

SALARY PREDICTION

SMART INTERNZ PROJECT REPORT SUBMITTED IN
PARTIAL FULFILMENT OF THE REQUIREMENT OF EXTERNSHIP

IN

APPLIED DATA SCIENCE

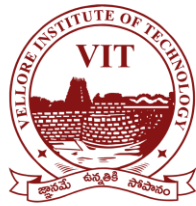
BY

20BCD7190 VARUN SAI

20BCE7129 ADI PRANAV

20BCR7109 AARON PAUL

NITHYA



VIT-AP
UNIVERSITY

ASSITED BY

Prof. SHER KHAN

DECLARATION

I the undersigned solemnly declare that the project report “ **Salary-Prediction**” that's able to predict the probabilities of getting a disease in organs like liver, kidney, heart, etc. relies on my very own work administered during the course of our study under the supervision of Dr. Yugal Kumar.

I assert the statements made and conclusions drawn are an outcome of my Project work. I further certify that

I. The work contained within the report is original and has been done by me under the overall supervision of my supervisor.

II. The work has not been submitted to the other Institution for the other degree/diploma/certificate during this university or the other University of India or abroad.

III. We've followed the rules provided by the university in writing the report.

IV. Whenever we've used materials (data, theoretical analysis, and text) from other sources, we've given due credit to them within the text of the report and given their details within the references.

Varun Sai & Team

20BCD7190

CERTIFICATE

This is to certify that the work which is being presented in the project report titled “ **Salary -Prediction** ” in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by “**Varun Sai and Team**” during the period from MAY 2023 to JUNE 2023 under the supervision of Dr. SHER KHAN.

VARUN SAI 20BCD7190
AARON PAUL 20BCR7109
JAYANA ADI PRANAV 20BCE7129

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

TABLE OF CONTENT

CHAPTER -1 INTRODUCTION	1
1.1 INTRODUCTION	1
1.2 PROBLEM STATEMENT	3
1.3 OBJECTIVE	3
1.4 METHODOLOGY	4
1.4.1 SUPERVISED LEARNING	5
1.4.2 UNSUPERVISED LEARNING	5
1.5 ORGANIZATION	6
CHAPTER-2 LITERATURE REVIEW	7
CHAPTER-3 SYSTEM DEVELOPMENT	9
3.1 FLOW CHART	9
3.3 TECHNICAL REQUIREMENT	13
3.4 DATA SET USED	14
3.5 DATA PROCESSING	15
3.6 DATA VISUALIZATION	16
3.7 DATA CLEANING	17
3.8 MODEL FITTING	17
3.9 VARIOUS STAGES OF THE PROJECT	19
CHAPTER-4 PERFORMANCE ANALYSIS	31
4.1 RESULTS CRITERIA	31
4.2 RESULTS ACHIEVED	34
CHAPTER-5 CONCLUSION	36
5.1 CONCLUSION	36
5.2 LIMITATIONS	36
REFERENCES	37

ABSTRACT

In the knowledge-based industry, compensation planning is an important area for growth and success strategies. In order to retain the most efficient employees, the provision of higher pay is essential. Determining such wages, based on various information about the current employee or prospective employee, is a challenge that organizations face on a regular basis. While HR executives often face such wage-raising issues and discussions in consultation with the appropriate departmental level managers, any automated system with such potential can be of great help to them. Considering employee qualifications (current or expected), including demographic profile and other information such as qualifications, performance level etc. Several well-known split algorithms can be used to predict the income category. But unfortunately, such employee data details of any organization are generally not publicly available to analyze the effectiveness of classification algorithms.

CHAPTER -1

INTRODUCTION

1.1 INTRODUCTION

Human compensation planning has become a key strategic area for firms to ensure sustained growth and profitability, with a heavy focus on the knowledge-based economy. One of the challenges that businesses confront today is retaining competent personnel and hiring skilled people from other companies. Salary becomes a crucial element in attracting current and new employees in both circumstances. As a result, providing full pay, which benefits both the current and future employee as well as the firm, is critical in retaining or attracting people to any organization. Employees have long known that a variety of factors influence an employee's salary and performance in the past, as well as his or her performance during an interview.

With the expansion of businesses, so does the number of employees. Firms essentially raise their employees' compensation in order to keep their talents and raise them. While there are no major obstacles in the way of salary increases in small businesses, this process should be carried out carefully in large firms in accordance with a number of guidelines to avoid negative consequences that could disrupt workforce mobilization. For organizations with a significant number of employees, developing a model in which market conditions are fully defined and all economic elements are taken into account necessitates a months-long procedure. In this regard, research developed a technique for estimating wage increases based on machine learning. Certain characteristics were determined and a specific scale was developed to determine the functional outcome of this study.

A worker's wage is presently the most common reason for them to leave a company.

employees switch occupations frequently with the intention to gain the promised income. And due to the fact that these consequences result in a loss for the corporation, we devised the idea of the employee receiving the desired/anticipated remuneration from the employer or corporation. In trendy aggressive surroundings, all people have elevated expectancies and dreams. We can not, but, offer all of us with their predicted wage at random; alternatively, a device must be in place to assess the worker's capacity to earn the predicted reimbursement. We can't expect precise income, but we are able to estimate it using unique facts units. A forecast is a knowledgeable assumption approximately what will arise within the future.

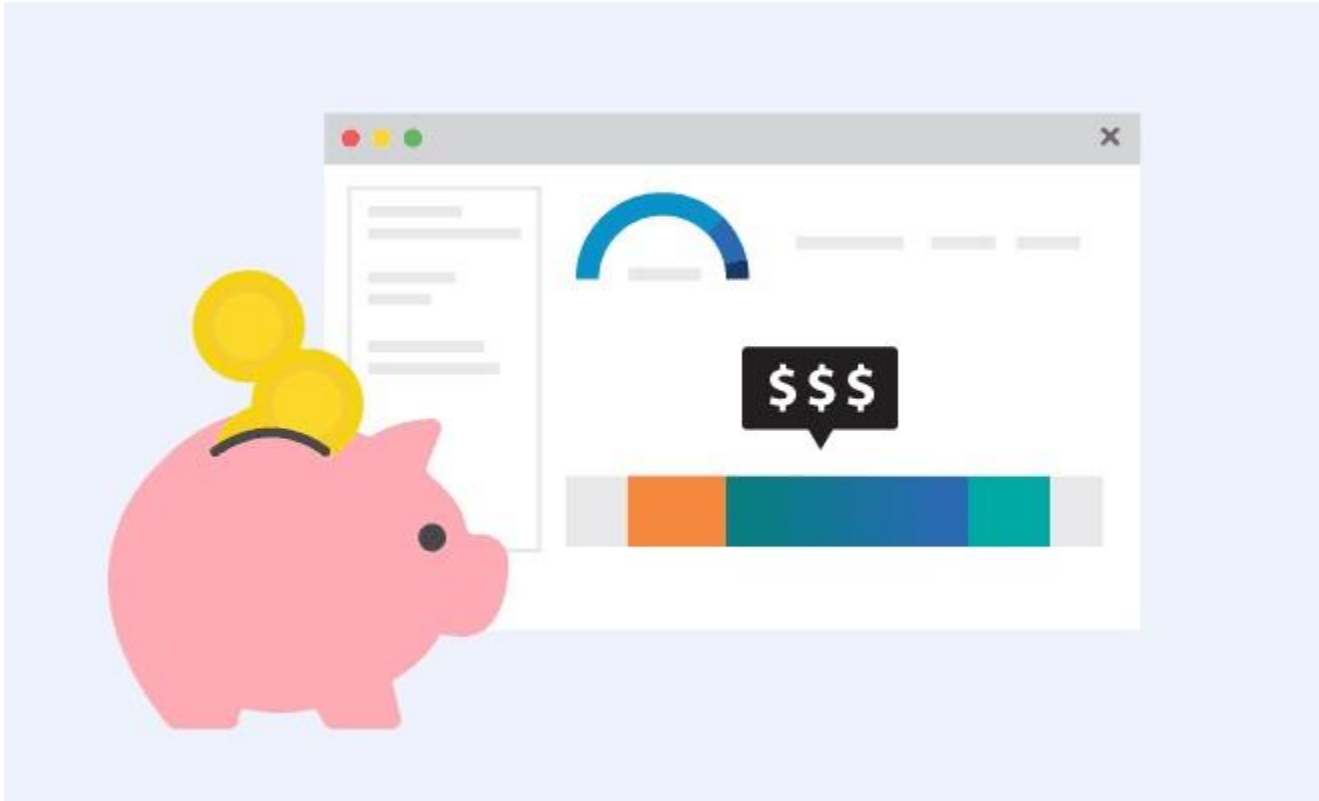


Fig 1 Salary Calculation

Human Resources has seen a significant increase in recent years, as the market has seen a surge in high-quality resources and capabilities, providing a competitive incentive for businesses to spend. Because of its impact on a company's competitiveness and efficiency, proper hiring is more important than ever. This research focused on estimating an employee's new compensation based on a variety of characteristics including age, job class, education, experience, former employment, previous income, and weekly hours. Salary may be successfully predicted using the aforementioned dataset criteria.

1.2 PROBLEM STATEMENT

In order to make the final gift on the job, employers must consider a variety of elements, including demographics and others. Although this work is performed by experienced human resource managers in cooperation with the appropriate departmental level management, it is always a difficult decision. For these decision-makers, any form of automated decision-making system can be highly useful in generating acceptable wage suggestions. Companies typically have their own compensation forecasting system, which uses internal data to forecast new hire earnings. In this regard, research developed a machine learning-based approach for estimating wage increases.

Overspending on new staff is a common concern for businesses. And getting a head start on the new employee's salary expectations will benefit both the person and the organization. Overspending to hire new employees, as we all know, will have a severe impact on the company's financial health.

1.3 OBJECTIVE

Salary is a continuous variable that may be calculated using regression analysis. A model for raw characteristics with two goals: straightforward wage prediction and illustrating the influence of embedded feature selection.

Using job pay datasets, the purpose of this research is to create a strong machine learning model that can predict future employee wages based on current employee salary data.

1.4 METHODOLOGY

In order to obtain useful information on online recruitment of IT professionals, we compare and contrast different strategies with machine learning models. The procedure follows the best practices in literature and industry, which includes different categories:

- I. Data Collection: A Python web browser is designed to analyze and collect information needed for a website.
- II. Correcting Inconsistencies: To remedy the inaccuracies and inconsistencies found in the survey, subject matter knowledge was required. We cleaned up the data by first converting all of the questions into variables, then translating all of the details into English. We put together several different words in the replies that refer to the same term. Corrected errors and misspellings. Finally, commas, apostrophes, quotation marks, question marks, and other punctuation symbols have been eliminated.
- III. Manual feature engineering: Features which are not important are discarded and some are suspended (e.g. converted into numerical features) by misusing domain information.
- IV. Data set definition: To give original and integrated data definition, mathematical techniques and basic models were applied.
- V. Deploying the model: The final project was deployed on localhost.

Machine learning is a branch of artificial intelligence that uses mathematical approaches to make machines more environmentally friendly. permits the active gadget to decide on hosting responsibilities. These programmes, or algorithms, may be used to test and expand over time by examining fresh records. The reason why computers can deduce meaning from records. Statistics are therefore the key to unlocking Machine Learning. A collection of ML rules may be highly useful as the most significant ML data you should have. to learn about a building's construction It's an intellectual property door founded on the premise that computers can search through records, discover styles, and make decisions with little human interaction.

Researchers using artificial intelligence information wish to examine if computers should study information based on sample recognition and the notion that computer programmes can test without being built to conduct good activities. When models are given fresh statistics, they may be able to adapt autonomously, which is a typical property of device knowledge. They study math in order to make consistent, repeatable judgments and outcomes. Not only is there new technological understanding now, but there is also fresh energy.

There are several perspectives on salary, including the use of despair and effective retreat, and the use of statistical sets to play statistics and enquiries. $\frac{3}{4}$ size of information is used to train the calculations, and $\frac{1}{4}$ size of information is designated as the Test set.

Highlighting, system efficiency and ingenuity have dramatically changed the world view of cost estimates with greater accuracy and forecasting. In addition, over the next few years significant improvements can be made to the use of these enhancements in anticipation of price payments. gadget learning has the following algorithms:

- Supervised learning
- Unsupervised learning

1.4.1 SUPERVISED LEARNING

This is a list of predictions. These predictions are unrelated variables. This learning algorithm's goal is to make predictions based on this set of independent variables. Variability in result predictability. This is a conditional variable. A function is formed from a group of independent variations that aids in the delivery of our intended output inputs. The machine is constantly trained in order to attain a particular level of accuracy in our training data. Linear regression, hindsight, KNN decision tree, random forest, and other guided readings are examples.

1.4.2 UNSUPERVISED LEARNING

No specific aim or outcome can be estimated or predicted with this approach. It is used to join many groups, which are then divided into various groups in order to interfere. K-Means are another example of unregulated learning.

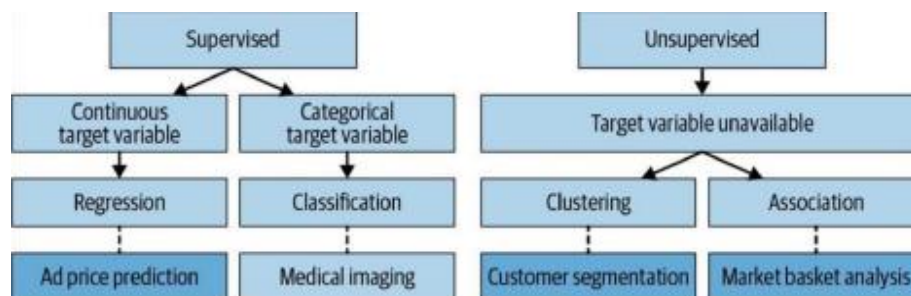


Fig 2 Supervised vs Unsupervised Learning

1.5 ORGANIZATION

Overall performance of controllable separators practical training is considered with limited memory accessible on web scraping . Accumulating training records and is often used directly in the steps of the section, reducing weight for customers. The report is composed of: section I describes the features used in determining the salary . Section II explains the proposed method and section IV defines test results and finally section V describes the conclusion.

CHAPTER 2

LITERATURE REVIEW

1. Lothe et al. propose a linear regression algorithm-based wage prediction system. The goal of the algorithm is to create a system that can anticipate salary based on various company parameters. The system's result is derived using an appropriate method and compared to other algorithms using standard scores and curves such as classification accuracy, F1 Score, ROC curve, and Precision-Recall curve. According to the statistical findings, linear regression with second order polynomial transformation produced the best predictions, with an MSE of 357 and a 76% accuracy.[1]
2. Dutta et al. design of a new Prediction Engine for Predicting Appropriate Salary for a Job. The technique is designed to predict salary for jobs where the compensation is not specified. Furthermore, the algorithm aims to assist freshers in estimating probable salaries for various organizations in various areas. A dataset given by ADZUNA serves as the study's foundation. The model is capable of accurately predicting values[2].
3. Khongchai and Songmuang presented a prediction model using Decision tree technique with seven features. The algorithm's output included not only a predicted wage, but also the three highest salaries of graduates with similar characteristics to the customers. They set up an experiment utilizing 13,541 records of actual graduated student data to test the system's efficiency. The overall accuracy score is 41.39 percent[3].
4. Ray delivered a brief assessment of several machine learning methods. The suggested technique is based on algorithms that are commonly used to tackle problems like classification, regression, and clustering. The benefits and drawbacks of various algorithms were explored, as well as a comparison of different algorithms in terms of performance, learning rate, and other factors. There have also been examples of practical applications of these algorithms presented[4].

5. Rhonda Magel and Michael Hoffman analyze the pay of Major League Baseball (MLB) players and if they are compensated based on their performance on the field. The researchers divided the experiment into two models based on the type of player, players and pitchers, and calculated using a regression model. Because the two models based on career production data had predictive r-squared values of at least 0.68, they were considered good predictive models. However, because their paper was limited to a single area, it could not be utilized to forecast compensation in other industries[5].

CHAPTER-3

SYSTEM DEVELOPMENT

3.1 FLOW CHART

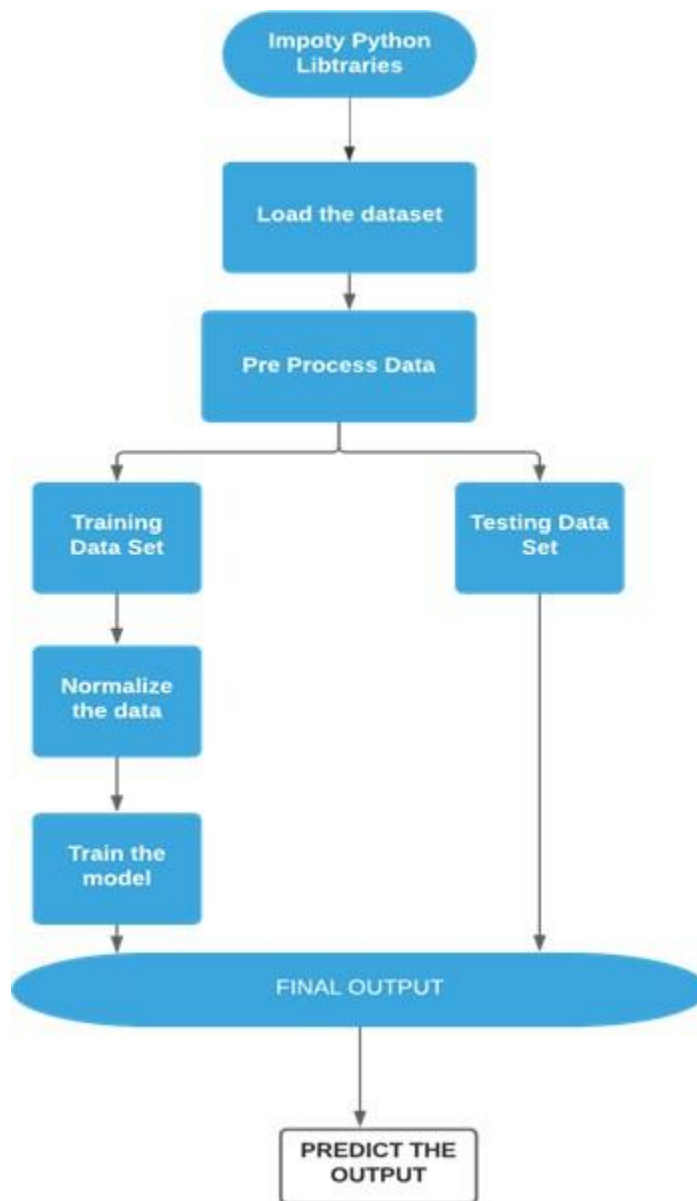


Fig 3 Flowchart of project

PYTHON

Python is a language-based dynamic semanticization system that is state-of-the-art. Its high level of integrated data structures, as well as dynamic typing and dynamic linking, are widely utilized in rapid application creation and as a scripting and pasting language to connect existing components. Python's straightforward, easy-to-understand syntax promotes readability while lowering programme maintenance costs.

Python encourages system flexibility and code reuse through modules and packages. Python interpreter and standard library that can be freely released in source or binary form across any big platform. Performance and consistency are two advantages of Python that make it ideal for AI and AI organizations. To perform the proper thing, there are links to specific archives, AI frameworks, and AI (ML) Private Fields.

HTML

HTML (HyperText Markup Language) is a markup language for displaying text in a web browser. Cascading Style Sheets (CSS) and programming languages like JavaScript can be beneficial. Web browsers display text on multimedia web pages by retrieving HTML documents from web servers or local archives. HTML describes the layout of a web page in terms of embedded numbers and symbols.

HTML elements are the components that make up HTML pages. Images and other items such as engagement forms can be integrated in a page using HTML standards. HTML allows you to produce well-organized texts by specifying semantics of text structure like titles, paragraphs, lists, links, and quotes, among other things. It creates organized texts by defining semantics of text structure such as titles, paragraphs, lists, links, quotes and more. HTML elements are defined by tags, written using angle brackets. Tags like ``, and `<input />` present the content directly on the page. Some tags like `<p>` provide information about document text and may include other tags as sub-elements. Browsers do not display HTML tags, but use them to translate page content.

CSS

CSS is a style sheet language that describes how text created with a markup language such as HTML is presented. CSS is a web-based language that works with HTML and JavaScript.

CSS is a collection of rules for presenting and displaying material, including layout, colours, and fonts. This section can improve content accessibility, give delivery items more flexibility and control, allow numerous web pages to share formatting by providing the required CSS in a separate.css file, which lowers content complexity and duplication, and save the.css file to improve page load performance between file sharing pages and their formatting.

FLASK

Flask is a lightweight Python web framework. Because they don't require any extra tools or libraries, they're categorized as microframeworks. It has no foundation for web page summaries, form verification, or any other category where third-party libraries offer equivalent services. However, Flask provides extensions, which allow you to add functionality to your app as though they were created in Flask. There are map extensions for objects, form verification, download management, and numerous open source projects, authentication systems, as well as a number of tools associated with the standard architecture.

When creating web applications, Flask provides the developer with a number of possibilities, tools, libraries, and other resources that allow you to create a web application without having to rely on them or describe how the project should be completed.

NUMPY

NumPy is a Python library that provides aid for massive, multi-dimensional collections and matrices, as well as a massive number of mathematical functions for interacting with them. Travis Oliphant invented NumPy in 2005 by integrating the competing capabilities of Numarray in Numbers, as well as a broad variety of other features. NumPy is free software with a large number of vendors. NumPy is a project supported by NumFOCUS.

NumPy in Python is similar to MATLAB in that both translate and allow customers to construct projects more quickly as long as many jobs are centered on clusters or networks rather than scales. There are various alternatives to these crucial ones.

SCIKIT-LEARN

The greatest Scikit-learn machine library for Python. Many useful machine learning techniques and mathematical modeling approaches, such as division, deceleration, integration, and size reduction, are available in the sklearn library. Sklearn uses machine learning models. Scikit-Learn is expensive and should not be used to read, manipulate, or summarize data. Some are there to assist you in translating the spread.

Scikit-learn comes with many features. Some of them are here to help us explain the spread:

- Supervised learning algorithms: Consider any professional reading algorithms you've examined that may be considered scientific. The science toolset includes anything from standard line models to SVM and decision trees. The expansion of machine learning algorithms is one of the key reasons for scientists' increasing usage. I started using scikit, and I would encourage young people to do so as well. I'll work on supervised learning issues.
- Unchecked learning algorithms: Compilation, feature analysis, key component analysis, and unsupervised neural networks are only some of the machine learning methods available.
- Contrary verification: a variety of methods are used by sklearn to ensure the accuracy of the models followed with discrete details.
- Feature removal: Scientific learning to remove images and text elements.
- Datasets for different toys: This was useful when studying science. I have studied SAS for different educational data sets. It helped them a lot to support when they read the new library.

PANDAS

Pandas is an open-source library that makes operating with relational or labelled statistics easy and intuitive. It includes a ramification of facts formats and strategies for working with numerical facts and time collection. The NumPy Python library gives the muse for this library. Pandas is quick and provides users with excellent performance and productivity. Pandas is a game-changer when it comes to analyzing data using Python, and it is one of the most popular and commonly used data munging/wrangling tools, if not THE most popular.

MATPLOTLIB

For 2D array charts, Matplotlib is an excellent Python visualization library. One of the most important benefits of visualization is that it allows us to see large amounts of data in easily understood images. Line, bar, scatter, histogram, and more graphs are available in Matplotlib. Matplotlib is free to use because it is open source. Matplotlib is primarily written in Python for platform portability, with a few pieces written in C, Objective-C, and Javascript.

3.3 TECHNICAL REQUIREMENT

HARDWARE REQUIREMENT

- Processor Intel(R) Celeron(R) J4005 CPU @ 2.00GHz 2.00 GHz
- Installed RAM 4.00 GB (3.85 GB usable)
- Device ID 5D876BDF-30FA-45D6-9F96-2D96BF4830B5
- System Type 64-bit working framework, x64-based processor

SOFTWARE REQUIREMENT

- Internet Browser: Microsoft Internet Explorer, Mozilla, Google Chrome or later
- Operating System: Windows XP / Windows7/ Windows Vista/ Windows 10
- PyCharm Version 2021.1
- Anaconda Version 2020.11
- ipython 7.19.0
- jupyterlab 2.2.6
- matplotlib 3.3.2
- notebook 6.1.4
- numpy 1.19.2
- pandas 1.1.3
- python 3.8.5
- scikit-learn 0.23.2
- scipy 1.5.2
- sqlalchemy 1.3.20
- statsmodels 0.12.0

3.4 DATA SET USED

While selecting or importing information sets, the type of method used, whether or not supervised, unsupervised, or semi-supervised, as well as the number of legitimate records and attributes, need to all be taken into consideration essential selection criteria. In this work, a Kaggle dataset with 50000 facts and 15 attributes is employed, in addition to the supervised technique. The range of statistics and attributes in this dataset have been chosen because they may be enormously sufficient to permit for the improvement of a green model and a huge range of options inside the prediction. The document layout is CSV. The dataset's houses are logical and competent, taking into consideration an accurate and green wage projection.

- The data is scraped from the website of Glass door using the scraper.py code.
- Job Description is broken into languages required for the job.[Python, R, Java, HTML, Excel,Spark, AWS]
- Salary is broken into hourly salary or employer provided.
- Company name simplified.
- Age of the company is simplified. If not available replace with the age calculated by DOB
- New CSV file created.
-

Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors	
0	0	Data Scientist	\$53K-\$91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense	\$50 to \$100 million (USD)	-1
1	1	Healthcare Data Scientist	\$63K-\$112K (Glassdoor est.)	What You Will Do\n\nInl. General Summary\n\nInThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care	\$2 to \$5 billion (USD)	-1
2	2	Data Scientist	\$80K-\$90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services	\$100 to \$500 million (USD)	-1
3	3	Data Scientist	\$56K-\$97K (Glassdoor est.)	*Organization and Job ID*\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	Government	Energy	Oil, Gas, Energy & Utilities	\$500 million to \$1 billion (USD)	Oak Ridge National Laboratory, National Renewa...
4	4	Data Scientist	\$86K-\$143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	Commerce Signals, Cardlytics, Yodlee

Fig 4 Raw Dataset

3.5 DATA PROCESSING

The data cleaning approach entails developing visualist and analytic methods that can handle multivariate, multidimensional data sets and are valuable in a variety of domains such as scientific visualization, business intelligence, machine learning and statistics. This method involves detecting the data, extracting it, cleaning it, and integrating it into a dataset that can be studied as needed. Because obtaining a perfect dataset that is noiseless is impossible, it is the programmers' obligation to remove as much noise, incompleteness, and other restrictions as possible to reduce errors. A decent enough model must exist.

It's a difficult task to improve the general accuracy of the ML model by removing all unneeded and useless features that don't contribute to the target variable. This study is based on data cleaning, which removes any null, missing, duplicate, and inconsequential values from the dataset. It not only cleans but also reduces the number of records and characteristics in the dataset. In ML, choosing the right feature is critical for improving the model's performance. The accuracy of the model can be improved by using the dataset attributes during training. Variables that are no longer necessary or are inconsequential must be eliminated. No data cleansing poses a risk of negatively affecting the model's forecast. There were attributes deleted from this model, and the remaining records were approximately which were efficient and sufficient enough to produce an accurate machine learning model capable of accurately predicting pay.

When data is collected and transformed into usable information, it is called data processing. Data processing is usually done by a data scientist or a team of data scientists, and it is critical that it is done correctly so that the final output, or data extraction, is not harmed.

Data processing begins with data in its immature form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context needed for computer translation and staffing throughout the organization.

FEATURES IN DATASET-

```
[ ] df.columns

Index(['Unnamed: 0', 'Job Title', 'Salary Estimate', 'Job Description',
      'Rating', 'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
      'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors',
      'hourly', 'employer_provided', 'min_salary', 'max_salary', 'avg_salary',
      'company_txt', 'job_state', 'same_state', 'age', 'python_yn', 'R_yn',
      'spark', 'aws', 'excel'],
      dtype='object')
```

Fig 5 Features in Dataset

3.6 DATA VISUALIZATION

In the previous 45 years, scientific data visualization has changed dramatically. Data visualization is the process of presenting information in the form of graphs and charts. It serves as a conduit for data and images. This is especially important as data visualization makes patterns and trends more visible. Regular analyses, such as predictive analytics, are aided by machine learning, which is a useful tool for displaying visualization. Data visualization is a field-independent technology that helps all tasks in some way. It is a useful tool for both demonstrating the relevance of huge data and assisting in its analysis.

Because the value of data is determined by its meaning, and visualization improves the meaning of data, this tool is essential. The ideal way to depict the age parameter is with a histogram, which represents the flow of the frequency, whereas parameters like educational level are best represented with a bar chart. We can explore three components at a time in a pretty simplistic method using interactive graphs using the Plotly package.

3.7 DATA CLEANING

- I. Only the numeric data from Salary was parsed.
- II. Separate columns were created for the employer's pay and hourly wages.
- III. Because the Salary column had a few empty entries, the rows lacking Salary were eliminated.
- IV. Made a separate column for the Company State by parsing the rating out of the company text.
- V. A new column was added to check if the job is at the company's headquarters. A new column was added to check the company's age by using the founding date.
- VI. Columns were added to check if the various talents were stated in the job description.
- VII. A new column for simplified job titles and seniority has been added.

3.8 MODEL FITTING

Without the useful information collected from the data, the data has no meaning. Predictive analytics is now concerned with data analysis in order to obtain useful information. All of this is only feasible thanks to techniques that allow the ML model to perform various jobs. This paper is driven using three machine learning algorithms i.e Multiple Linear Regression, Lasso Regression and Random forest.

MULTIPLE LINEAR REGRESSION

The general form of the equation for linear regression is:

$$y = A * x + B$$

where, y is the dependent variable, x is the independent variable, and A and B are the equation's coefficients. The distinction between linear regression and multiple regression is that multiple regression requires the ability to handle many inputs, whereas linear regression only requires one. To account for this modification, the multiple regression equation takes the following form:

$$y = A1 * x1 + A2 * x2 + \dots + An * xn + B$$

LASSO REGRESSION

Lasso regression is a sort of shrinkage-based linear regression. Data values are shrunk towards a central point, such as the mean, in shrinkage. Simple, sparse models are encouraged by the lasso approach (i.e. models with fewer parameters). This form of Residual Sum of Squares $+\lambda$ is unique (Sum of the absolute value of the magnitude of coefficients)

Where,

$$\sum_{i=1}^n (y_i - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

λ denotes the amount of shrinkage.

$\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model

$\lambda = \infty$ implies no feature is considered i.e, as λ closes to infinity it eliminates more and more features

The bias increases with increase in λ

Variance increases with decrease in λ regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

RANDOM FOREST

Random Forest is a well-known supervised learning technique-based machine learning algorithm. It can be used for both classification and regression problems in machine learning. It is based on ensemble learning, which is a way of combining multiple classifiers to solve a complicated problem and improve the performance of the model.

A random forest is made up of a large number of individual decision trees that work together as an ensemble, as the name suggests. The random forest creates a class prediction for each tree, and the class with the highest votes becomes our model's forecast.

3.9 VARIOUS STAGES OF THE PROJECT

Before training our model first of all we must have to balance our Data set.

```
import pandas as pd
from sklearn import metrics
import warnings
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
warnings.filterwarnings('ignore')
```

Fig 6 Imported Libraries

IMPORTING DATASETS

```
df.head()
```

salary , rating , name ,size, founded , tye of owner , sector , revenue , or maybe no . of cometitors on 7/5
#get this code ready tommorow and then get spark to input these

Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector
0	0	Data Scientist	\$53K-\$91K (Glassdoor est.)	3.8	Tecolote Research	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense
1	1	Healthcare Data Scientist	\$63K-\$112K (Glassdoor est.)	3.4	University of Maryland Medical System	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care
2	2	Data Scientist	\$80K-\$90K (Glassdoor est.)	4.8	KnowBe4	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services
3	3	Data Scientist	\$56K-\$97K (Glassdoor est.)	3.8	PNNL	Richland, WA	Richland, WA	1001 to 5000 employees	1965	Government	Energy	Oil, Gas, Energy & Utilities
4	4	Data Scientist	\$86K-\$143K (Glassdoor est.)	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services

Fig 7 Imported Dataset

We can see from the data description that there are no missing values. But if you check the dataset the missing values are replaced with invalid values like '?'. Let's replace such values with 'nan' and check for missing values again.

```
[4] df.shape
(956, 15)

[5] df.columns
Index(['Unnamed: 0', 'Job Title', 'Salary Estimate', 'Job Description',
      'Rating', 'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
      'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors'],
      dtype='object')

[6] df.isna().sum()
Unnamed: 0      0
Job Title      0
Salary Estimate 0
Job Description 0
Rating         0
Company Name   0
Location       0
Headquarters   0
Size          0
Founded        0
Type of ownership 0
Industry       0
Sector         0
Revenue        0
Competitors    0
dtype: int64
```

Fig 8 Calculating Nan Values

OBJECT VALUE TO NUMERIC

```
[10]

[11] df = df[df['Salary Estimate'] != '-1']
salary = df['Salary Estimate'].apply(lambda x: x.split('(')[0])
minus_Kd = salary.apply(lambda x: x.replace('K','').replace('$',''))

[12] min_hr = minus_Kd.apply(lambda x: x.lower().replace('per hour','').replace('employer provided salary:',''))

[13] df['min_salary'] = min_hr.apply(lambda x: int(x.split('-')[0]))
df['max_salary'] = min_hr.apply(lambda x: int(x.split('-')[1]))
df['avg_salary'] = (df.min_salary+df.max_salary)/2

[14] df["Company Name"].head(5)
0      Tecolote Research\n3.8
1  University of Maryland Medical System\n3.4
2      KnowBe4\n4.8
3      PNNL\n3.8
4  Affinity Solutions\n2.9
Name: Company Name, dtype: object
```

Fig 9 Converted object values to numeric values

CATEGORIZING ON THE BASIS OF SENIORITY LEVEL

```
def seniority(title):
    if 'sr' in title.lower() or 'senior' in title.lower() or 'sr.' in title.lower() or 'lead' in title.lower() or 'principal' in title.lower():
        return 'senior'
    elif 'jr' in title.lower() or 'jr.' in title.lower():
        return 'jr'
    else:
        return 'na'

df['job'] = df['Job Title'].apply(job_simple)
df['seniority'] = df['Job Title'].apply(seniority)

df.job.value_counts()

data scientist    279
na                184
data engineer     119
analyst           102
manager           22
mle               22
director          14
Name: job, dtype: int64
```

Fig 10 Categorized seniority level

FINDING COMPETITORS

```
Name: desc_len, Length: 742, dtype: int64

[37] df['num_comp'] = df['Competitors'].apply(lambda x: len(x.split(',')) if x != '-1' else 0)

[38] df['Competitors']

0          -1
1          -1
2          -1
3  Oak Ridge National Laboratory, National Renewa...
4  Commerce Signals, Cardlytics, Yodlee
...
737  Pfizer, AstraZeneca, Merck
738  See Tickets, TicketWeb, Vendini
739          -1
740          -1
741          -1
Name: Competitors, Length: 742, dtype: object

[39] df["num_comp"].head(5)

0    0
1    0
2    0
3    3
4    3
Name: num_comp, dtype: int64
```

Fig 11 Competitors of the company

DIFFERENT PLOTS

- Histograms for various fields that might contribute to the salary predictions.
- Correlation between these fields plotted using diverging palette
- Bar plots of different location and company textas and Headquarter

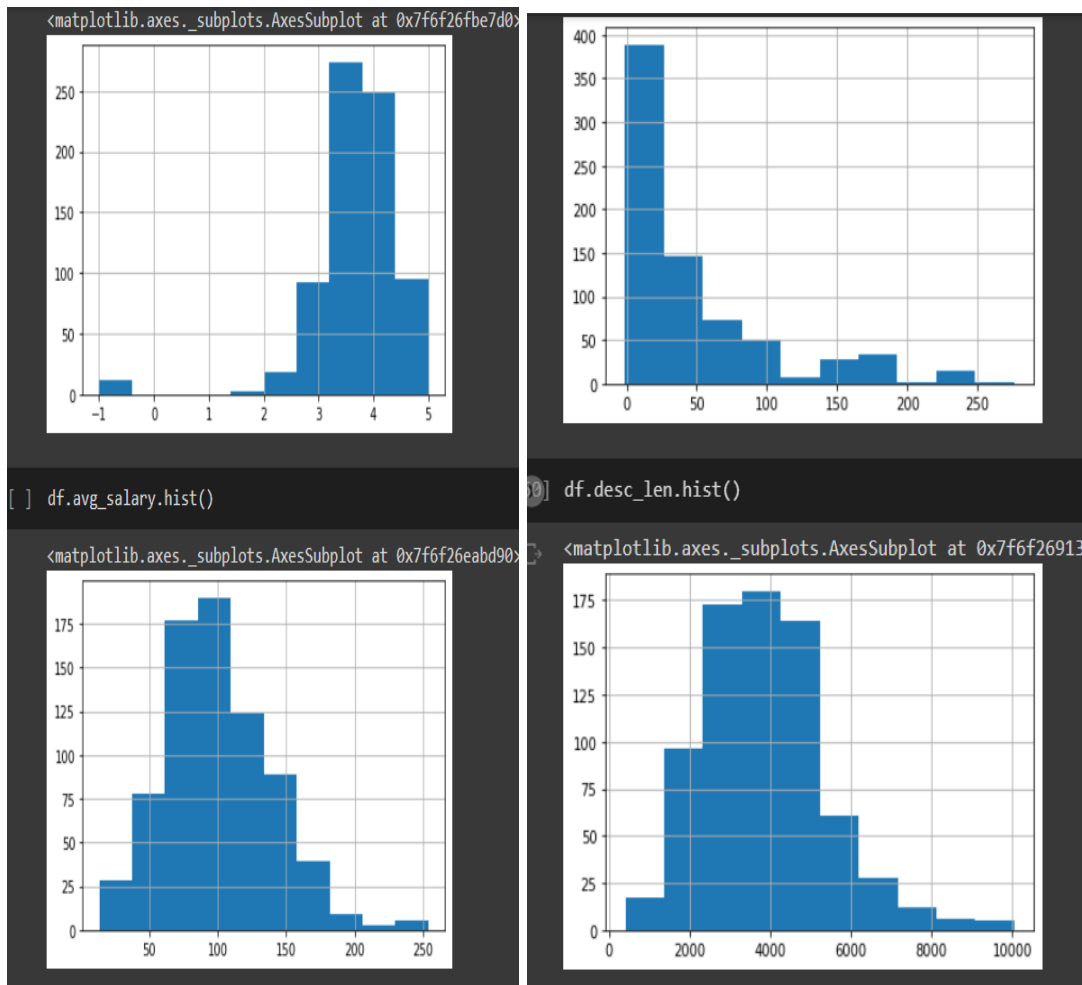


Fig 12 (a) Rating of company (b) Average salary (c) Company's Age (d) Length of job description

CORRELATION MAP

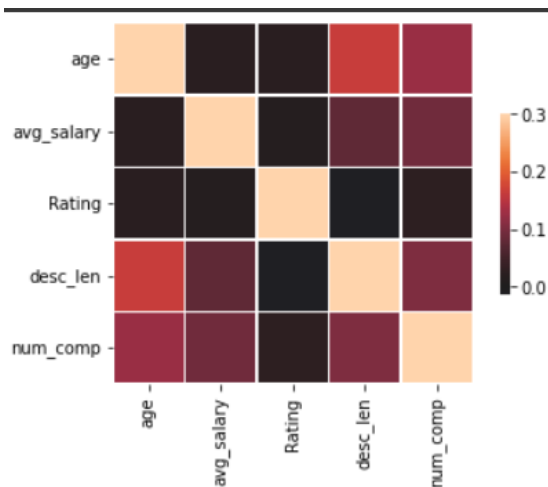


Fig 12 Correlation Map

We must convert the categorical data to numerical before applying any imputation algorithms. We could use get dummies, but since most of the columns only have two categories, we'll use mapping instead. Why? Because there are just two categories, the two columns created after get_dummies will have a high degree of correlation because they both explain the same thing. In any event, one of the columns will have to be removed. As a result, we'll employ mapping for these columns. We'll use get dummies for columns with more than two categories.

```
[78] df_model['Type of ownership'].unique()


array(['Company - Private', 'Other Organization', 'Government',
      'Company - Public', 'Hospital', 'Nonprofit Organization',
      'Subsidiary or Business Segment', 'Unknown',
      'School / School District'], dtype=object)

[79] df_model.drop(df_model[df_model['Type of ownership']=='Unknown'].index,inplace =True)

[80] def owner(a):
    b=0
    if a =='Company - Private':
        b=1
    elif a =='Company - Public':
        b=2
    elif a=='Government':
        b=3
    else:
        b=4
    return b
```

Fig 13 Type of ownership

Salary is a continuous variable for which regression analysis can be used. By avoiding overfitting and selecting the most relevant features using regularisation algorithms, it is possible to do embedded feature selection and maybe improve the models.



df_model.head()

	avg_salary	Rating	Size	Type of ownership	Sector	Revenue	employer_provided	same_state	age
0	72.0	3.8	750	1	5	75.0	0	0	47
1	87.5	3.4	7	4	5	3500.0	0	0	36
2	85.0	4.8	750	1	5	300.0	0	1	10
3	76.5	3.8	3000	3	2	750.0	0	1	55
5	95.0	3.4	350	2	4	1500.0	0	1	20

Fig 14 Cleaned data

CREATING PICKLE FILE

The Python pickle module is used to serialize and deserialize a Python object structure. In Python, pickling an object allows it to be stored to disc. Before writing to the file, Pickle "serialize" the object. Pickling is a Python function that turns a character stream from a list, dict, or other Python object. This character stream contains all of the information needed to rebuild the object in another Python procedure.

```
import pickle
filename = 'model.pkl'
pickle.dump(rf, open(filename, 'wb'))
```

Fig 15 Pickle File

INDEX.HTML

```
<div class="login">
<div>
</div>

<!-- Main Input For Receiving Query to our ML -->
<form action="{{ url_for('predict')}}"method="post">
  <input type="text" name="Rating" placeholder="Rating i.e. 1.0-5.0" required="required" />
  <input type="text" name="Size" placeholder="Number of employees i.e. 1-10000" required="required" />
  <input type="text" name="Type of ownership" placeholder="1.0 :private , 2.0: public , 3.0 : government ,others :4.0" required="required" />
  <input type="text" name="Sector" placeholder="Sector IT : 1 , Oil : 2 , Government : 3 , Realstate : 4 , others : 5" required="required" />
  <input type="text" name="Revenue" placeholder="Revenue in million dollars " required="required" />
  <input type="text" name="empprovided" placeholder="employer provided salary : 1 else 0" required="required" />
  <input type="text" name="state" placeholder="same state : 1 else 0)" required="required" />
  <input type="text" name="Age" placeholder="Company age : " required="required" />

  <button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>
</form>

<br>
<br>
{{ prediction_text }}
</div>
```

Fig 16 Code of Html

← → ↻ ⓘ File | C:/Users/jigya/OneDrive/Desktop/glassdoor_prediction/templates/index.html

SALARY PREDICTION

Rating i.e. 1.0-5.0	Number of employees i.e. 1	1.0 :private , 2.0: public , 3.0	Sector IT : 1 , Oil : 2 , Gove	Revenue in million dol
employer provided salary :	same state : 1 else 0)	Company age :	Predict	

{{ prediction_text }}

Fig 17 Output of Html

STYLE.CSS

```
body {  
  width: 100%;  
  height: 1000px;  
  font-family: 'Open Sans', sans-serif;  
  background: #092756;  
  color: #fff;  
  font-size: 18px;  
  text-align: center;  
  letter-spacing: 1.2px;  
  overflow-y: scroll !important;  
  overflow-x: hidden;  
}  
  
.login {  
  position: absolute;  
  top: 5%;  
  left: 38%;  
  
  width: 400px;  
  height: 800px;  
}  
  
.login h1 { color: #fff; text-shadow: 0 0 10px rgba(0,0,0,0.3); letter-spacing: 1px; text-align: center; }  
  
input {  
  width: 100%;  
  margin-bottom: 10px;  
  background: rgba(0,0,0,0.3);  
  border: none;  
  outline: none;  
  padding: 10px;  
  font-size: 13px;  
  color: #fff;  
  text-shadow: 1px 1px 1px rgba(0,0,0,0.3);  
  border: 1px solid rgba(0,0,0,0.3);  
  border-radius: 4px;  
  box-shadow: inset 0 -5px 45px rgba(100,100,100,0.2), 0 1px 1px rgba(255,255,255,0.2);  
  -webkit-transition: box-shadow .5s ease;  
  -moz-transition: box-shadow .5s ease;  
  -o-transition: box-shadow .5s ease;  
  -ms-transition: box-shadow .5s ease;  
  transition: box-shadow .5s ease;  
}  
  
input:focus { box-shadow: inset 0 -5px 45px rgba(100,100,100,0.4), 0 1px 1px rgba(255,255,255,0.2); }
```

Fig 18 Css Styling

APP.PY

```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle

app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))

@app.route('/')
def home():
    return render_template('index.html')

@app.route('/predict',methods=['POST'])
def predict():
    """
    For rendering results on HTML GUI
    """
    int_features = [float(x) for x in request.form.values()]
    final_features = np.array(int_features).reshape(1, -1)
    prediction = model.predict(final_features)

    output = round(prediction[0], 2)

    return render_template('index.html', prediction_text='Salary is {} K $ '.format(output))

@app.route('/predict_api',methods=['POST'])
def predict_api():
    """
    For direct API calls through request
    """
    data = request.get_json(force=True)
    prediction = model.predict([np.array(list(data.values()))])

    output = prediction[0]
    return jsonify(output)

if __name__ == "__main__":
    app.run(debug=True)
```

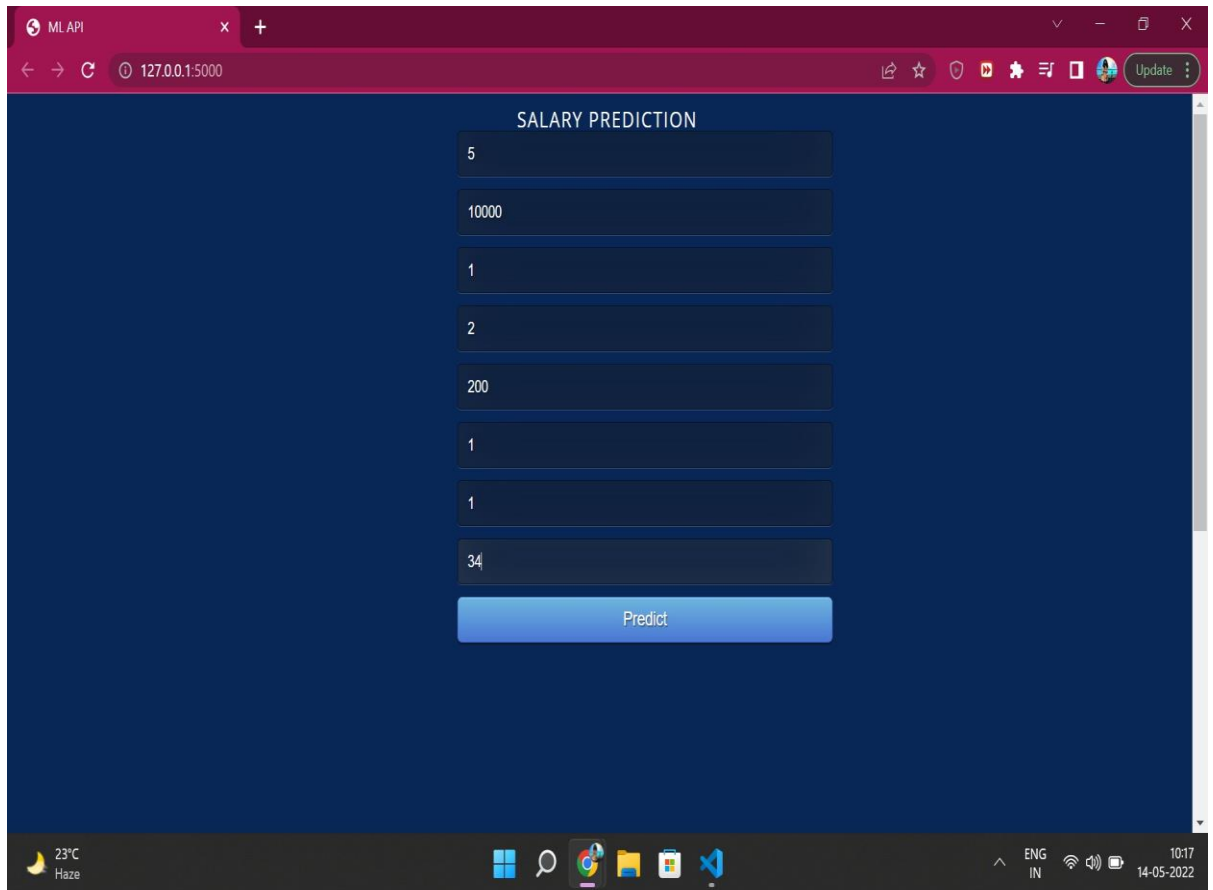
Fig 20 Flask Code

SERVER.PY

```
1  """
2  This code takes the JSON data while POST request and performs the prediction using loaded
3  the results in JSON format.
4  """
5
6
7  # Import libraries
8  import numpy as np
9  from flask import Flask, request, jsonify
10 import pickle
11
12 app = Flask(__name__)
13
14 # Load the model
15 model = pickle.load(open('model.pkl', 'rb'))
16
17 @app.route('/api/', methods=['POST'])
18 def predict():
19     # Get the data from the POST request.
20     data = request.get_json(force=True)
21
22     # Make prediction using model loaded from disk as per the data.
23     prediction = model.predict([[np.array(data['exp'])]])
24
25     # Take the first value of prediction
26     output = prediction[0]
27
28     return jsonify(output)
29
30 if __name__ == '__main__':
31     try:
32         app.run(port=5000, debug=True)
33     except:
34         print("Server is exited unexpectedly. Please contact server admin.")
```

Fig 21 Flask Code for server connection

OUTPUT:



The screenshot shows a web browser window with a single tab titled 'ML API'. The address bar displays '127.0.0.1:5000'. The webpage has a dark blue background and is titled 'SALARY PREDICTION'. It contains eight input fields with the following values: 5, 10000, 1, 2, 200, 1, 1, and 34. Below these fields is a light blue button labeled 'Predict'. The browser's status bar at the bottom shows the Windows taskbar with icons for the Start menu, Search, Edge, File Explorer, Task View, and Teams. The system tray on the right indicates the language is 'ENG IN', and the date and time are '10:17 14-05-2022'.

Input Field	Value
1	5
2	10000
3	1
4	2
5	200
6	1
7	1
8	34

Fig 22 Webpage for input data

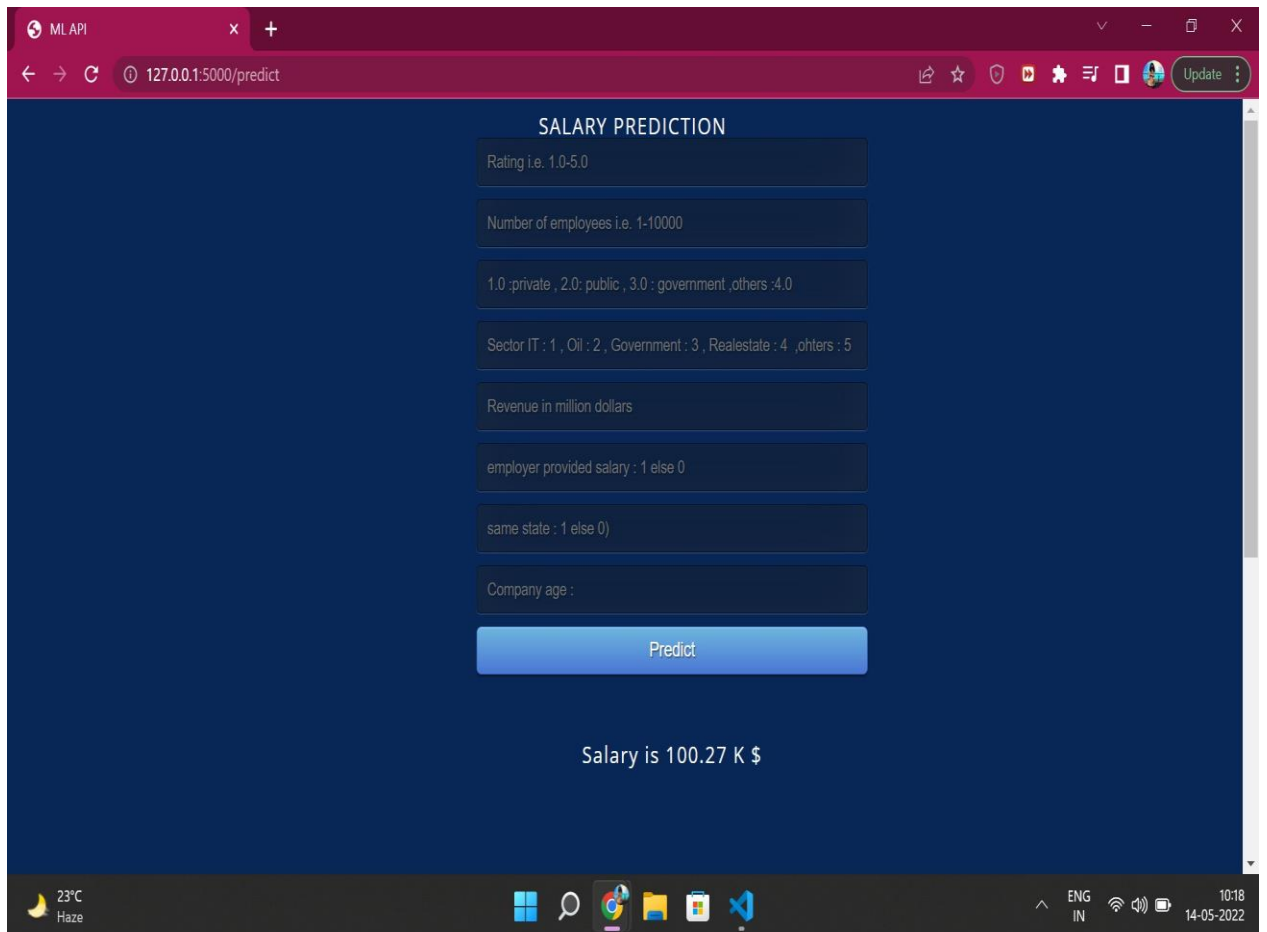


Fig 23 Webpage for output

CHAPTER-4

PERFORMANCE ANALYSIS

4.1 RESULTS CRITERIA

After all the above steps , now we will discuss the result produced by all the algorithms implemented in the project .The table below shows the result recorded from different algorithms used in the project.

MEAN ABSOLUTE ERROR

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}.$$

The MAE calculates the average size of forecast errors without considering the direction of the mistakes. It determines how precise continuous variables are. The equation is presented in the library references. In other words, over the verification sample, the MAE is the average of the absolute values of the differences between the forecast and the relevant observation. Because the MAE is a linear score, all individual differences are equally weighted in the average.

MEAN SQUARE ERROR

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

The mean squared error (MSE) of a regression line reflects how close it is to a set of points. By squaring the distances between the points and the regression line, it achieves this (these lengths are the "errors"). To eliminate any negative signals, squaring is required. Significant differences are also given more weight. It's called the mean squared error since you're calculating the average of a sequence of errors. The lower the MSE, the better the forecast.

ROOT MEAN SQUARE ERROR

$$\text{RMSE}_{fo} = \left[\sum_{i=1}^N (z_{fi} - z_{oi})^2 / N \right]^{1/2}$$

The RMSE is a quadratic scoring rule that calculates the average magnitude of an error. The RMSE equation is stated in both references. To put it another way, the difference between the forecast and the actual values is squared before being averaged throughout the sample. Finally, the square root of the average is determined. The RMSE gives enormous errors a lot of weight because the errors are squared before being averaged. The RMSE is thus most useful when large errors are really undesirable.

R SQUARE

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

R^2 is a coefficient of determination that determines how variations in one variable can explain differences in another. When a woman becomes pregnant, for example, the date she gives birth is directly related. The percentage of variation in y that can be explained by x-variables is calculated using R-squared. The range is 0 to 1. (i.e. 0 percent to 100 percent of the variation in y can be explained by the x-variables).

EXPLAINED VARIANCE SCORE

$$\text{explained variance}(y, \hat{y}) = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$$

$\text{Var}(y)$ represents the variance between prediction errors and actual values. Closer to 1.0 scores are preferred, as they indicate lower squared standard deviations of errors.

Explained variance is used to measure the difference between a model and real data (also known as explained variation). To put it another way, it's the proportion of total variation explained by actual components as opposed to error variance.

MEAN ABSOLUTE PERCENTAGE ERROR

The accuracy of a forecast system is determined by the mean absolute percentage error (MAPE). It is given as a percentage and is calculated as the average absolute percent inaccuracy for each time period minus actual values divided by real values.

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

where:

- where n is the number of fitted points
- A_t denotes the actual value
- F_t denotes the predicted value.
- Σ is the summarization notation (the absolute value is summed for every forecasted point in time).

4.2 RESULTS ACHIEVED

MULTIPLE LINEAR REGRESSION

```
Mean Absolute Error (MAE): 10.554747877609152
Mean Squared Error (MSE): 309.63434708417515
Root Mean Squared Error (RMSE): 17.596429952810745
Mean Absolute Percentage Error (MAPE): 0.12336445677786868
Explained Variance Score: 0.8013053352058359
R^2: 0.8012173670184801
```

Fig 24 Result of Multiple Linear Regression

LASSO REGRESSION

```
Mean Absolute Error (MAE): 10.554747877609152
Mean Squared Error (MSE): 309.63434708417515
Root Mean Squared Error (RMSE): 17.596429952810745
Mean Absolute Percentage Error (MAPE): 0.12336445677786868
Explained Variance Score: 0.8013053352058359
R^2: 0.8012173670184801
```

Fig 25 Result of Lasso Regression

RANDOM FOREST

```
Mean Absolute Error (MAE): 10.57032473230167  
Mean Squared Error (MSE): 306.3559366378329  
Root Mean Squared Error (RMSE): 17.50302649937527  
Mean Absolute Percentage Error (MAPE): 0.12145854308599474  
Explained Variance Score: 0.8033507855216122  
R^2: 0.8033220788072561
```

Fig 26 Result of Random Forest

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

The major goal of this project is to determine an applicant's appropriate future wage based on various domain-specific parameters. To train and perform predictions using the ML model, Multiple linear regression, Lasso regression and Random forest. The classification report is used as a comparison criteria to evaluate the overall efficiency of both algorithms once they have been imported. Random Forest outperforms the other THREE algorithms. The best prediction can then be picked. It can be improved by doing the following:

1. It can provide more advanced software for tallying salary mediums
2. It will host the platform on web servers
3. It can handle larger databases and curves than the sample above.

5.2 LIMITATIONS

1. The number of records is never adequate to accurately identify and estimate income, as the more data there is, the more accurate the ML Model becomes. The dataset's column count is presumed to be correct, however it is insufficient in comparison to the records.
2. Because the class distribution is either skewed or imbalanced, the binary classification used does not have the same number of samples from each class.
3. Based on personal details such as marital status, education level, and other factors, the dataset forecasted personal income levels to be above or below 50,000 per year. Although the skew is not significant, there could have been many more cases with incomes less than \$50,000 or greater than \$50,000

REFERENCES

1. D.M Lothe,Prakash Tiwari, Nikhil Patil, Sanjana Patil, Vishwajeet Patil, "SALARY PREDICTION USING MACHINE LEARNING" 2021 6 International Journal Of Advanced Scientific Research and Engineering Trends (IJASRET)
2. Sananda Dutta, Airiddha Halder, Kousik Dasgupta," Design of a novel Prediction Engine for predicting suitable salary for a job" 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN).
3. Pornthep Khongchai, Pokpong Songmuang, "Improving Students' Motivation to Study using Salary Prediction System" 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)
4. Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (ComIT-Con), India, 14th -16th Feb 2019
5. Magel, Rhonda, and Michael Hoffman. "Predicting salaries of major league baseball players." *International Journal of Sports Science* 5, no. 2 (2015): 51-58.