JOB THREAT INDEX USING MACHINE LEARNING

A Summer Internship Project Report Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY IN
CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Submitted by

B Navaneeth	21071A6611
G Ruchitha	21071A6620
G Aravind	21071A6623
K Yashwanth	21071A6626

Under the guidance of

Dr. B. Venkatesh
(Sr. Assistant Professor, Department of CSE-AIML & IoT, VNR VJIET)

DEPARTMENT OF CSE(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi, Affiliated to JNTUH Recognized as “College with Potential for Excellence” by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi, Affiliated to JNTUH Recognized as "College with Potential for Excellence" by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

TAMASŌMA JYOTIRGAMAYA

CERTIFICATE

This is to certify that the project report entitled "**JOB THREAT INDEX USING MACHINE LEARNING**" is a bonafide work done under our supervision and is being submitted by **B Navaneeth (21071A6611),G Ruchitha (21071A6620) , G Aravind (21071A6623) , K Yashwanth (21071A6626)** in partial fulfillment for the award of the degree of Bachelor of Technology in CSE(Artificial Intelligence and Machine Learning), of the VNRVJIET, Hyderabad during the academic year 2023-2024. Certified further that to the best of our knowledge the work presented in this thesis has not been submitted to any other University or Institute for the award of any Degree or Diploma.

Dr. B. Venkatesh **Sr. Assistant Professor** **Dept. of CSE (AIML & IoT)** **VNRVJIET**	**Dr. N. Sandhya** **Professor& HOD** **Dept. of CSE (AIML&IoT)** **VNR VJIET**

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi, Affiliated to JNTUH Recognized as "College with Potential for Excellence" by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

DECLARATION

We declare that the major project work entitled "**JOB THREAT INDEX USING MACHINE LEARNING**" submitted in the department of CSE-Artificial Intelligence and Machine Learning, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **CSE-Artificial Intelligence and Machine Learning** is a bonafide record of our own work carried out under the supervision of **Dr. B. Venkatesh , Sr.Assistant Professor, Department of CSE(AIML & IoT), VNRVJIET**. Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.

Place: Hyderabad

B Navaneeth	**G Ruchitha**	**G Aravind**	**K Yashwanth**
(21071A6611)	(21071A6620)	(21071A6623)	(21071A6626)

ACKNOWLEDGEMENT

Firstly, we would like to express our immense gratitude towards our institution VNR Vignana Jyothi Institute of Engineering and Technology, which created a great platform to attain profound technical skills in the field of Computer Science, thereby fulfilling our most cherished goal.

We are very much thankful to our Principal, **Dr. Challa Dhanunjaya Naidu,** and our Head of Department, **Dr. N. Sandhya**, for extending their cooperation in doing this project within the stipulated time.

We extend our heartfelt thanks to our guide, **Dr. B Venkatesh,** and the project coordinators **Miss. J. Pushpa Kumari** and **Miss. Preety Singh** for their enthusiastic guidance throughout the course of our project.

Last but not least, our appreciable obligation also goes to all the staff members of the CSE (AIML & IoT) and to our fellow classmates who directly or indirectly helped us.

B Navaneeth	(21071A6611)
G Ruchitha	(21071A6620)
G Aravind	(21071A6623)
K Yashwanth	(21071A6626)

ABSTRACT

"This work offers an insight into research investigating the utilization of machine learning algorithms, as outlined in the study titled 'Anticipating Job Disruption: An AI-Enhanced Model for Predicting Automation-Induced Threat Levels”. In this work, we are predicting how much percentage of a job can be replaced by AI

The rapid advancement of automation technologies has raised concerns about the potential impact on employment across various sectors. In response to this, our project, the "Job Threat Index," focuses on predicting the degree of job threat in different fields attributable to automation. Leveraging machine learning, specifically the K-Nearest Neighbors (KNN) algorithm, we aim to provide a quantitative measure of the vulnerability of specific occupations to automation-induced displacement.

Through feature engineering and data preprocessing, we extract relevant parameters to input into the KNN algorithm. KNN, a supervised learning algorithm, is chosen for its ability to classify data points based on the similarity of their features to those in the training dataset. In our context, this allows us to predict the level of automation threat faced by different jobs.

Preliminary results indicate the effectiveness of the KNN algorithm in predicting job threat levels, with a promising level of accuracy. Ongoing refinement of the model and continuous integration of updated datasets will enhance the accuracy and reliability of our predictions.

INDEX

1. INTRODUCTION

- The "Job Threat Index ML Project" is a data-driven initiative aimed at providing valuable insights into the ever-evolving landscape of employment and workforce dynamics.
- This model aims to provide concise and accurate predictions, helping businesses, policymakers, and individuals navigate the evolving employment landscape effectively.
- By analyzing various factors, it offers valuable insights into which job roles and industries are most vulnerable to automation, aiding in proactive workforce planning and reskilling efforts.
- In an era of rapid technological change, this model serves as a vital tool for informed decision-making and long-term workforce sustainability.

1.1 Problem statement:

In the world of predicting job threats, many attempts have been made using fancy computer models. However, these models often struggle because they don't consider all the important stuff about jobs. We need to think about more than just machines taking over jobs. We need to look at things like how adaptable people's skills are, how strong industries are, and how ready society is for changes.

The big issue with the existing models is that they don't look at the full picture. They focus too much on specific things and can't see the whole story. The challenge is to make a better model—one that looks at many different aspects of jobs, like what the jobs are like, what happened in the past, and what skills are needed..

Our goal is to make a strong model that can give us a number called the Job Threat Index. This number will help us see how likely it is that certain jobs might be at risk because of machines taking over.

2. LITERATURE SURVEY

FEASIBILITY STUDY

The feasibility study is both technically and economically viable. The organization's capabilities, coupled with positive cost-benefit indicators and user-centric strategies, position the project for successful development and deployment.

ORGANIZATIONAL FEASIBILITY

Expertise and Resources:

The organization possesses the requisite expertise in machine learning, cybersecurity, and data science. A skilled workforce is available, and collaboration with external experts is feasible if necessary. The organizational culture is conducive to embracing innovation in the cybersecurity domain.

ECONOMIC FEASIBILITY

Cost-Benefit Analysis:

A thorough cost-benefit analysis indicates that the benefits, including reduced security incidents and associated costs, outweigh the development, data acquisition, and deployment expenses. The estimated return on investment (ROI) over time is positive, aligning well with the organization's budget constraints.

TECHNICAL FEASIBILITY

Technology Infrastructure:

The existing technology infrastructure is robust enough to support the implementation and maintenance of machine learning algorithms for phishing detection. The necessary data for training and testing are available, and scalability considerations have been addressed to handle increasing web traffic and evolving phishing techniques. Integration with current security systems is seamless.

BEHAVIORAL FEASIBILITY

User Training and Acceptance:

User training programs are planned to educate users on recognizing and reporting potential phishing threats. The phishing detection system aligns with current user practices and workflows, ensuring high user acceptance. A feedback mechanism is established for users to report issues, fostering a collaborative approach to system improvement. Change management strategies are in place to facilitate a smooth transition.

Literature Review

S.NO	Title of the Paper	Models	Pros	cons	Year
1	The Automation of Jobs: A Threat for Employment or a Source of New Entrepreneurial Opportunities?	Data: 1)German Socio-Economic Panel (SOEP) data Models used 1)Probit regression model is used to analyse transitions from paid employment to unemployment and self-employment. It is a binary regression model, appropriate when the dependent variable is binary (0 or 1). 2)fractional response models (FRM) :The FRM is used to estimate the probability of an occupational change occurring within the next two years	Pros: 1)Descriptive results 2)Multivariate Analysis 3)Data Source: The use of the German Socio-Economic Panel (SOEP) data, a large and representative household survey, enhances the credibility and generalizability of the study's findings.	lCons: 1) Probit regression model Interpretation of coefficients is based on the cumulative normal distribution, which may be less intuitive than linear regression coefficients. 2)Fractional response models (FRM) Assumes a specific distribution for the error term, which may not always be appropriate for all datasets. 3)While the study examines the impact of automation on job transitions, it does not explore the broader socioeconomic implications or consequences of automation, which could be relevant for policymakers and society.	May 2015
2	Are Robots/AI Viewed as More of a Workforce Threat in Unequal	1. ***Data Source***: The study utilised individual-level data from the Eurobarometer 87.1 survey conducted in 2017.	Pros: 1)**Multilevel Modelling:** Using multilevel models is appropriate for analysing data where observations are nested within higher-level units	**Cons:** 1)Limited Information on Model Fit 2)Control Variable Complexity 3)Interpretability of Coefficients:	7 July 2021

	Societies?	Eurobarometer is a series of public opinion surveys conducted by the European Commission, and this particular survey focused on topics related to the impact of digitalization and automation on daily life. 2. **Sample Selection**: The sample was limited to employed participants, resulting in a final sample size of 13,294 employed individuals. The decision to focus on employed participants was likely due to the study's specific interest in the effects of advanced technology on the workforce. 3)Regression and Multi level model	2)**Full-maximum likelihood estimation,** ideal for extensive datasets and multiple Level 2 groups like countries, provides more precise model parameter estimates than other methods. 3) **The Intraclass Correlation Coefficient** (ICC) quantifies the share of outcome variance attributed to country-level factors, pivotal for assessing significant cross-country differences as hypothesised in the study.	4)Data Source Limitations	
3	Robots Worldwide: The Impact of Automation on Employment and Trade	The analysis seems to rely on econometric models and statistical techniques, rather than ML. The section discusses the regression setting and econometric issues in the analysis of the impact of robots on employment. Here are the key methods and approaches used in the study: 1)Cobb-Douglas Production Function 2)Log-Linearization 3)Inclusion of Robot Stock 4)Dummy Variable for Labour Intensity 5)Cross-Country Trends 6)Instrumental Variable (TP Index) 7)Stylized Analytical	**Pros:** **1)Causality Inference:** Instrumental variable (IV) analysis is valuable for establishing causal relationships in observational data by mitigating problems related to endogeneity and omitted variables, enhancing the validity of causal inferences. **2)Endogeneity Mitigation:** IVs are chosen to be exogenous and unrelated to the outcome variable, reducing endogeneity by isolating the impact of the independent variable (e.g., robot usage) from potentially confounding omitted variables, improving causal	**Cons:** **1)Assumption Dependence:** IV analysis relies on several assumptions, including the relevance and exogeneity of the instrument. **2)Instrument Selection:** Choosing a valid instrument can be challenging. **3)Limited Generalizability:** IV estimates may only apply to specific contexts where the instrumental variable is relevant.	2010

		Framework 8)Plausibility Checks	inference. **3)Improved Validity:** When the assumptions of IV analysis are met, the estimates of causal effects are more valid and less biased compared to traditional regression methods. **4)Policy Implications:** IV analysis is commonly used in economics and policy research to assess the impact of interventions or policies when randomized controlled trials are not feasible or ethical. It allows researchers to estimate causal effects in real-world settings.	**4)Precision and Sample Size:** IV estimates can be less precise and may require larger sample sizes than traditional regression analysis. **5)Complexity:** IV analysis is more complex than ordinary least squares (OLS) regression	
4	Jobs at Risk of Automation in the USA: Implications for Community College	**1)Data Collection:** The study collects occupation data from the 2019 Current Population Survey, which is a well-established source of labor statistics in the USA. **2)Risk Estimation:** The risk of automation estimates for different job categories are adopted from Frey and Osborne (2017). These estimates are based on their research, which may have involved machine learning and data analysis techniques to assess the automation potential of various occupations. **3) Data Matching and Calculation:** The study matches the risk of automation data with the	**Pros:** **1)Data Reliability:** The study uses data from the 2019 Current Population Survey, which is a widely recognized and reliable source for labour statistics in the USA. **2)Informative Insights**: The study provides valuable insights into the risk of job automation, which is a significant concern in the context of advancing technologies. It highlights the potential impact on various industries and age groups. **3)Policy Relevance:** The research offers important information for policy and decision-	**Cons:** **1)Lack of Detailed Information** **2)Limited Scope:** The study's methods focus on statistical analysis and reporting, rather than delving into the underlying machine learning algorithms or data processing techniques. **3)Data Lag:** The use of older data may not fully capture the most recent developments in automation technology and its impact on the job market.	January 25, 2021

		Current Population Survey data and calculates the numbers and proportions of jobs at risk for different age groups and industries. These calculations likely involve standard statistical methods. **4) Descriptive Statistics:** The study reports its findings using descriptive statistics such as percentages and numbers to summarise the results. **5) Comparative Analysis:** The study compares the risks of automation across different industries and age groups, which would typically involve statistical comparisons and possibly regression analysis.	makers, particularly in the field of education and workforce development, by emphasising the role of community colleges in addressing automation-related challenges. **4)Transparency**: The study outlines its data sources, methods, and findings, which enhances transparency and allows for potential replication or further research.	**4)Assumption of Static Risk:** The study assumes that the risk of automation estimated by Frey and Osborne remains constant over the next few decades. **5)Lack of Methodological Details**	
5	Predicting Challenge and Threat Appraisal of Job Demands among Nurses: The Role of Matching Job Resources	**1)Data Collection:**The study was conducted as part of a larger research project on nurses' working conditions in Luxembourg. Data collection was done through an online survey in the year 2021. Data collection occurred during the COVID-19 pandemic. **2) Participants :** Inclusion criteria for participants required them to work a minimum of 20 hours per week and have at least 6 months of experience in a nursing profession.	**Pros :** **1)Clarity in Objectives:** The limitations section provides clear insights into the scope and objectives of the study, which helps readers understand the context in which the research was conducted. **2)Transparency:** The authors openly acknowledge the limitations of their study, demonstrating a commitment to transparency and academic rigor.	**Cons:** **1)Cross-Sectional Design:** While the authors defend the use of a cross-sectional design for their study, it remains a limitation in terms of drawing causal conclusions. Longitudinal or experimental designs could have provided more robust insights into causality. **2)Sampling Bias:** The use of	January 11, 2023

		3) Measures: The study used established and validated scales to measure the relevant variables	**3)Suggestion for Future Research:** The authors offer valuable suggestions for future research, indicating areas where further investigation could enhance the understanding of the subject matter.	convenience sampling is acknowledged as a limitation. The lack of representativeness in the sample limits the generalizability of the findings. **3)Potential Impact of COVID-19:** The acknowledgment of the impact of the COVID-19 pandemic on nurses' job stress is valid, but it could also be seen as a limitation, as it may have confounded the results. This factor could have been controlled for or investigated in more detail.	
6	Job Insecurity, Employability and Financial Threat during COVID-19	Data Collection: Survey through Amazon's Mechanical turk Methods: Correlation among the attributes to find the effect of one attribute on the other attribute	**Pros**: Attribute Importance: The considered attributes are very essential with respect to the desired output. Path Analysis: The path analysis done in the paper from the obtained calculations are easy to correlate among the attributes used by the researcher.	**Cons**: The dataset is not big enough to conclude various conclusions. The dataset is obtained during the lockdown where the psychological conditions of the people is not in a good state which effects the quality of the dataset.	March 2023

7	The Potential for Artificial Intelligence in healthcare	Data: Dataset is not present but they gave the future prospects of AI in healthcare Models: Actual ML or AI models are not implemented but in which domain of the healthcare industry AI is infused is explained in detail	**Pros:** Gives insights on how AI will play an essential part in the healthcare industry.	**Cons:** It would have been implemented on a sample dataset to show which domain is going to adapt the most amount of automation.	June 2019

3.EXISTING SYSTEM

1. McKinsey Global Institute's Job Displacement Impact Models:

   McKinsey has developed models to assess the potential impact of automation on different occupations

2. World Economic Forum (WEF) Future of Jobs Report:

   The WEF produces reports that analyze the impact of technological changes on employment, including predictions about job displacement and emerging job roles.

3. Burning Glass Technologies:

   Companies like Burning Glass use big data analytics to provide insights into labor market trends, including emerging job roles and potential job threats.

4. AI-Based Predictive Analytics Platforms:

   Companies and platforms focused on predictive analytics, using machine learning algorithms to analyze trends and forecast job market changes.

4. SYSTEM REQUIREMENTS

Functional Requirements

- Data Collection and Integration: various sources, including job market data, industry reports, and automation statistics.
- Data Preprocessing: data cleaning, normalization, and feature extraction, to prepare the data for analysis.
- Feature Selection: selecting most relevant features or variables that contribute to job threat assessment.
- Machine Learning Algorithms: (e.g., regression, classification) to analyze the data and make predictions.
- Model Training: On historical data to learn patterns and relationships between job market conditions, automation trends, and job threat levels.
- Real-time Data Updates: real-time or periodic updates of data to ensure that the model remains current and relevant.
- Prediction: Providing the capability to predict job threat indexes for different sectors, regions, or timeframes based on user input or predefined scenarios.
- User Interface: Developing a user-friendly interface that allows users to input parameters, select criteria, and view job threat predictions.
- Update Mechanism: periodically updating the data as per new & updated information

Non Functional Requirements

- Scalability: The system should be able to handle an increasing amount of data and users without a significant decrease in performance.
- Performance: The machine learning model must provide results within an acceptable response time, even as the dataset grows. For example, the system should be able to generate threat index scores quickly.
- Reliability: The system should be highly reliable, with minimal downtime. It should also have mechanisms for disaster recovery.
- Security: Protect sensitive personal and job-related data, ensuring that access is restricted only to

authorized personnel. This includes data encryption and compliance with relevant data protection regulations.

- Privacy: Ensure that the system complies with privacy regulations and best practices for handling sensitive information. Minimize the risk of data breaches.
- Usability: The user interface should be intuitive and easy to use, with a low learning curve. Users should be able to interact with the system without extensive training.
- Compatibility
- Maintainability
- Data Quality: Non-functional requirements should include data quality standards to ensure that the data used to train the model is accurate, reliable, and up to date.
- Compliance: Ensure that the project complies with legal and regulatory requirements relevant to job threat assessment and data handling.
- Resource Utilization: Optimize resource usage, such as memory and CPU, to maximize efficiency and reduce operational costs.

5. SOFTWARE DESIGN

4.1 UML DIAGRAMS

The Device Architecture Manual describes the application requirements, operating state, application and subsystem functionality, documents and repository setup, input locations, yield types, human-machine interfaces, management reasoning, and external interfaces. The Unified Modeling Language (UML) assists software developers in expressing an analysis model through documents that contain a plethora of syntactic and semantic instructions. A UML context is defined as five distinct viewpoints that present the system from a particularly different point of view.

The components are similar to modules that can be combined in a variety of ways to create a complete UML diagram. As a result, comprehension of the various diagrams is essential for utilizing the knowledge in real-world systems. The best method to understand any complex system is to draw diagrams or images of it. These designs have a bigger influence on our understanding. Looking around, we can see that info-graphics are not a new concept, but they are frequently utilized in a variety of businesses in various ways.

User Model View
The perspective refers to the system from the clients' point of view. The exam's depiction depicts a situation of utilization from the perspective of end-clients. The user view provides a window into the system from the perspective of the user, with the system's operation defined in light of the user and what the user wants from it.

Structural model view
This layout represents the details and functionality of the device. This software design maps out the static structures. This view includes activity diagrams, sequence diagrams and state machine diagrams

Behavioral Model View
It refers to the social dynamics as framework components, delineating the assortment cooperation between various auxiliary components depicted in the client model and basic model view. UML Behavioral Diagrams illustrate time-dependent aspects of a system and communicate the system's dynamics and how they interact. Behavioral diagrams include interaction diagrams, use case diagrams, activity diagrams and state–chart diagrams.

Implementation Model View
The essential and actions as frame pieces are discussed in this when they are to be

manufactured. This is also referred to as the implementation view. It uses the UML Component diagram to describe system components. One of the UML diagrams used to illustrate the development view is the Package diagram.

Environmental Model View

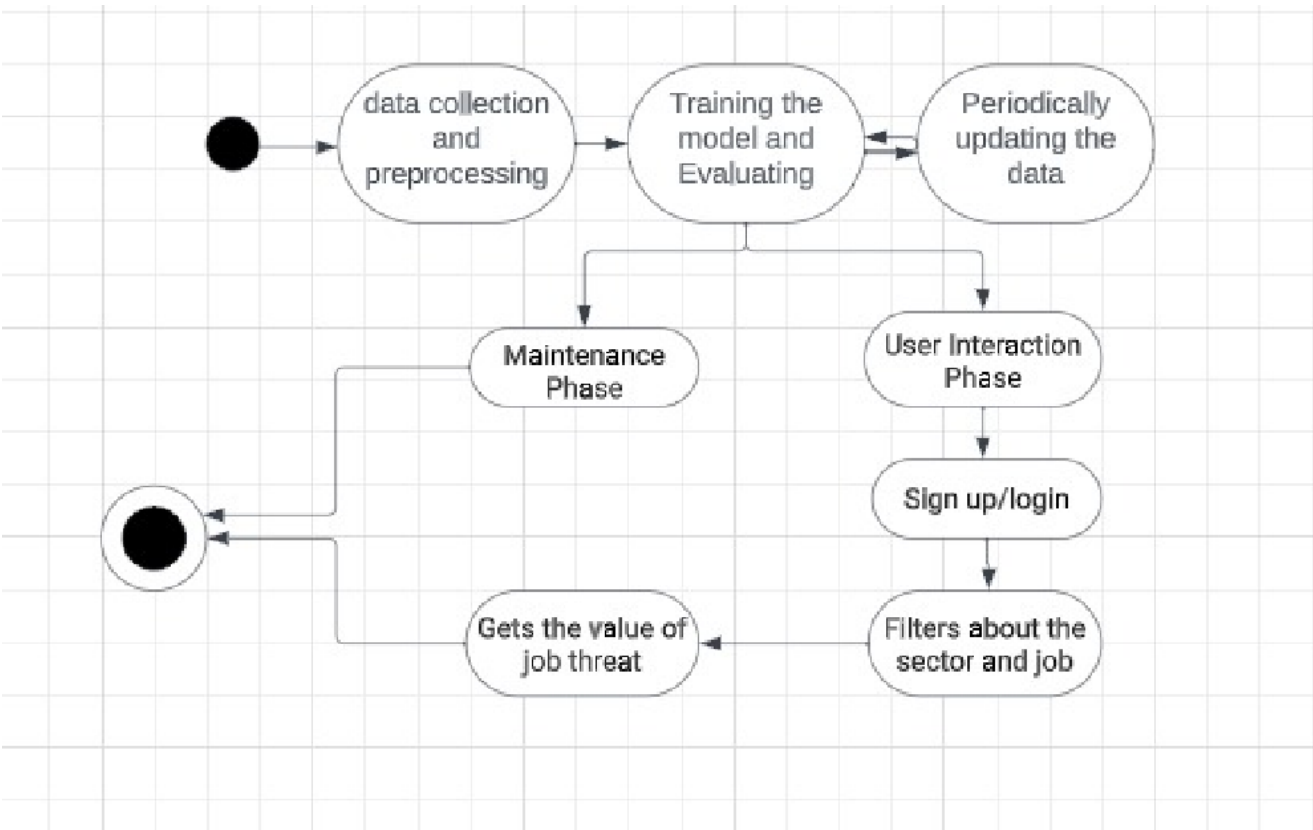
The systemic and functional component of the world where the program is to be introduced was expressed within this. The diagram in the environmental view explains the software model's after-deployment behavior. This diagram typically explains user interactions and the effects of software on the system. The following diagrams are included in the environmental model: Diagram of deployment.

The UML model is made up of two separate domains:

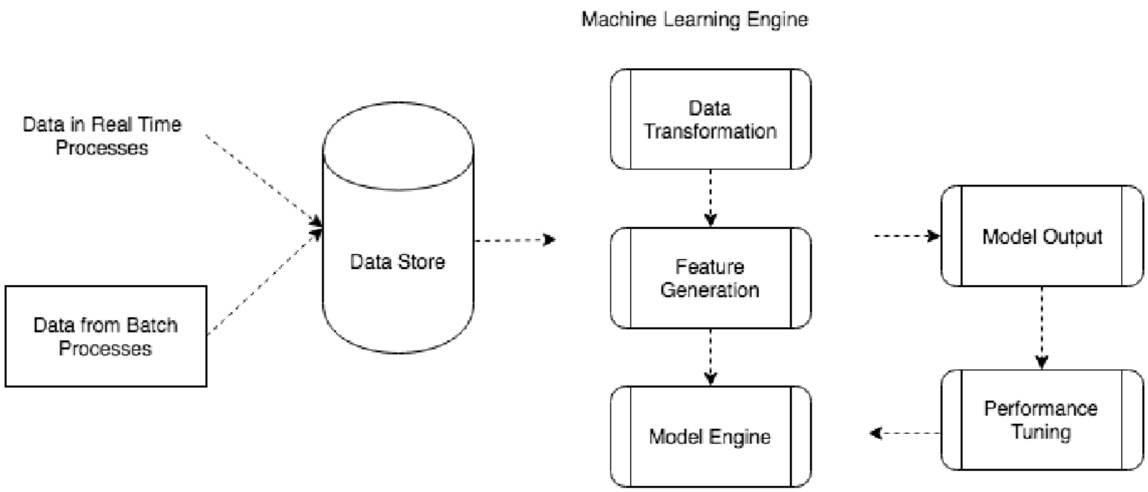
- Demonstration of UML Analysis, with a focus on the client model and auxiliary model perspectives on the framework.
- UML configuration presenting, which focuses on demonstrations, usage, and natural model perspectives.

USE CASE DIAGRAM

A use case diagram is a kind of behavioral diagram that is used in the Unified Modeling Language (UML). This type of diagram is defined by and developed from use case research. Its purpose is to provide a graphical representation of a system's functionality in terms of its actors, the goals of the actors that they want to achieve (which are stated as use cases), and any relationships that exist between those use cases. The primary objective of a use case diagram is to specify which system functions are carried out for particular actor.



System Architechture



6. PROPOSED SYSTEM

We gathered historical data from various sources , such as Employment rates, job vacancies, industry trends, and wage data, GDP growth, inflation rates, unemployment rates, regional economic data. This dataset will be used for training and evaluating the model. The data is Cleaned and processed to remove outliers, missing values and inconsistencies. Relevant features, such as historical Job replaced data and usable variables(Potential AI innovation areas) are extracted from the raw data. Split the data into training and testing sets for model evaluation. With the extracted feature data we train a Random forest, KNN, Decision tree model. After training the model use it to make job threat index predictions on testing data and calculate the accuracy and forecast the errors by comparing the predicted Job threat index to the actual observations. Based on the error analysis, refine the model to reduce Job threat index errors. Consider using ensemble methods, such as bagging or boosting, for the Random Forest algorithm to improve Job threat index accuracy and error estimation.

MODULES

- **Data collection**: This module is responsible for collecting historical data on IPL matches, including match location, team lineup, player stats, and match outcomes. This data can be obtained from various sources such as IPL websites, APIs, and scraping tools.
- **Data preprocessing**: This module is responsible for cleaning and preparing the collected data for machine learning. This might involve removing irrelevant or incomplete data, transforming the data into a consistent format, and normalizing the data.
- **Feature engineering**: This module is responsible for creating new features from the existing data that might be more informative for the machine learning model. For example, features could be created to represent team strength, player form, and pitch conditions.
- **Feature selection**: This module is responsible for selecting the most relevant features to use for model training. This can be done using domain knowledge or feature selection techniques such as principal component analysis (PCA) or recursive feature elimination (RFE).
- **Model selection**: This module is responsible for selecting the most appropriate machine learning model for the task. There are many different machine learning algorithms available, such as decision trees, random forests, support vector machines (SVMs), and artificial neural networks (ANNs). The choice of model depends on the type of data and the problem at hand.
- **Model training**: This module is responsible for training the selected machine learning model on the preprocessed data. The training process involves splitting the data into training and validation sets, fitting the model on the training set, and evaluating the model on the validation set.
- **Model evaluation**: This module is responsible for evaluating the trained machine learning model on held-out test set to assess its performance on unseen data. This helps in determining how well the model is likely to generalize to new data.
- **Model deployment**: This module is responsible for deploying the trained and evaluated machine learning model to production so that it can be used to make predictions on new data. This might involve saving the model to a file, deploying it to a cloud platform, or embedding it in a software application.
- **Visualization:** This module is responsible for creating data visualizations to help you understand the data and evaluate the performance of the machine learning model

7. CODING AND IMPLEMENTATION

Imports and Loading Dataset:

Reading Data Set

```
In [3]: import pandas as pd
```

```
In [4]: df2=pd.read_csv('Dataset1.csv')
```

```
In [5]: df3=df2
```

```
In [6]: df2.head()
```

Out[6]:

	Job titiles	AI Impact	Tasks	AI models	AI_Workload_Ratio	Domain
0	Communications Manager	98%	365	2546	0.143362	Communication & PR
1	Data Collector	95%	299	2148	0.139199	Data & IT
2	Data Entry	95%	325	2278	0.142669	Administrative & Clerical
3	Mail Clerk	95%	193	1366	0.141288	Leadership & Strategy
4	Compliance Officer	92%	194	1369	0.141709	Medical & Healthcare

Removing unnecessary columns:

```
In [7]: a1=df2['Job titiles']
```

```
In [8]: df2.drop('Job titiles',axis=1,inplace=True)
```

```
In [9]: df2.head()
```

Out[9]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Domain
0	98%	365	2546	0.143362	Communication & PR
1	95%	299	2148	0.139199	Data & IT
2	95%	325	2278	0.142669	Administrative & Clerical
3	95%	193	1366	0.141288	Leadership & Strategy
4	92%	194	1369	0.141709	Medical & Healthcare

Concating df2,a1 into a dataframe

```
In [10]: df3 = pd.concat([df2,a1],axis=1)
```

```
In [11]: df3.head()
```

Out[11]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Domain	Job titiles
0	98%	365	2546	0.143362	Communication & PR	Communications Manager
1	95%	299	2148	0.139199	Data & IT	Data Collector
2	95%	325	2278	0.142669	Administrative & Clerical	Data Entry
3	95%	193	1366	0.141288	Leadership & Strategy	Mail Clerk
4	92%	194	1369	0.141709	Medical & Healthcare	Compliance Officer

Dataframe for titles , display, type

```
In [12]: job=df3['Job titiles']

In [13]: job.head()

Out[13]: 0    Communications Manager
         1            Data Collector
         2                Data Entry
         3                Mail Clerk
         4        Compliance Officer
         Name: Job titiles, dtype: object

In [14]: type(job)

Out[14]: pandas.core.series.Series

In [15]: jobtitles=[]
         for i in job:
             jobtitles.append(i)
```

Label Encoding

In [16]:
```
print(jobtitles)
```

```
Analyst', 'It Specialist', 'Junior Network Engineer', 'Network Consultant', 'Network Support Specialist', 'Salesforce Adminis
trator', 'Sap Functional Consultant', 'Support Technician', 'Crisis Counselor', 'Mechanical Drafter', 'Production Engineer',
'Security Engineer', 'Simulation Engineer', 'Underground Miner', 'Unix Engineer', 'Verification Engineer', 'Voice Engineer',
'Chief Administrative Officer', 'Emergency Management Specialist', 'Loss Prevention Manager', 'Proposal Coordinator', 'Provid
er Relations Representative', 'Secretary', 'Trading Assistant', 'Volunteer Coordinator', 'Car Driver', 'Carrier', 'Constructi
on Driver', 'Contract Driver', 'Courier', 'Forklift Driver', 'Tanker', 'Taxi Driver', 'Assistant Restaurant Manager', 'Banque
t Captain', 'Brewer', 'Chef Manager', 'Executive Chef', 'Food Production Worker', 'Front Of House Manager', 'Front Office Sup
ervisor', 'Greeter', 'Hotel Front Office Manager', 'Hotel General Manager', 'Hotel Operations Manager', 'Contract Negotiato
r', 'Housing Specialist', 'Proposal Specialist', 'Analytical Chemist', 'Cytogenetic Technologist', 'Food Scientist', 'Food Te
chnologist', 'Laboratory Assistant', 'Patent Agent', 'Taxonomist', 'Accounting Consultant', 'Accounting Technician', 'Account
s Receivable Clerk', 'Appraiser', 'Broker Assistant', 'Commercial Loan Officer', 'Commodity Trader', 'Corporate Accountant',
'Entry Level Accountant', 'Equity Trader', 'Financial Accountant', 'Fund Accounting Manager', 'General Accountant', 'General
Ledger Accountant', 'Investment Banking Analyst', 'Private Equity Analyst', 'Risk Analyst', 'Treasury Manager', 'Clinic Recep
tionist', 'Dental Receptionist', 'Front Office Assistant', 'Yard Jockey', 'Associate Consultant', 'Consultant', 'Financial Co
nsultant', 'Technical Consultant', 'Technology Consultant', 'Product Demonstrator', 'Promotions Manager', 'Purchasing Agent',
'Special Events Coordinator', 'Assistant Project Manager', 'Associate Product Manager', 'Product Director', 'Product Owner',
'Project Estimator', 'Project Lead', 'Project Leader', 'Project Management Specialist', 'Inventory Specialist', 'Merchandisin
g Assistant', 'Stock Manager', 'Architectural Drafter', 'Drilling Supervisor', 'General Contractor', 'Safety Professional',
'Safety Specialist', 'Adjuster', 'Claim Specialist', 'Claims Examiner', 'Claims Representative', 'Insurance Broker', 'Policy
Advisor', 'Mobile Phlebotomist', 'Orthopedic Technician', 'Aircraft Assembler', 'Aircraft Maintenance Technician', 'Aircraft
```

In [17]:
```
from sklearn.preprocessing import LabelEncoder
```

In [18]:
```
le = LabelEncoder()
```

In [19]:
```
jobcode={}
```

In [20]:
```
j=1
for i in jobtitles:
    jobcode[i]=j
    j+=1
```

Mapping the jobcode

In [21]:
```
print(jobcode)
```

```
ce Analyst': 331, 'Risk Management Analyst': 332, 'Salesforce Business Analyst': 333, 'Data Analyst
t': 335, 'Health Data Analyst': 336, 'Sql Dba': 337, 'Sql Server Dba': 338, 'Shipping Clerk': 339,
stration': 341, 'Administrative Director': 342, 'Assistant Administrator': 343, 'Compliance Analyst
nator': 345, 'Office Clerk': 346, 'Operations Clerk': 347, 'Procurement Clerk': 348, 'Receiver': 34
50, 'Registration Specialist': 351, 'Unit Secretary': 352, 'Healthcare Business Analyst': 353, 'Sur
t Office': 355, 'Demand Planner': 356, 'Logistics Coordinator': 357, 'Procurement Agent': 358, 'Pur
'Shipping Coordinator': 360, 'Supply Chain Coordinator': 361, 'Supply Coordinator': 362, 'Assignmer
k Agent': 364, 'Radio Operator': 365, 'Manual Qa Tester': 366, 'Manual Tester': 367, 'Coldfusion De
cientist': 369, 'Drupal Developer': 370, 'Java': 371, 'Java Developer': 372, 'Java Engineer': 373,
ava Software Developer': 375, 'Home Inspector': 376, 'Biomedical Engineer': 377, 'Biotechnology': 3
r': 379, 'Industrial Organizational Psychologist': 380, 'Accounts Payable': 381, 'Auditor': 382, 'E
ection Agent': 384, 'Collection Representative': 385, 'Collection Specialist': 386, 'Collections Sp
e Auditor': 388, 'Cost Controller': 389, 'Credit Controller': 390, 'Debt Collector': 391, 'Exchange
l Engineer': 393, 'Financial Examiner': 394, 'Loan Closer': 395, 'Mortgage Closer': 396, 'Mortgage
sor': 398, 'Trader': 399, 'Treasury Accountant': 400, 'Clinical Analyst': 401, 'Meter Reader': 402,
403, 'Lactation Consultant': 404, 'Ticket Taker': 405, 'Usher': 406, '911 Dispatcher': 407, '911 Op
Technician': 409, 'Forensic Examiner': 410, 'Forensic Scientist': 411, 'Intelligence': 412, 'Milita
Coordinator': 414, 'Security Technician': 415, 'Skip Tracer': 416, 'Information Analyst': 417, 'Inf
'Network Systems Administrator': 419, 'Senior Systems Engineer': 420, 'Software Engineer': 421, 'Sy
'Systems Analyst': 423, 'Unix Administrator': 424, 'Unix System Administrator': 425, 'Vmware Admini
```

In [22]:
```
df3['job_title_code']=df3['Job titiles'].map(jobcode)
```

In [23]:
```
df3.head()
```

Out[23]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Domain	Job titiles	job_title_code
0	98%	365	2546	0.143362	Communication & PR	Communications Manager	1
1	95%	299	2148	0.139199	Data & IT	Data Collector	2
2	95%	325	2278	0.142669	Administrative & Clerical	Data Entry	3
3	95%	193	1366	0.141288	Leadership & Strategy	Mail Clerk	4
4	92%	194	1369	0.141709	Medical & Healthcare	Compliance Officer	5

Dropping the job titles

In [24]:
```
df3=df3.drop('Domain',axis=1)
df3.head()
```

Out[24]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Job titiles	job_title_code
0	98%	365	2546	0.143362	Communications Manager	1
1	95%	299	2148	0.139199	Data Collector	2
2	95%	325	2278	0.142669	Data Entry	3
3	95%	193	1366	0.141288	Mail Clerk	4
4	92%	194	1369	0.141709	Compliance Officer	5

In [25]:
```
dff=df3.drop('Job titiles',axis=1)
```

In [26]:
```
df3
```

Out[26]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Job titiles	job_title_code
0	98%	365	2546	0.143362	Communications Manager	1
1	95%	299	2148	0.139199	Data Collector	2
2	95%	325	2278	0.142669	Data Entry	3
3	95%	193	1366	0.141288	Mail Clerk	4
4	92%	194	1369	0.141709	Compliance Officer	5
...	...	...	...	...	...	...
4701	5%	686	2798	0.245175	Singer	4702
4702	5%	556	2206	0.252040	Airport	4703
4703	5%	1316	4695	0.280298	Director	4704
4704	5%	710	2594	0.273709	Nurse	4705

Displaying df3

```
In [27]: df3.head()
```

Out[27]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Job titiles	job_title_code
0	98%	365	2546	0.143362	Communications Manager	1
1	95%	299	2148	0.139199	Data Collector	2
2	95%	325	2278	0.142669	Data Entry	3
3	95%	193	1366	0.141288	Mail Clerk	4
4	92%	194	1369	0.141709	Compliance Officer	5

```
In [28]: #Applying the model on the dataset
```

```
In [29]: df3.head()
```

Out[29]:

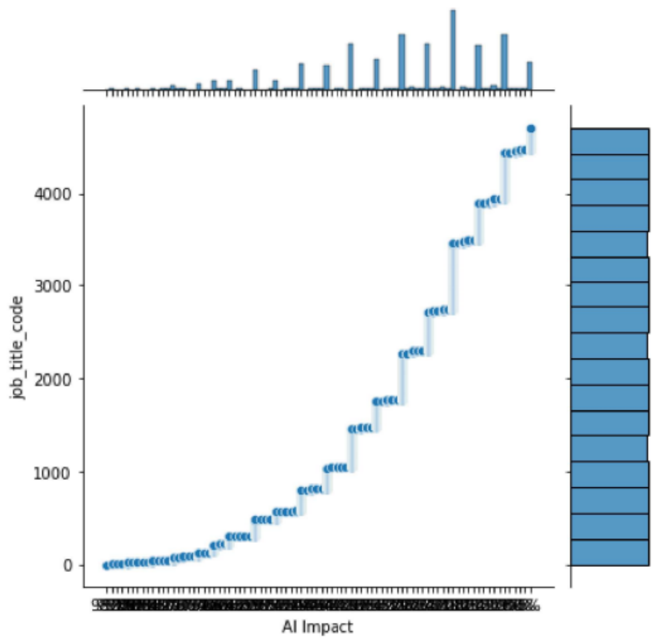
	AI Impact	Tasks	AI models	AI_Workload_Ratio	Job titiles	job_title_code
0	98%	365	2546	0.143362	Communications Manager	1
1	95%	299	2148	0.139199	Data Collector	2
2	95%	325	2278	0.142669	Data Entry	3
3	95%	193	1366	0.141288	Mail Clerk	4
4	92%	194	1369	0.141709	Compliance Officer	5

```
In [31]: import seaborn as sns
```

Data Visualiation

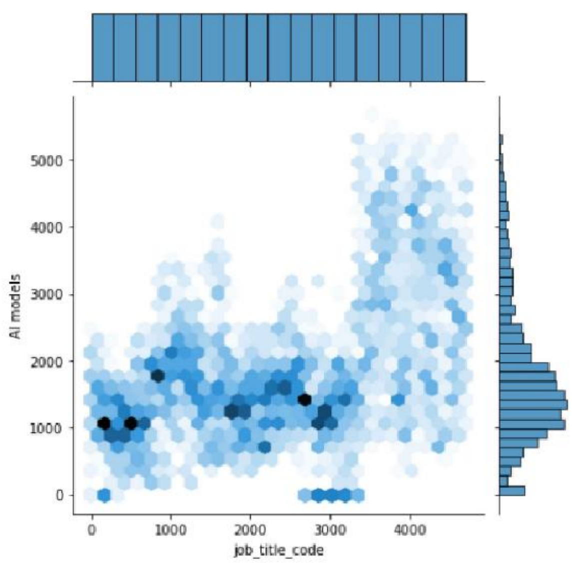
```
In [32]: sns.jointplot(x='AI Impact',y='job_title_code',data=df3)
```

```
Out[32]: <seaborn.axisgrid.JointGrid at 0x258f5faffa0>
```



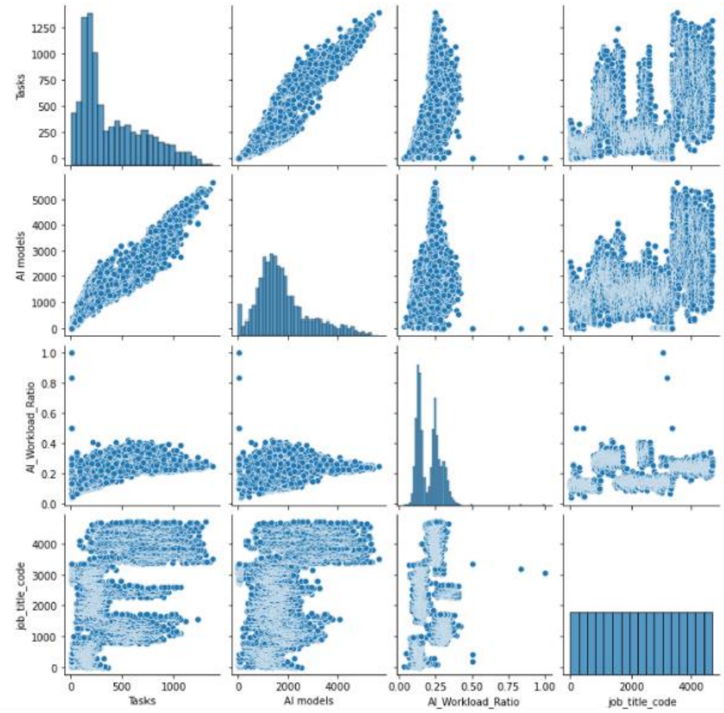

```
In [33]: sns.jointplot(x='job_title_code',y='AI models',kind='hex',data=df3)
```

```
Out[33]: <seaborn.axisgrid.JointGrid at 0x25892f38580>
```



```
In [34]: sns.pairplot(df3)
```

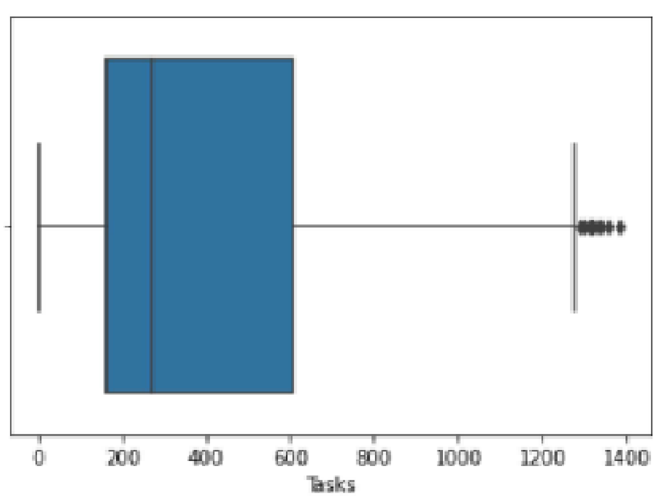
Out[34]: <seaborn.axisgrid.PairGrid at 0x258943d2760>



Remove Outliners

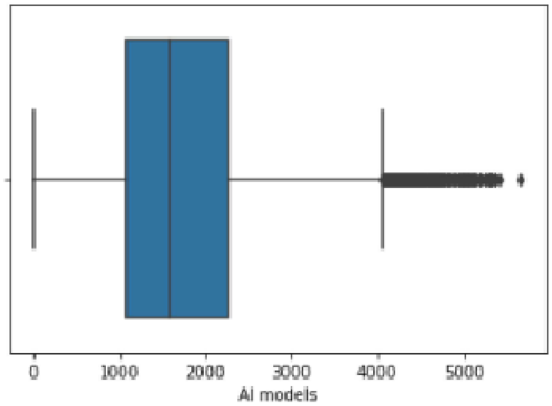
```
In [35]: import seaborn as sns
         import matplotlib.pyplot as plt
         sns.boxplot(x='Tasks',data=df3)
```

Out[35]: <AxesSubplot:xlabel='Tasks'>



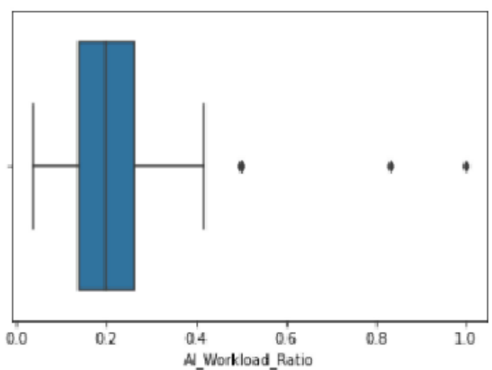
In [36]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x='AI models',data=df3)
```

Out[36]: <AxesSubplot:xlabel='AI models'>



In [38]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x='AI_Workload_Ratio',data=df3)
```

Out[38]: <AxesSubplot:xlabel='AI_Workload_Ratio'>

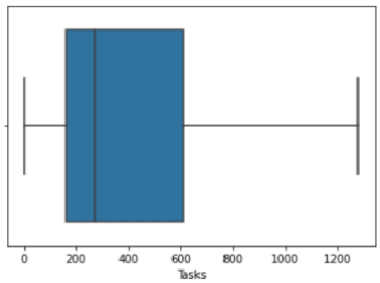


In [40]:
```
#REMOVING OUTLIERS
```

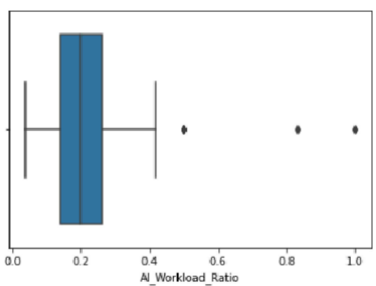
In [42]:
```
import numpy as np
```

In [43]:
```
percentile25=df3['Tasks'].quantile(0.25)
percentile75=df3['Tasks'].quantile(0.75)
iqr=percentile75-percentile25
upperlimitpm1=percentile75+1.5*iqr
lowerlimitpm1=percentile25-1.5*iqr
df3['Tasks']=np.where(df3['Tasks']>upperlimitpm1,upperlimitpm1,np.where(df3['Tasks']<lowerlimitpm1,lowerlimitpm1,df3['Tasks']))
```

In [44]:
```
sns.boxplot(x='Tasks',data=df3)
plt.show()
```



In [60]:
```
sns.boxplot(x='AI_Workload_Ratio',data=df3)
plt.show()
```



In [45]:
```
5=df3['AI models'].quantile(0.25)
5=df3['AI models'].quantile(0.75)
ile75-percentile25
m1=percentile75+1.5*iqr
m1=percentile25-1.5*iqr
els']=np.where(df3['AI models']>upperlimitpm1,upperlimitpm1,np.where(df3['AI models']<lowerlimitpm1,lowerlimitpm1,df3['AI models'
```

Converting AI Impact from String datatype to numerical

```
In [47]: df3.head()
Out[47]:
```

	AI Impact	Tasks	AI models	AI_Workload_Ratio	Job titiles	job_title_code
0	98%	365.0	2546.0	0.143362	Communications Manager	1
1	95%	299.0	2148.0	0.139199	Data Collector	2
2	95%	325.0	2278.0	0.142669	Data Entry	3
3	95%	193.0	1366.0	0.141288	Mail Clerk	4
4	92%	194.0	1369.0	0.141709	Compliance Officer	5

```
In [49]: df3=df3.drop('Job titiles',axis=1)
```

```
In [54]: for i in range(len(df3['AI Impact'])):
             s=df3['AI Impact'][i]
             n=len(s)
             a=s[0:n-1]
             a=int(a)
             df3['AI Impact'][i]=a
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_23032\2622634572.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df3['AI Impact'][i]=a
```

Rounding off the AI_Workload_Ratio

```
In [55]: df3.head()
Out[55]:
```

	AI Impact	Tasks	AI models	AI_Workload_Ratio	job_title_code
0	98	365.0	2546.0	0.143362	1
1	95	299.0	2148.0	0.139199	2
2	95	325.0	2278.0	0.142669	3
3	95	193.0	1366.0	0.141288	4
4	92	194.0	1369.0	0.141709	5

```
In [63]: for i in range(len(df3['AI_Workload_Ratio'])):
             a=df3['AI_Workload_Ratio'][i]
             b=round(a,2)
             df3['AI_Workload_Ratio'][i]=b
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_23032\3517472379.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  df3['AI_Workload_Ratio'][i]=b
```

Replacing null values with nan's

```
In [67]: import numpy as np

         # Check for NaN values
         nan_indices = np.isnan(df3['AI_Workload_Ratio'])

         # Check for infinite values
         inf_indices = np.isinf(df3['AI_Workload_Ratio'])

         # Print the indices where NaN or infinite values are present
         print("NaN indices:", np.where(nan_indices))
         print("Infinite indices:", np.where(inf_indices))

         NaN indices: (array([], dtype=int64),)
         Infinite indices: (array([3034, 3035, 3036, 3037, 3184, 3211, 3322], dtype=int64),)

In [74]: df3['AI_Workload_Ratio'][3034]

Out[74]: inf

In [75]: df3.replace([np.inf, -np.inf], np.nan, inplace=True)
```

Replacing null values with nan's

```
In [76]: import numpy as np

         # Check for NaN values
         nan_indices = np.isnan(df3['Tasks'])

         # Check for infinite values
         inf_indices = np.isinf(df3['Tasks'])

         # Print the indices where NaN or infinite values are present
         print("NaN indices:", np.where(nan_indices))
         print("Infinite indices:", np.where(inf_indices))

         NaN indices: (array([], dtype=int64),)
         Infinite indices: (array([], dtype=int64),)
```

Dropping the converted nan values from the dataframe

In [78]:
```
df3=df3.dropna()
```

In [79]:
```
df3.head()
```

Out[79]:

	AI Impact	Tasks	AI models	AI_Workload_Ratio	job_title_code
0	98	365.0	2546.0	0.14	1
1	95	299.0	2148.0	0.14	2
2	95	325.0	2278.0	0.14	3
3	95	193.0	1366.0	0.14	4
4	92	194.0	1369.0	0.14	5

Splitting the dataframe

In [81]:
```
from sklearn.model_selection import train_test_split
```

In [82]:
```
X_train, X_test, y_train, y_test = train_test_split(df3.drop('job_title_code',axis=1),
                                                    df3['job_title_code'], test_size=0.30)
```

In [105]:
```
from sklearn.preprocessing import MinMaxScaler

# Assuming 'data' is your DataFrame or array
# Replace 'data' with your actual data

# Create a MinMaxScaler
scaler = MinMaxScaler()

# Fit the scaler on the data and transform the data
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)
```

Applying the KNN Classifier on the split data

```
In [106]: from sklearn.neighbors import KNeighborsClassifier

In [107]: knn = KNeighborsClassifier(n_neighbors=1)

In [108]: knn.fit(X_train_normalized,y_train)

Out[108]: KNeighborsClassifier(n_neighbors=1)

In [109]: y_pred = knn.predict(X_test_normalized)

In [111]: from sklearn.metrics import classification_report,confusion_matrix

In [116]: from sklearn.metrics import mean_absolute_error
          mae = mean_absolute_error(y_test, y_pred)
          print(mae)

          95.92056737588652
```

```
In [113]: from sklearn.metrics import r2_score

          r2=r2_score(y_test,y_pred)
          print(r2)

          0.9872609984254223

In [120]: mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
          print(mape)

          6.025627249843691

In [121]: from sklearn.metrics import explained_variance_score
          explained_var = explained_variance_score(y_test, y_pred)
          print(explained_var)

          0.9872624088748995
```

8. RESULTS


Job Threat Index
Tasks
AIModels
AI_Workload_Ratio
Job_Title_Code
Probability of Automation


Job Threat Index
Tasks
AIModels
AI_Workload_Ratio
Job_Title_Code
Probability of Automation
Probability of taking over by AI is 85

9. CONCLUSION AND FUTURE SCOPE

In conclusion, our predictive model for job threat indexing, incorporating features such as job titles, AI impact percentages, task counts, AI model presence, and AI workload ratios, has yielded insightful results. Specifically, for data entry jobs, our model predicts an 89% likelihood of replacement due to the influence of artificial intelligence. This finding underscores the vulnerability of certain job roles to technological advancements and the need for proactive measures to address potential workforce disruptions.

The major key findings are

High AI Impact: The substantial AI impact percentage indicates a significant influence on data entry jobs, suggesting a notable potential for automation.

Task Replacement: The high predicted replacement percentage implies that a substantial portion of tasks associated with data entry roles could be automated by AI systems.

AI Model Integration: The presence of AI models associated with data entry jobs suggests an ongoing integration of advanced technologies in this field.

Workload Redistribution: The computed AI workload ratio indicates a considerable shift in workload distribution, highlighting the transformation of traditional data entry tasks through the adoption of AI.

Future Scope:

Skill Enhancement Programs: Given the high predicted replacement percentage, there is a clear need for skill enhancement programs targeting individuals in data entry roles. Focusing on developing skills that complement AI capabilities can enhance employability and job resilience.

Job Redefinition Strategies: Organizations can explore strategies to redefine the roles of data entry professionals, emphasizing tasks that complement AI capabilities and require human expertise, such as data quality assurance and complex problem-solving.

Continuous Monitoring: As technology evolves, continuous monitoring and updates to the predictive model will be essential to adapt to changing job dynamics and ensure the accuracy of predictions.

Expanding the Model: Future iterations of the model could incorporate additional features, such as educational qualifications, industry-specific trends, and geographical considerations, to provide a more comprehensive assessment of job threats.

10. REFERENCES

1. https://www.timextender.com/blog/data-empowered-leadership/rise-of-the-machines-impact-artificial-intelligence-machine-learning-data-automation-integration
2. https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for
3. https://www.sciencedirect.com/science/article/pii/S2667241323000113
4. https://www.run.ai/guides/machine-learning-engineering/machine-learning-automation
5. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Driving%20impact%20at%20scale%20from%20automation%20and%20AI/Driving-impact-at-scale-from-automation-and-AI.ashx
6. https://hubtgi.com/the-benefits-of-machine-learning-and-ai-in-process-automation/
7. https://emeritus.org/blog/ai-and-ml-benefits-of-ai-automation/
8. https://www.machinemetrics.com/blog/the-impact-of-ai-and-machine-learning-on-manufacturing
9. https://www.concur.com/blog/article/how-automation-artificial-intelligence-and-machine-learning-are-helping-organizations
10. https://builtin.com/machine-learning/automl-automated-machine-learning