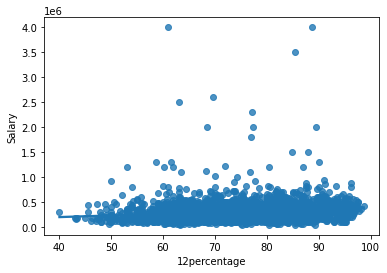
1. **Scatterplot between Quant vs Salary**

****

1. **Scatterplot between 10th percentage vs Salary**

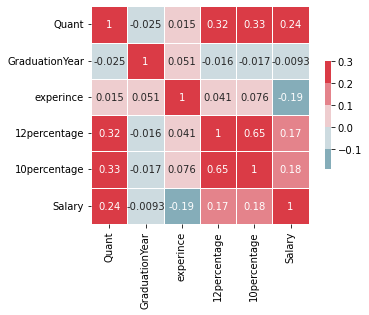
****

1. **Scatterplot between 12th percentage vs Salary**

****

**12.5** **HEATMAP**

1. **Heatmap of the top 5 features with salary**

****

1. **MODEL COMPRISSION**

During building the project we went through many regression-based algorithms like

* 1. **LINEAR REGRESSION**

Linear Regression is a supervised system studying a set of rules in which the anticipated output is non-stop and has a regular slope. It’s used to expect values inside a non-stop range, (e.g. sales, price) as opposed to seeking to classify them into categories (e.g. cat, dog). There are essential types:

**Simple regression**

Simple linear regression makes use of conventional slope-intercept form, in which m and b are the variables our set of rules will attempt to “learn” to provide the maximum correct predictions. x represents our entered information and y represents our prediction.

Text

Description automatically generated

**Multivariable regression**

A extra complex, the multi-variable linear equation would possibly appear like this, wherein w represents the coefficients or weights, our version will try and learn.

**A picture containing text

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated

**Model Evaluation:**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Gradient Descent**

To decrease MSE we use Gradient Descent to calculate the gradient of our fee feature. Gradient descent includes searching at the mistake that our weight presently offers us, the use of the by-product of the fee feature to discover the gradient (The slope of the fee feature the use of our modern-day weight), after which converting our weight to transport withinside the path contrary of the gradient. We want to transport withinside the contrary path of the gradient for the reason that gradient factors up the slope as opposed to down it, so we flow withinside the contrary path to attempt to lower our error.

Diagram

Description automatically generated

To know more about linear regression please click [here](https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86) .

* 1. **SUPPORT VECTOR REGRESSION**

SUPPORT VECTOR REGRESSION,Those who are in Machine Learning or Data Science are quite familiar with the term SVM or Support Vector Machine. But SVR is a bit different from SVM. As the name suggest the SVR is an regression algorithm , so we can use SVR for working with continuous Values instead of Classification which is SVM.

The terms that we are going to be using frequently in this post

1. **Kernel**: The function used to map a lower dimensional data into a higher dimensional data.
2. **Hyper Plane**: In SVM this is basically the separation line between the data classes. Although in SVR we are going to define it as the line that will will help us predict the continuous value or target value
3. **Boundary line**: In SVM there are two lines other than Hyper Plane which creates a margin . The support vectors can be on the Boundary lines or outside it. This boundary line separates the two classes. In SVR the concept is same.
4. **Support vectors**: This are the data points which are closest to the boundary. The distance of the points is minimum or least.

Graphical user interface, text, application, email

Description automatically generated

To know more about SVR please click [here](https://www.mygreatlearning.com/blog/support-vector-regression/#:~:text=SVR%20is%20built%20based%20on,data%20is%20not%20linearly%20separable.).

* 1. **DECISION TREE’S REGRESSION TECHNIQUE**

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

Before learning more about decision trees let’s get familiar with some of the terminologies.

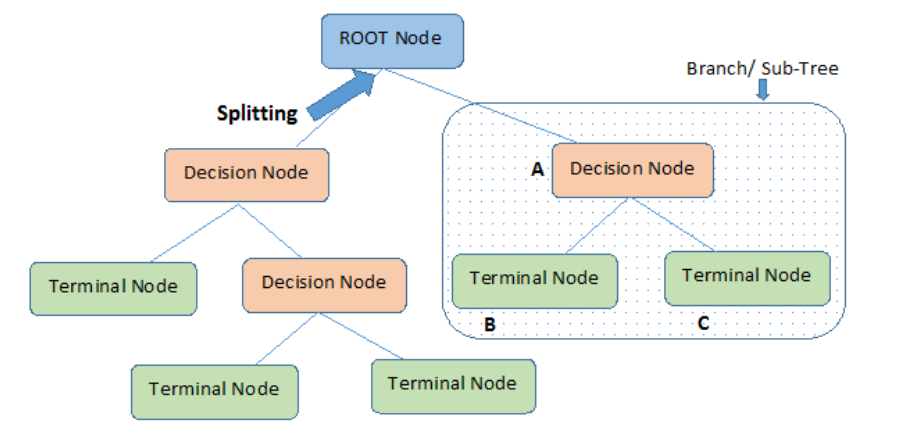
**Root Nodes** – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.

**Decision Nodes** – the nodes we get after splitting the root nodes are called Decision Node.

**Leaf Nodes** – the nodes where further splitting is not possible are called leaf nodes or terminal nodes.

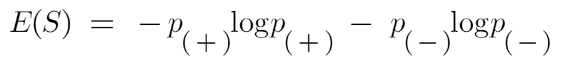
**Sub-tree** – just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.

**Pruning** – is nothing but cutting down some nodes to stop overfitting.



**Entropy:**

Entropy helps us to build an appropriate decision tree for selecting the best splitter. Entropy can be defined as **a measure of the purity of the sub split**. Entropy always lies between 0 to 1. The entropy of any split can be calculated by this formula.



Here p+ is the probability of positive class

p– is the probability of negative class

S is the subset of the training example

**Information Gain**

The information gain is **the amount of information gained about a random variable or signal from observing another random** variable.

information gain Decision tree algorithm

Graphical user interface, text, application

Description automatically generated

To know more about Decision Tree Regression please click [here](https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda).

**13.4** **RANDOM FOREST REGRESSION**

Every selection tree has an excessive variance, however, whilst we integrate them all collectively in parallel then the ensuing variance is low as every selection tree receives flawlessly skilled on that precise pattern records and for this reason the output doesn’t rely upon one selection tree however more than one selection trees. In the case of a class problem, the very last output is taken through the use of the bulk vote casting classifier. In the case of a regression problem, the very last output is the suggestion of all of the outputs. This component is Aggregation.

Diagram

Description automatically generated

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Graphical user interface, text, application, email

Description automatically generated

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique

* Design a specific question or data and get the source to determine the required data.
* Make sure the data is in an accessible format else convert it to the required format.
* Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
* Create a machine learning model
* Set the baseline model that you want to achieve
* Train the data machine learning model.
* Provide an insight into the model with test data
* Now compare the performance metrics of both the test data and the predicted data from the model.
* If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data or using another data modelling technique.

To know more about Random Forest Regression, please click [here](https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/).

At this stage, you interpret the data you have gained and report accordingly.

But among all the features Random Forest algorithm fits the dataset with less Root\_Mean\_Square\_Error (RMSE). RMSE is a best way among all other ways of finding the errors such as Mean\_Absolute\_Error (RME).

**Root\_Mean\_Squared\_Error of all algorithms**

Graphical user interface, application

Description automatically generated

**Root\_Mean\_Squared\_Error best three algorithms**

Graphical user interface, application

Description automatically generated

To know about the various ways Error finding, please click [here](https://scikit-learn.org/stable/modules/model_evaluation.html).

1. **Graphical User Interface ( GUI )**

What is a Graphical User Interface?

A graphics-based operating system interface that uses icons, menus and a mouse (to click on the icon or pull down the menus) to manage interaction with the system.

Graphical user interface

Description automatically generated

In this project we used **Streamlit** plugin to build Graphical User Interface (GUI) of our project. This helps us to makes users to interact with model and tool.

15. **STREAMLIT**

A picture containing text, clipart

Description automatically generated

Streamlit is an open-source python library for creating and sharing web apps for data science and machine learning projects. The library can help you create and deploy your data science solution in a few minutes with a few lines of code.

For accessing streamlit cheat sheet, please click [here](https://docs.streamlit.io/library/cheatsheet).

**15.1** **COMMANDS TO RUN THE STREAMLIT WEB-APPLICATION:**

A screenshot of a computer

Description automatically generated with medium confidence

**15.2** **PREDICATION PAGE GUI**

Graphical user interface, text, application

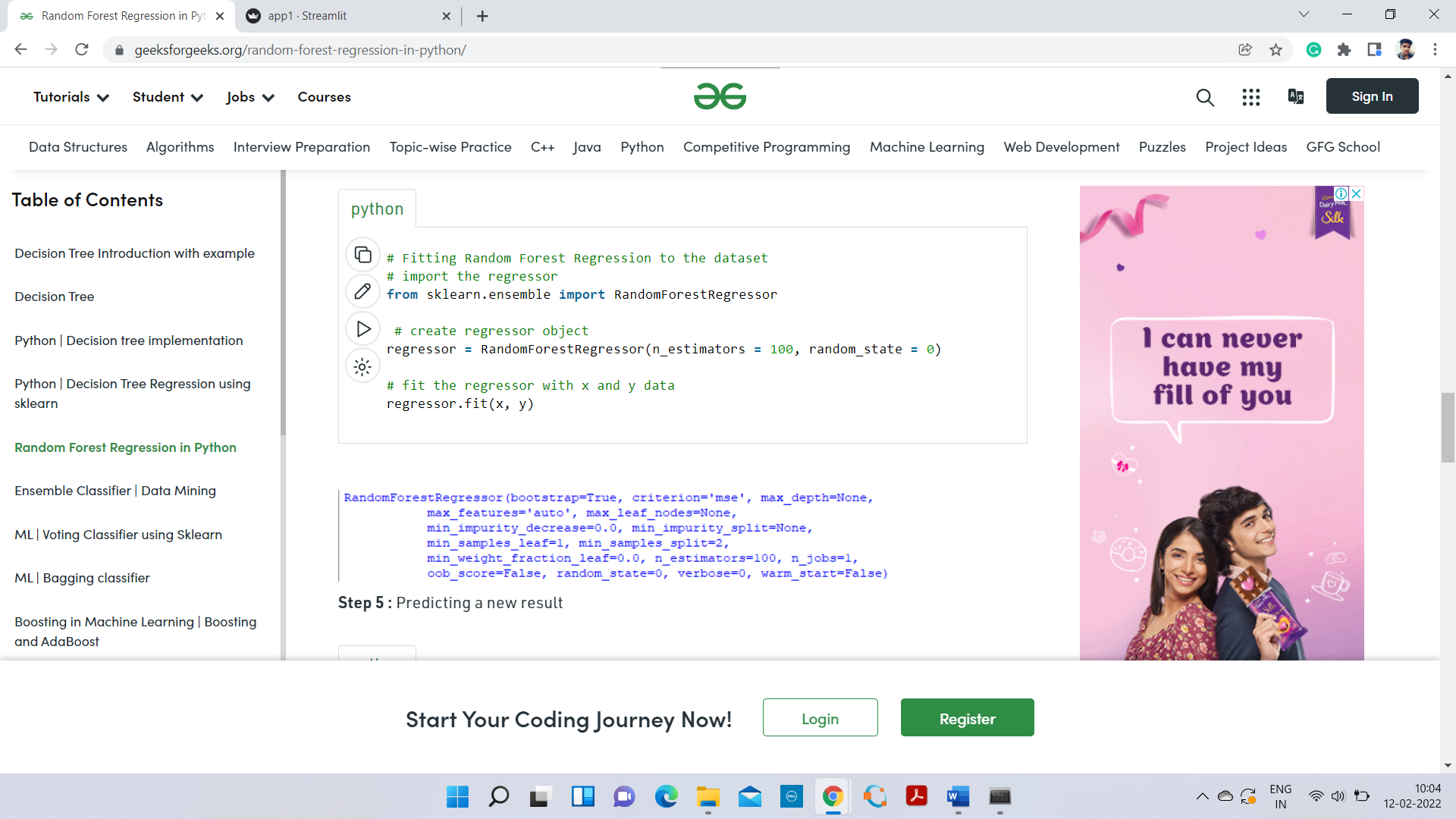
Description automatically generated

**15.3** **EDA PAGE GUI**

Chart

Description automatically generated

1. **SAMPLE CODE**



1. **CONCLUSION**

Finally, we created a tool that predicts the salary of a students when he enters the data like his 10th grade score, 12 grade score and so on. We used Random Forest because it predicts the salary very well and with less RMSE as 11.02 .