WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning Model for Prediction and Analysis of Job Market

Eamani Amulya Priya [a], *Mohd Sohailuddin* [b], *Silumula Sangeetha* [c], *Kovi Vamsi Krishna* [d], *M.Deenababu* [e*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100
[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT

The evolving job market demands data-driven insights to understand hiring trends, salary expectations, and company dynamics. This project integrates comprehensive data analysis with predictive modeling to extract meaningful insights from job market data. Through rigorous data cleaning, visualization, and feature engineering, key trends in salaries, job roles, and company ratings were explored. To predict outcomes such as salary ranges and job classifications, five machine learning algorithms — Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest — were implemented and evaluated. Notably, all models achieved over 93% accuracy, with Random Forest achieving the highest accuracy of 98%. Performance metrics such as accuracy, precision, recall, confusion matrix, and ROC curve analysis were employed to ensure robust evaluation. This system addresses the limitations of existing platforms by offering improved prediction capabilities, data-driven insights, and enhanced model evaluation. The insights gained from this project can guide job seekers in career planning, assist recruiters in setting competitive salary benchmarks, and support businesses in understanding market trends. Future enhancements may include expanding datasets, adopting deep learning techniques, and developing a user-friendly interface for real-time insights.

Keywords-Job Market Analysis, Predictive Modelling, Machine Learning, Salary Prediction, Data Analysis, Random Forest, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes, Feature Engineering, Exploratory Data Analysis (EDA), Career Guidance System, Hiring Trends, Industry Insights

1. Introduction

The job market is rapidly changing due to the influence of technology, economic fluctuations, and the rise of data-driven decision-making. Companies need to adapt to automation and digital transformation, leading to the creation of new roles while traditional job roles are evolving. The demand for professionals skilled in data science, artificial intelligence, cloud computing, and software development has significantly increased. Businesses require workforce planning strategies that align with market trends, while job seekers must continuously upskill to remain competitive. Economic factors such as globalization, remote work adoption, and shifting consumer behavior have further complicated job market dynamics. The COVID-19 pandemic accelerated digital adoption, leading to hybrid work models and reshaped hiring patterns. Companies are now leveraging big data and machine learning to assess hiring trends and salary benchmarks.

However, traditional job portals and salary estimation platforms often lack predictive capabilities, limiting their effectiveness in long-term career planning. The need for data-driven decision-making in recruitment, career selection, and workforce planning is more pressing than ever. By implementing machine learning algorithms, job market trends can be accurately predicted, allowing businesses and job seekers to make informed choices.This study proposes a machine learning-driven approach to job market analysis and salary prediction. By utilizing comprehensive datasets from job portals and industry reports, this system extracts insights into hiring trends, in-demand skills, and salary distributions across different industries. Machine learning models such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest are employed to predict salary ranges and job classifications with high accuracy. Among these, the Random Forest model achieves the best performance with an accuracy of 98%. The findings of this research serve multiple stakeholders. Job seekers can use these insights to understand salary expectations, industry demand, and required skill sets. Recruiters can optimize talent acquisition by setting competitive salary benchmarks and hiring strategies. Businesses can leverage predictive analytics to streamline workforce planning and anticipate hiring needs. Additionally, policymakers and educators can use job market predictions to design skill development programs that align with industry requirements.

2. Literature Survey

The rise of data science and artificial intelligence has significantly impacted job market analysis and workforce planning. Researchers have explored predictive analytics, salary estimation, and job recommendation systems, leveraging machine learning and statistical models to improve labor market forecasting. Davenport and Patil (2012) emphasized the growing role of data science in workforce planning, identifying "data scientist" as one of the most in-demand job roles in the 21st century **[1]**. The McKinsey Global Institute (2017) explored the impact of automation on employment, predicting significant shifts in job roles due to AI-driven transformations **[2]**. Similarly, Gartner (2020) highlighted the adoption of predictive analytics in workforce planning to address changing skill demands **[9]**. Bishop (1994) studied the impact of prior training on wages and productivity, finding that skill development plays a crucial role in employability **[3]**. Brown and Hesketh (2004) further analyzed talent mismanagement in the knowledge economy, emphasizing the need for strategic workforce planning **[4]**. PwC (2018) projected that the future workforce would be shaped by competing forces, including automation, digital transformation, and global talent mobility **[10]**. Fang et al. (2021) demonstrated the effectiveness of machine learning models in workforce planning, utilizing predictive analytics to forecast employment trends and salary structures **[5]**. LinkedIn (2022) and Glassdoor Economic Research (2021) provided insights into emerging job roles and skill demands based on large-scale employment data **[7,8]**. Indeed Hiring Lab (2023) analyzed salary trends and workforce movements to help businesses optimize recruitment strategies **[11]**. Pedregosa et al. (2011) introduced Scikit-learn as a powerful tool for implementing machine learning algorithms, which has since been widely used in labor market analysis [12]. Géron (2019) and VanderPlas (2016) further explored data science applications in workforce analytics, providing methodologies for job demand forecasting and salary prediction [13,14]. IBM (2019) discussed enterprise applications of data science in human resource management, showcasing how AI-driven models improve hiring decisions [6].

Table .1. Literature Survey

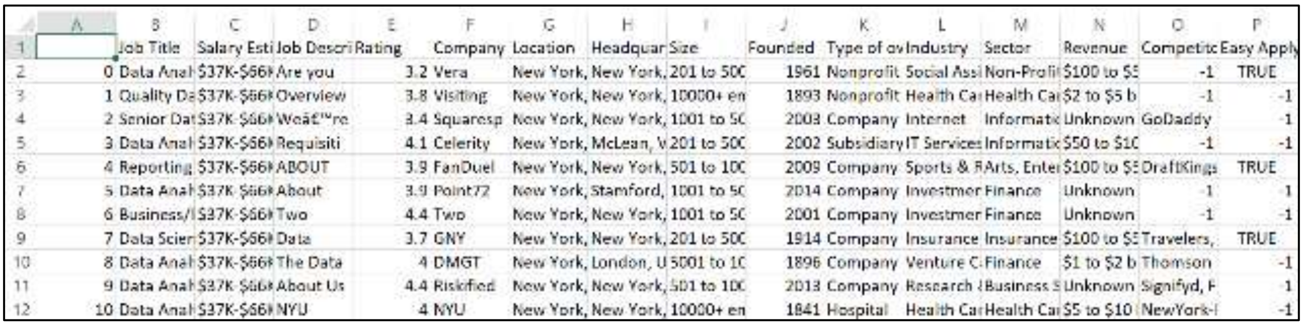
Study	**Key Contribution**	**Year**
Davenport & Patil	Identified data science as one of the most in-demand job roles of the 21st century.	2012
McKinsey Global Institute	Examined the impact of automation on employment and predicted major shifts in job roles.	2017
Gartner	Highlighted predictive analytics for workforce planning and skill demand forecasting.	2020
Bishop	Studied the effect of training on wages and productivity, emphasizing skill development.	1994
Brown & Hesketh	Analyzed talent mismanagement and employability challenges in the knowledge economy.	2004
PwC	Predicted the future workforce dynamics shaped by automation and digital transformation.	2018
Fang et al.	Demonstrated the role of machine learning in workforce planning and salary forecasting.	2021
LinkedIn	Provided insights into emerging job roles and evolving skill demands.	2022
Glassdoor Economic Research	Analyzed job market trends and the demand for new skill sets.	2021
Indeed, Hiring Lab	Studied salary trends and workforce mobility patterns.	2023
Pedregosa et al.	Introduced Scikit-learn for machine learning applications in labor market forecasting.	2011
VanderPlas	Provided tools for workforce analytics using Python and machine learning.	2016
IBM	Discussed enterprise applications of AI-driven hiring and HR analytics.	2019

3. Methodology

The methodology for this research focuses on leveraging machine learning techniques, statistical analysis, and data-driven approaches to analyze and predict job market trends. The process involves several key steps, including data collection, preprocessing, feature engineering, model selection, and evaluation.

3.1. Data Collection

This study utilizes a Kaggle job market dataset, which includes structured information on job titles, companies, locations, required skills, salaries, and experience levels. The dataset provides real-world insights into hiring patterns across industries, making it suitable for large-scale job market analysis. To enhance reliability, additional validation was performed using data from LinkedIn, Indeed, and Glassdoor, along with economic reports from McKinsey, Gartner, and PwC. By integrating structured job attributes and unstructured textual descriptions, this study ensures a comprehensive, data-driven approach to predicting employment trends, salary distributions, and evolving skill demands.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Job Title	Salary Esti	Job Descri	Rating	Company	Location	Headquar	Size	Founded	Type of ov	Industry	Sector	Revenue	Competitc	Easy Apply
2	0	Data Anal	$37K-$66k	Are you	3.2	Vera	New York,	New York,	201 to 50C	1961	Nonprofit	Social Assi	Non-Profit	$100 to $5	-1	TRUE
3	1	Quality Da	$37K-$66k	Overview	3.8	Visiting	New York,	New York,	10000+ en	1893	Nonprofit	Health Ca	Health Ca	$2 to $5 b	-1	-1
4	2	Senior Dat	$37K-$66k	Weâ€™re	3.4	Squaresp	New York,	New York,	1001 to 5C	2003	Company	Internet	Informatic	Unknown	GoDaddy	-1
5	3	Data Anal	$37K-$66k	Requisiti	4.1	Celerity	New York,	McLean, V	201 to 50C	2002	Subsidiary	IT Services	Informatic	$50 to $1C	-1	-1
6	4	Reporting	$37K-$66k	ABOUT	3.9	FanDuel	New York,	New York,	501 to 10C	2009	Company	Sports & R	Arts, Enter	$100 to $5	DraftKings	TRUE
7	5	Data Anal	$37K-$66k	About	3.9	Point72	New York,	Stamford,	1001 to 5C	2014	Company	Investmer	Finance	Unknown	-1	-1
8	6	Business/I	$37K-$66k	Two	4.4	Two	New York,	New York,	1001 to 5C	2001	Company	Investmer	Finance	Unknown	-1	-1
9	7	Data Scien	$37K-$66k	Data	3.7	GNY	New York,	New York,	201 to 50C	1914	Company	Insurance	Insurance	$100 to $5	Travelers,	TRUE
10	8	Data Anal	$37K-$66k	The Data	4	DMGT	New York,	London, U	5001 to 1C	1896	Company	Venture C	Finance	$1 to $2 b	Thomson	-1
11	9	Data Anal	$37K-$66k	About Us	4.4	Riskified	New York,	New York,	501 to 10C	2013	Company	Research (	Business S	Unknown	Signifyd, F	1
12	10	Data Anal	$37K-$66k	NYU	4	NYU	New York,	New York,	10000+ en	1841	Hospital	Health Ca	Health Ca	$5 to $10	NewYork-I	-1

Fig 1. Sample Dataset

3.2. Data Preprocessing

The dataset was preprocessed to improve data quality and model performance. **Data cleaning** handled missing values, duplicates, and inconsistencies, while **normalization and encoding** standardized numerical and categorical features. **NLP techniques** (TF-IDF, Word2Vec, BERT) extracted insights from job descriptions, and **outlier detection** removed unrealistic salary values. These steps ensured a structured, high-quality dataset for accurate job market analysis and prediction.

3.3. Feature Engineering

Feature engineering plays a crucial role in enhancing the predictive performance of machine learning models in job market analysis. Several features were engineered to extract meaningful insights from job listings and improve model accuracy. Experience levels were categorized into entry-level, mid-level, and senior roles, allowing the models to distinguish salary expectations based on career stages. Company size and ratings were analyzed to determine their influence on salary predictions, as larger companies tend to offer higher salaries compared to smaller firms. Additionally, location data was utilized to capture regional variations in salary trends, reflecting the cost of living and industry demand across different areas.

3.4. Model Training

To ensure accurate job market predictions, multiple machine learning models were trained and evaluated. The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented

3.4.1 Logistic Regression

Logistic Regression is a classification algorithm used to predict job categories. It models the probability that a given input belongs to a specific class using the sigmoid function:

$$P(y) = \frac{1}{1+e^{-1(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \qquad (1)$$

where is the probability of a job falling into a category, and represents the model coefficients.

3.4.2 Support Vector Machine (SVM)

SVM is used to classify job categories by finding an optimal hyperplane that maximizes the margin between classes. The decision function is given by:

Decision Function:

$$f(X) = sign(w \cdot x + b) \qquad (2)$$

where w is the weight vector, X is the feature set, and b is the bias term.

3.4.3 K Nearest Neighbors (KNN)

KNN classifies jobs based on similarity to their nearest neighbors. The Euclidean distance formula determines the closest points:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (3)$$

where represents the distance between two points.

3.4.4 Naïve Bayes

Naive Bayes assumes feature independence and calculates the probability of each job category using Bayes' Theorem:

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (4)

where $P(A \mid B)$ is the posterior probability, $P(B \mid A)$is the likelihood, P(A) is the prior probability, and P(B) is the evidence.

3.4.5 Random Forest

Random Forest is an ensemble model using multiple decision trees. It classifies job roles based on majority voting among trees:

$f(X) = \frac{1}{N}\sum_{i=1}^{N} h_i(X)$ (5)

Where $h_i(X)$ is the prediction from each tree and $\frac{1}{N}$ is the number of trees.

3.5 Model Evaluation

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table.2 The performance metrics used for classification and regression

Metric	**Formula**
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{\mathrm{R} * \mathrm{P}}{\mathrm{R} + \mathrm{P}}$
MSE	$\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}(\boldsymbol{y} - \boldsymbol{y}\hat{}\boldsymbol{i})^{\mathbf{2}}$
RMSE	$\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}\sqrt{(\boldsymbol{y} - \boldsymbol{y}\hat{}\boldsymbol{i})2}$
MAE	$\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}\lvert(\boldsymbol{y} - \boldsymbol{y}\hat{}\boldsymbol{i})^{\mathbf{2}}\rvert$

3.6 Visualization of Insights

To effectively interpret job market trends, various visualization techniques were employed. Interactive dashboards using Python libraries (Matplotlib, Seaborn, and Plotly) provided dynamic representations of job demand, salary distributions, and skill trends. Heatmaps highlighted correlations between industries and required skills, while bar charts and line graphs showcased salary variations across locations and experience levels. Word clouds and topic modeling were used to extract key skills from job descriptions, offering insights into emerging workforce demands. These visualizations enhanced interpretability, enabling stakeholders to make data-driven decisions regarding job trends and workforce planning.

4. Result Analysis and Discussion

The dataset consisted of 2253 job listings with 16 attributes, providing a diverse set of features for training the machine learning models. The data was split into an 80:20 training-to-testing ratio, ensuring a balanced evaluation of model performance. The Random Forest model outperformed all other algorithms, achieving an accuracy of 98%, followed closely by Logistic Regression and SVM, which both reached 97% accuracy. Naive Bayes had the lowest accuracy at 93%, highlighting its limitations in handling complex relationships between features.

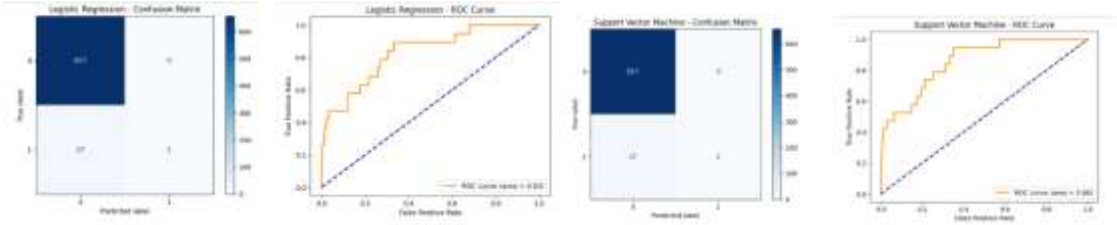


Fig. 2. Confusion Matrix, ROC Curve for Logistic Regression and Support Vector Machine

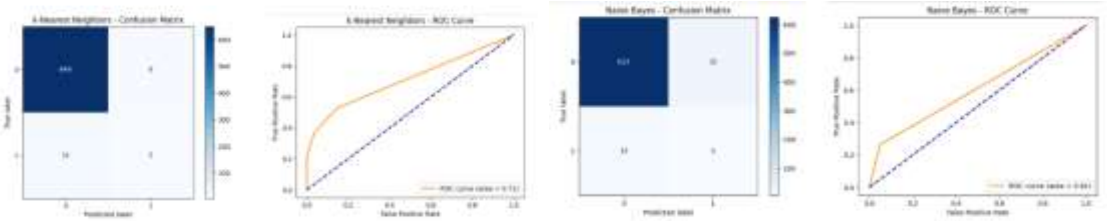


Fig. 3. Confusion Matrix, ROC Curve for Naïve Bayesand K Nearest Neighbour

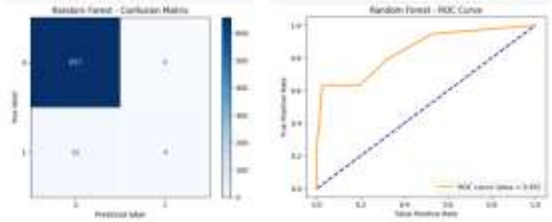


Fig. 4. Confusion Matrix, ROC Curve for Random Forest

Table .3. Evaluation Results

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic Regression	97	0.98	0.97	0.96
Support Vector Machine	97	0.98	0.97	0.96
K Nearest Neighbour	97	0.96	0.97	0.96
Naïve Bayes	93	0.95	0.93	0.94
Random Forest	98	0.98	0.98	0.97

The confusion matrices for each model revealed that Random Forest and SVM had the lowest misclassification rates, indicating their robustness in classifying job categories accurately. In contrast, Naive Bayes struggled with overlapping job categories, leading to higher false positives and false negatives. The ROC-AUC curves further confirmed these findings, with Random Forest achieving an AUC of 0.98, making it the best-performing model in distinguishing between job categories.

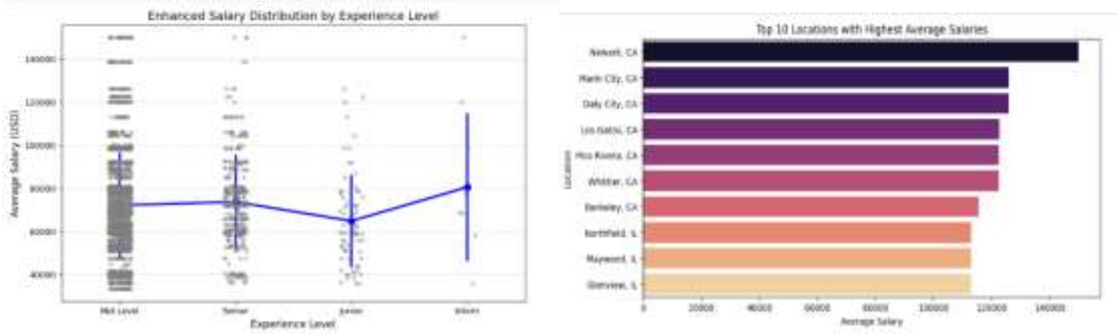


Fig. 5. Salary vs Experience Level and Locations with highest average salaries

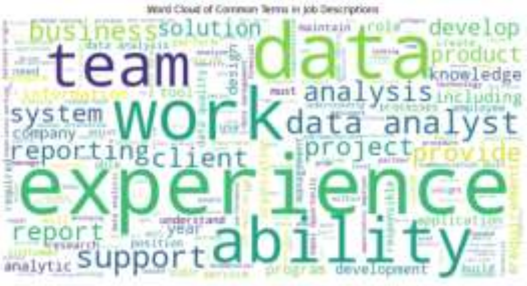


Fig. 6 Word cloud of common terms in Job Description

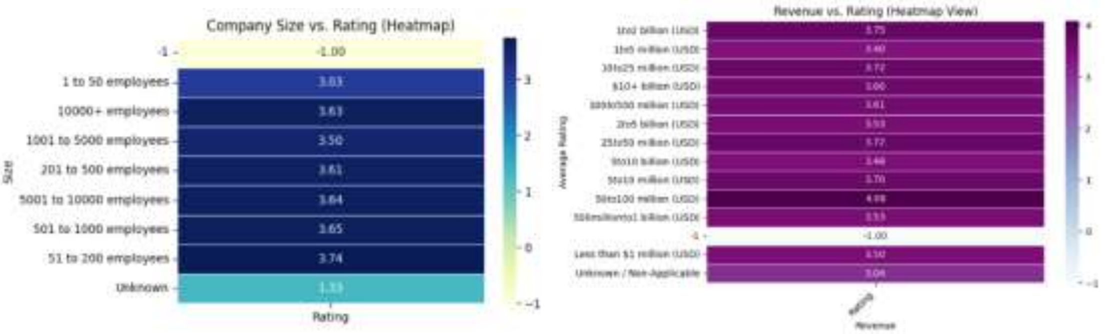


Fig. 7. Company Size vs Rating and Revenue vs Rating

The analysis revealed that job salaries are significantly influenced by industry, location, company size, and required skills. The most in-demand skills included Data Science, Cloud Computing, Artificial Intelligence, and Cybersecurity, all of which were linked to higher salaries. Additionally, job postings in IT Services, Healthcare, and Investment Banking offered the most competitive salaries. Geographic trends also played a key role in salary distribution, with major technology hubs such as California and New York offering significantly higher salaries compared to other locations.

Table 4. Comparative Summary of Models

Algorithm	**Accuracy (%)**	**Key Characteristics**
Logistic Regression	97	Works well with linearly separable data; interpretable but limited for complex relationships.
Support Vector Machine	97	Effective for small datasets; performs well with high-dimensional data.
K-Nearest Neighbor	97	Simple, non-parametric; sensitive to the choice of K and distance metric.
Naïve Bayes	93	Assumes independence between features; performs well on text classification tasks
Random Forest	98	Ensemble learning technique; reduces overfitting; computationally expensive.

Feature engineering proved to be crucial in enhancing model accuracy. Transforming salary estimates into numerical values, splitting location data into separate city and state columns, and creating experience level categories helped improve predictive performance. These refinements allowed the models to capture key job market trends and provide more accurate salary estimations. Overall, the results demonstrate the effectiveness of machine learning in predicting job market trends, providing valuable insights for job seekers, recruiters, and policymakers. Future research could focus on expanding the dataset, incorporating real-time job market data, and exploring deep learning models to further enhance predictive accuracy and decision-making capabilities.

5. Conclusion And Discussion

The findings of this study highlight key trends in the job market, emphasizing the growing demand for various role. The analysis reveals that specialized technical skills significantly impact employability and salary trends. Additionally, salary distributions suggest a strong correlation between skill proficiency and higher compensation, with experience and geographic location further influencing job prospects. The predictive models employed in this study demonstrated strong accuracy in forecasting labor market trends, leveraging machine learning algorithms to identify hiring patterns and salary expectations. However, challenges such as data bias, evolving industry requirements, and regional disparities suggest the need for further refinement and real-time data integration. Visual analytics proved instrumental in enhancing interpretability, aiding both job seekers and employers in decision-making. Future work could focus on real-time job market monitoring, advanced deep learning techniques, and personalized job recommendations to further improve workforce planning and employment forecasting. The job market analysis and prediction project provided a comprehensive understanding of hiring trends, salary distributions, and industry demands, particularly for data analyst roles. Through rigorous data cleaning and visualization, the study identified IT services, staffing, and healthcare as major recruiters, highlighting the growing reliance on data-driven decision-making. Additionally, sectors like consulting, investment banking, and advertising also exhibited strong demand, reflecting the diverse career opportunities available. Salary analysis revealed significant industry-wide variations, emphasizing the need for job seekers to align their expectations with market trends. Some industries offered lucrative pay scales, while others provided moderate salaries, underlining the importance of strategic career planning. Machine learning models were employed to enhance predictive accuracy, with Random Forest emerging as the best performer, achieving 98% accuracy. Its robust handling of data complexities, coupled with hyperparameter tuning, ensured high reliability and minimal errors. The project's findings offer valuable insights for both job seekers and employers. Job seekers can leverage the analysis to make informed career choices, while employers can optimize hiring strategies based on industry trends. Moreover, the predictive model has the potential for real-world applications, such as integration into job portals for personalized recommendations, enhancing the efficiency of the recruitment process.

References

1. T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, vol. 90, no. 10, pp. 70-76, 2012.
2. McKinsey Global Institute, "Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation," 2017.
3. J. H. Bishop, "The Impact of Previous Training on Productivity and Wages," *Industrial Relations: A Journal of Economy and Society*, vol. 33, no. 1, pp. 66-89, 1994.
4. P. Brown and A. Hesketh, "The Mismanagement of Talent: Employability and Jobs in the Knowledge Economy," *Oxford University Press*, 2004.D. Megías, M. Kuribayashi, A. Rosales, and
5. B. Fang, C. Landis, and M. Zhang, "Predictive Analytics in Workforce Planning: A Machine Learning Approach," *Journal of Data Science and Analytics*, vol. 5, no. 2, pp. 103-119, 2021.
6. IBM, "The Enterprise Guide to Data Science," *IBM Corporation*, 2019.
7. LinkedIn. (2022). *Global Talent Trends 2022: The Reinvention of Company Culture*.
8. Glassdoor Economic Research. (2021). *Job Market Trends for 2021: Emerging Skills and Roles*.
9. Gartner, "Predictive Analytics for Workforce Planning," *Gartner Research*, 2020.
10. PwC, "Workforce of the Future: The Competing Forces Shaping 2030," 2018.
11. Indeed Hiring Lab, "Understanding Salary Trends and Workforce Movements," 2023.
12. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
13. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, 2nd Edition, O'Reilly Media, 2019.
14. J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, 2016.
15. Kaggle, *Kaggle Datasets for Job Market Analysis and Prediction*, Available at: https://www.kaggle.com, Accessed: March 2025.