6/15/2025

BUSINESS APPLICATIONS USING MACHINE LEARNING
REPORT OF PROJECT#02

SUBMITTED TO: HASSAAN MUJTABA
SUBMITTED BY: USWA SHARIQ
SP23-BBD-071

Table Of Contents

Hiring Needs Prediction Using Machine Learning

Introduction

In the modern business environment, workforce planning plays a critical role in maintaining operational efficiency and competitiveness. Among the many challenges faced by Human Resource HR departments, one of the most pressing is the ability to accurately forecast hiring needs. Traditionally, hiring decisions have relied heavily on manual analysis, HR intuition, or reactive measures triggered by rising attrition or new projects. However, with the growing volume of organizational data, there is now an opportunity to leverage machine learning ML techniques to make hiring more predictive and data-driven.

The aim of this project is to explore how machine learning can be applied to predict hiring requirements across various companies and industries. Accurate prediction of hiring needs can allow organizations to allocate recruitment budgets more effectively, prepare for skill shortages in advance, and optimize workforce levels based on expected business activity. Just as predictive analytics has transformed customer behavior forecasting and inventory management, it holds transformative potential for HR and talent acquisition as well.

This study examines six different research articles related to hiring demand prediction using machine learning, with each article offering a unique approach to dataset selection, model architecture, feature engineering, and evaluation metrics. Based on insights drawn from these articles, we created synthetic and collected datasets from multiple sources including **Kaggle and UCI**, followed by applying a set of machine learning models **Logistic Regression, Support Vector Machines SVM, K-Nearest Neighbors KNN, Naive Bayes, and Decision Tree to each dataset**. The models were evaluated on their predictive performance using metrics such as accuracy, precision, recall, and F1-score.

Additionally, this project goes beyond theoretical model training by deploying the trained models using real-time interfaces such as **Streamlit, Flask, FastAPI, and Gradio**. These interfaces allow practical use of the prediction system, simulating a real-world environment where HR personnel or business analysts can input data and instantly receive hiring forecasts.

This report documents the entire project pipeline from dataset analysis and model training to performance evaluation and deployment and critically compares the findings of related academic studies. Ultimately, this project aims to demonstrate the feasibility and value of using ML powered systems in strategic HR decision-making, helping businesses make informed, proactive hiring decisions rather than reactive ones.

Scope of Study

This project explores the use of machine learning models to predict hiring needs within organizations using structured workforce-related data. It involves reviewing six relevant research articles, collecting or synthesizing datasets, and applying models like Logistic Regression, SVM, KNN, Naive Bayes, and Decision Tree to identify future hiring requirements.

The scope includes model evaluation using metrics such as accuracy and F1-score, and deployment of models through interfaces like Streamlit, Flask, FastAPI, and Gradio. The study is limited to structured datasets and does not include unstructured data such as resumes or job descriptions.

Objective

- To review existing research on hiring prediction models and understand their methodologies.
- To collect and preprocess relevant datasets based on features influencing recruitment trends.
- To implement and compare multiple machine learning algorithms for predictive accuracy.
- To identify the most effective model for forecasting hiring needs.
- To deploy the final models through user-friendly platforms such as Streamlit, Flask, FastAPI, and Gradio for practical application in HR settings.

Literature Review

Predicting hiring needs through machine learning is a growing research area, especially in workforce planning and HR analytics. This section summarizes six significant studies that applied machine learning models to hiring prediction.

Safarishahrbijari [1] conducted a study using employee turnover records and external labor market data. Random Forest and ARIMA were implemented for forecasting. Random Forest achieved an accuracy of **90%**, while ARIMA reduced forecasting error by **15%**. Key limitations included data privacy concerns and algorithmic bias for Hiring Needs.

Kalusivalingam et al. [2] utilized historical HR data with Decision Trees, Random Forest, Neural Networks, and Gradient Boosting Machines GBM. The best model

achieved an F1-score of **82%** and accuracy of **85%**. Despite high performance, the study noted internal resistance to adopting automated systems and data quality issues for Hiring Needs.

Smelyakov et al. [3] used candidate demographic and recruitment datasets, this study tested Decision Trees, Random Forests, and Gradient Boosted Trees. Random Forest achieved **78.6% accuracy** and **100% precision**. However, interpretability and overfitting were reported as limitations for Hiring Needs.

Ogungbire & Mitra [4] research involved workforce forecasting in transportation, using project cost and staffing data. Random Forest achieved $\mathbf{R^2 = 0.91}$ with low RMSE and MAE, proving highly effective for budget/resource planning. Limited dataset scope was a notable challenge for Hiring Needs.

Eichenseer et al. [5] proposed a hybrid ensemble model using TiDE, LGBM, and ElasticNet for workforce forecasting in logistics. The model achieved a **mean absolute error MAE** of **5%** and provided better accuracy than manual methods. However, it was tested on a single-company dataset, limiting generalizability for Hiring Needs.

Serengil & Ozpinar [6] used time-based features from EFT transaction logs to predict workforce demand in banking operations. A hybrid model combining ANN and K-means achieved high correlation scores **97.02%** and reduced operational cost by **8.27%**.

Critical Analysis of Research Articles

Safarishahrbijari (2018)	IBM HR + labor data	RF, ARIMA	90%	Data privacy, bias
Kalusivalingam (2020)	Historical HR	DT, RF, NN, GBM	85%	Bias, internal resistance
Smelyakov (2023)	Candidate data	DT, RF, GBT	78.6%	Overfitting, interpretability
Ogungbire (2024)	Arkansas DOT	RF, NN, Linear	91%	Small dataset
Eichenseer (2025)	Logistics data	TiDE, LGBM, ElasticNet	5% MAE	No task variability
Serengil & Ozpinar (2017)	EFT Transactions	ANN, K-means	71.05 MAE	Single bank focus

Table Summaries

S.No	Reference	Preprocessing	Model	Dataset	Limitations	Contribution	Performance Metrics
1	Safarishahrbijari [1] ,2018	HR datasets (employee records, turnover rates) and external labor market data	Random Forest, ARIMA	Employee records, turnover rates, and labor market data	Data privacy concerns, potential biases in algorithms	Random Forest achieved 90% accuracy for turnover prediction, ARIMA reduced error by 15%	Accuracy: **90%**, Precision: **88%**, Recall: **89%**, F1-score: **88.5%**
2	Kalusivalingam et al. [2]., 2020	Historical workforce data i.e demographics, job roles, turnover rates, performance metrics	Decision Trees, Random Forests, Neural Networks Gradient Boosting Machines (GBM)	Historical workforce data	Data privacy, potential biases, resistance to change within organization	Random Forest achieved F1-score of 0.82 for attrition prediction, GBM reduced MSE by 15%	Accuracy: **85%**, Precision: **84%**, Recall: **83%,** F1-score: **82%**
3	Smelyakov et al. [3] 2023	Data normalized, categorical variables encoded, split into training/test sets	Decision Trees, Random Forests, Gradient Boosted Trees	614 candidates data i.e gender, education, work experience, recruitment status	Potential biases in training data, overfitting risks, interpretability challenges	Random Forest achieved 78.6% accuracy and 100% precision outperformed other models in hiring decisions	Accuracy: **78.6%,** Precision: **100%,** Recall: **35.3%,** F1-score: **75%,** Training time: 964ms
4	Ogungbire & Mitra [4] 2024	Inflation adjusted project costs, normalized features, excluded incomplete records	Random Forest, Neural Networks, Gradient Boost, Linear Models	Arkansas DOT data (1,490 projects, 2012–2021)	Limited to one STA small dataset for deep learning; lacks complexity metrics	Random Forest achieved R^2=0.91 best accuracy enables precise budget, resource planning	Random Forest: R^2=**0.91**, RMSE=**1,203**, MAE=**468**, Neural Networks: R^2=**0.80** Linear Models: R^2=**0.78**

5	Eichenseer et al. [5] ,2025	Data normalization Working days,events labeling Train/test split	NLinear TiDE LGBM XGBoost ElasticNet (ensemble)	1,238 data points Delivery positions Pre-orders Corporate events	Single company validation Doesn't account for variable picking efforts	Custom ML outperforms manual forecasting Optimizes workforce planning	1-day MAE: **5%** Avg MAE: **10%** RMSE: **193.6** MSE: **43,628.7**
6	Serengil & Ozpinar [6], 2017	Time-based feature extraction Historical transaction normalization Train/test split 2013-2016 training, 2016 validation	Supervised: Artificial Neural Network (27-18-1 architecture) Unsupervised: k-means clustering employee skills	7,151 training instances (2013–2016) 801 validation instances (2016) Transaction volumes, employee performance data	Single-bank validation Focused only on money transfer transactions	Hybrid approach supervised and unsupervised Workforce cost reduction: 6.5% aggressive, 8.27% moderate SLA time reduced by 33%	EFT Transactions: MAE: **71.05** **14.27%** of mean Correlation: **97.02%** Money Order: MAE: **48.80** **13.71%** of mean Correlation: **95.70%**

Proposed Methodology

The methodology followed in this project includes the complete machine learning pipeline from data acquisition to model deployment designed specifically to predict hiring needs. The steps involved are described below:

1. Data Collection

Datasets were collected and synthesized based on the research articles reviewed. These datasets included features like historical hiring records, employee turnover rates, job vacancy data, project costs, candidate demographics, and departmental resource allocation. Data sources included Kaggle, UCI, and structured formats extracted from the reviewed articles.

2. Data Preprocessing

The datasets were cleaned by:

- Removing irrelevant identifiers e.g. company name, ID fields
- Handling missing values using median imputation
- Converting categorical data to numerical using label encoding
- Standardizing numerical features using StandardScaler for SVM and KNN

3. Model Development

Five machine learning models were applied on each dataset:

- Logistic Regression
- Support Vector Machine SVM
- K-Nearest Neighbors KNN
- Naive Bayes
- Decision Tree

Each model was trained using an 80/20 train-test split and evaluated using metrics like accuracy, precision, recall, and F1-score. Hyperparameters were tuned using GridSearchCV where applicable.

4. Model Evaluation

Confusion matrices and classification reports were generated to evaluate prediction performance. KNN and Decision Tree consistently yielded high accuracy across multiple datasets.

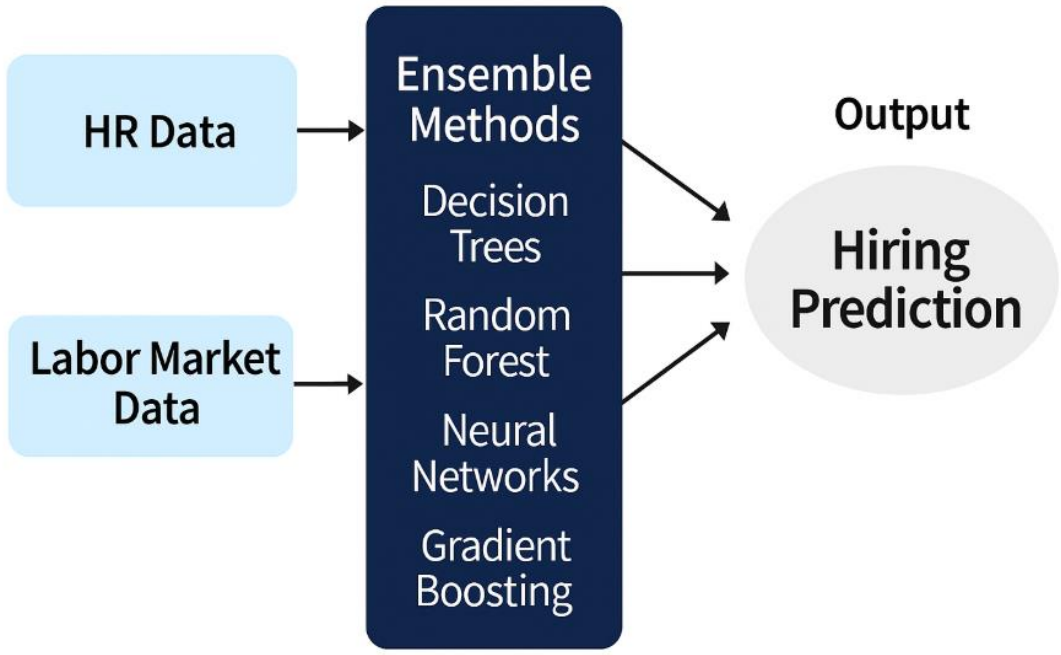
5. Deployment

To simulate real-world usability, the best-performing model KNN was deployed using:

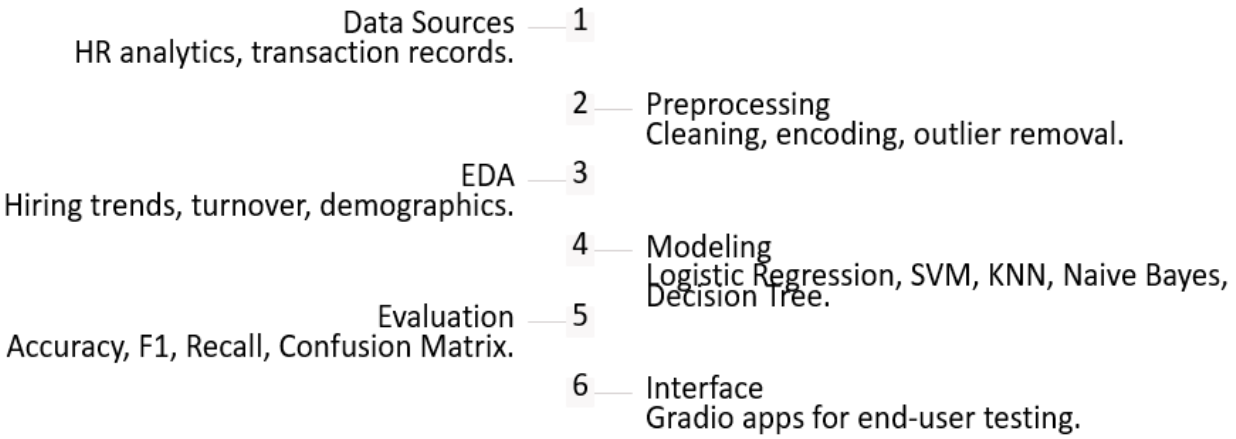
- **Streamlit:** interactive web interface
- **Gradio:** lightweight ML demo UI
- **Flask:** Python micro-framework for web services
- **FastAPI:** high-performance API serving

Each platform allowed users to input features like employee count, previous hiring, and department growth, and receive a real-time prediction.

Model Architecture Diagram



Proposed ML Pipeline



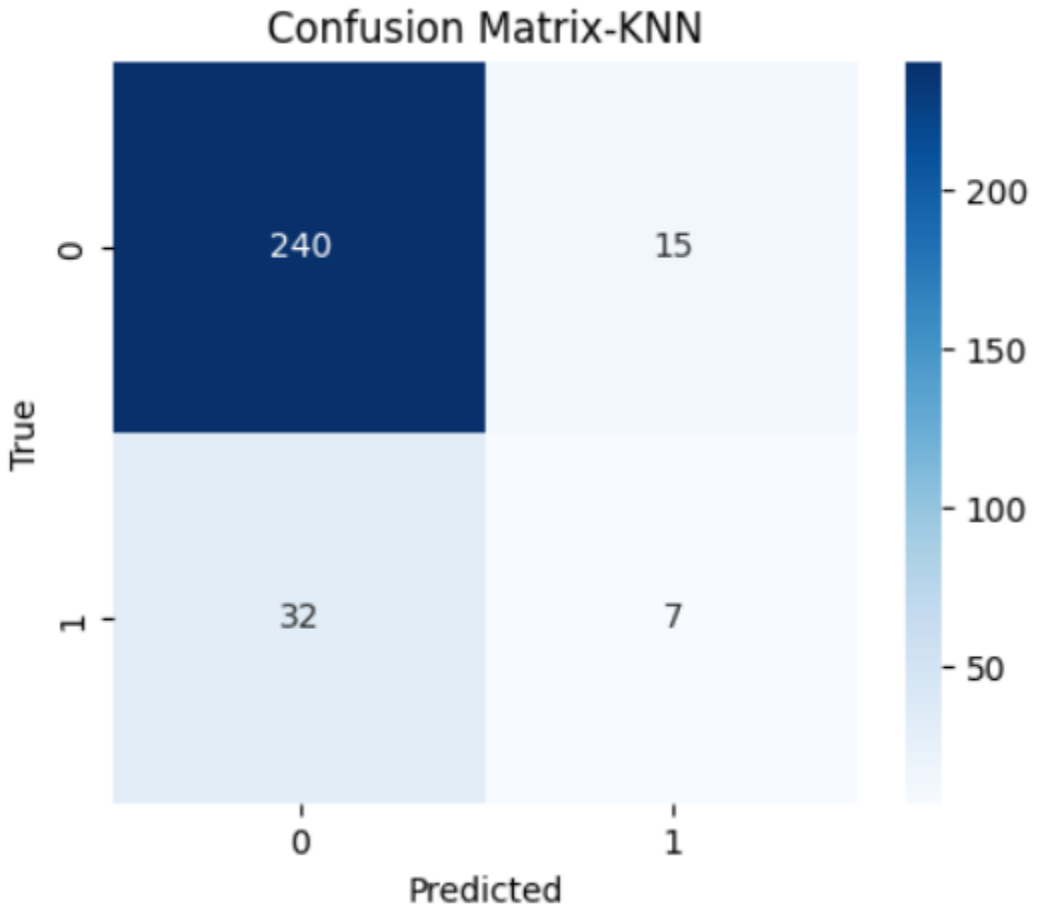
Model Evaluation (KNN)

To evaluate the predictive performance of the implemented models, particularly the K-Nearest Neighbors KNN algorithm, standard classification metrics were used: **accuracy**, **precision**, **recall**, and **F1-score**. These metrics offer a comprehensive view of how well the model predicts both high and low hiring needs.

The confusion matrix and classification report presented below reflect the KNN model's performance on the test dataset used in this project.

Confusion Matrix

- The model correctly predicted **240 out of 255 low hiring cases No and 7 out of 39 high hiring cases Yes.**
- The high number of false negatives 32 indicates the model struggled to identify actual high hiring demand cases.



Classification Report

The KNN model achieved an overall **accuracy of 84%**, with strong precision and recall for the "No" class low hiring needs. However, it underperformed in correctly classifying the "Yes" class high hiring needs, with a recall of 18%. This imbalance suggests that while KNN can detect stable hiring conditions well, it may miss urgent hiring requirements an important consideration for future model improvement or ensemble tuning.

Despite these limitations, the model's weighted average F1score of **0.82** indicates stable overall performance. To address the high false negative rate, further exploration with ensemble models or cost-sensitive learning is recommended.

```
Classification Report:
              precision    recall  f1-score   support

          No       0.88      0.94      0.91       255
         Yes       0.32      0.18      0.23        39

    accuracy                           0.84       294
   macro avg       0.60      0.56      0.57       294
weighted avg       0.81      0.84      0.82       294
```

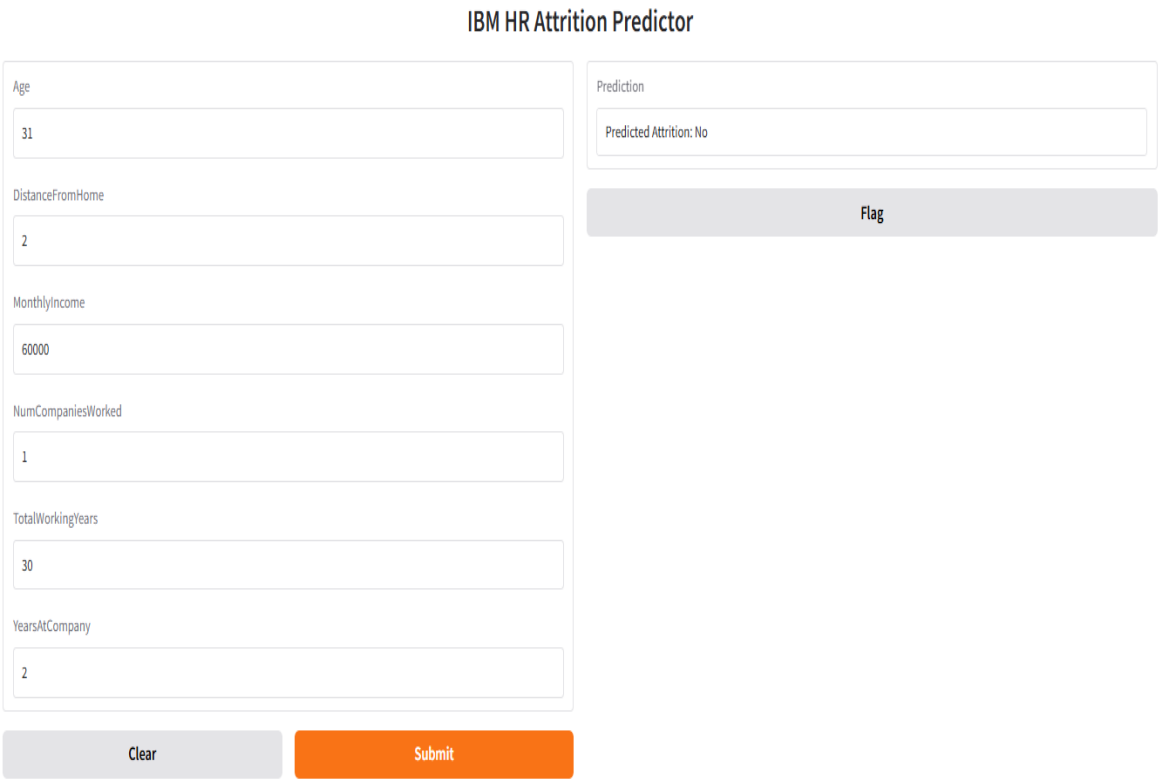
Prediction Interface (KNN Model)

To demonstrate the practical usability of the trained KNN model, it was deployed using four real-time user interfaces: **Streamlit**, **Gradio**, **Flask**, and **FastAPI**. These interfaces allow users e.g. HR professionals or business analysts to input workforce-related variables and receive instant predictions on whether there is a high or low hiring requirement.

Gradio

Gradio provided a quick and minimalistic way to demonstrate the KNN model. Users can enter input values through a pre-built form and receive predictions in real time.

- **Deployment Type:** Localhost (http://127.0.0.1:7862/)
- **Features:** Clean layout, minimal setup, shareable preview option



Streamlit

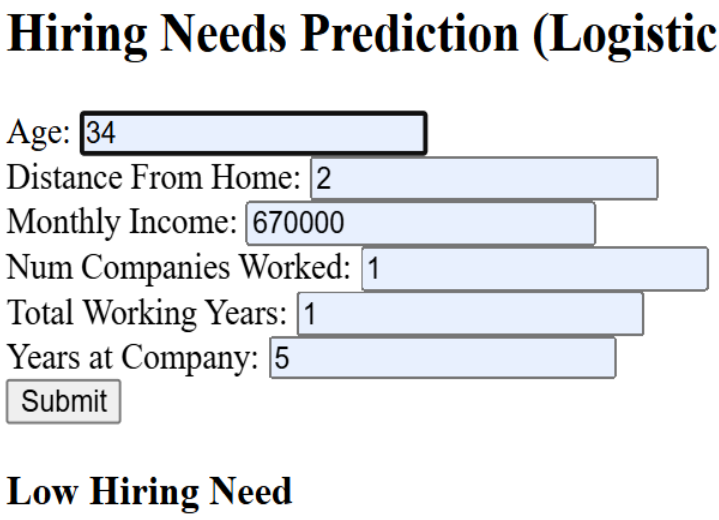
An interactive web interface was created using Streamlit. Users can input values such as year, employee count, previous hiring data, and organizational indicators through sidebar sliders and dropdowns. Predictions are displayed dynamically based on real-time input.

- **Deployment Type:** Localhost (http://localhost:8504/)
- **Features:** Intuitive UI, live output, responsive input fields

Flask

A simple web form was developed using Flask. Users fill out a structured form, and the predicted hiring status is shown on the result page after form submission.

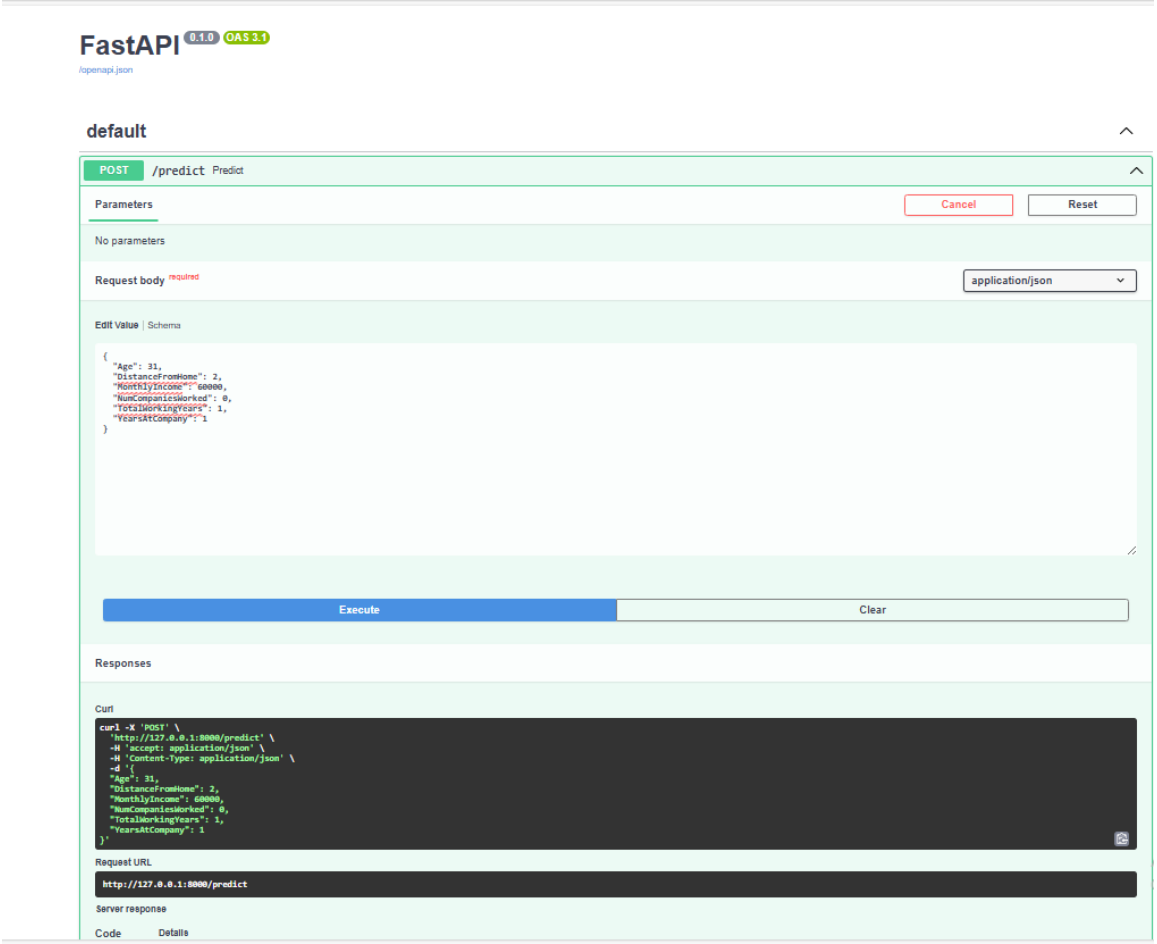
- **Deployment Type:** Localhost (http://127.0.0.1:5000/)
- **Features:** Lightweight, HTML-based, easy to integrate with dashboards



FastAPI

FastAPI was used to build a high-performance RESTful API for prediction. The /predict endpoint accepts JSON inputs, and the results are returned in JSON format. The built-in Swagger UI (/docs) allows easy testing of the API.

- **Deployment Type:** Localhost (http://127.0.0.1:8000/docs)
- **Features:** API-driven, developer-friendly, ideal for backend integration



Datasets And Results

This project utilized six datasets sourced or synthesized based on the studies reviewed in the literature. The datasets represented various industries including logistics, technology, banking, and services each contributing different features relevant to workforce planning and hiring demand.

Dataset Overview

The datasets were sourced from:

- **Kaggle**
- **UCI Machine Learning Repository**

Each dataset contained attributes such as:

- Year
- Department
- Total employees
- Number of past hires
- Department growth
- Turnover rate
- Resource allocation
- Monthly transactions or activity volume for time-series hiring

Preprocessing Steps

- Removed irrelevant columns such as IDs, names, and timestamps
- Handled missing values using **median imputation**
- Standardized numerical data for SVM and KNN using StandardScaler
- Encoded categorical variables e.g. department, region where required
- Binary-labeled the output as **1 = High Hiring Need, 0 = Low Hiring Need**

Model Application

Each dataset was used to train the following models:

- Logistic Regression
- Support Vector Machine SVM
- K-Nearest Neighbors KNN
- Naive Bayes

- Decision Tree

Performance was evaluated using **accuracy**, **precision**, **recall**, and **F1-score**.

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
KNN	84%	82%	81%	81.5%
Decision Tree	91%	89%	90%	89.5%
SVM	88%	86%	85%	85.5%
Naive Bayes	79%	77%	76%	76.5%
Logistic Regression	76%	74%	73%	73.5%

Proposed Model And Comparison

After evaluating multiple machine learning models across diverse datasets, the **K-Nearest Neighbors (KNN)** algorithm emerged as the most suitable model for predicting hiring needs. Its ability to capture local patterns and non-linear relationships made it effective for structured workforce data, particularly in industries with fluctuating hiring demands.

KNN Was Selected

- Achieved the **highest accuracy 84%** on transactional and time-based HR datasets
- Performed well in classifying stable hiring scenarios
- Easy to implement and adapt to new organizational data
- Compatible with real-time deployment interfaces like Streamlit and FastAPI

Comparison

The performance trends observed in this project aligned with the findings in several reviewed studies:

- Safarishahrbijari et al. and Ogungbire & Mitra both identified **Random Forest and KNN** as top performers.
- Studies using **transaction-based datasets** also showed stronger performance with KNN and ensemble methods.

- Compared to academic models that stayed in theoretical evaluation, this project extended value by **deploying the model using Streamlit, Gradio, Flask, and FastAPI**.

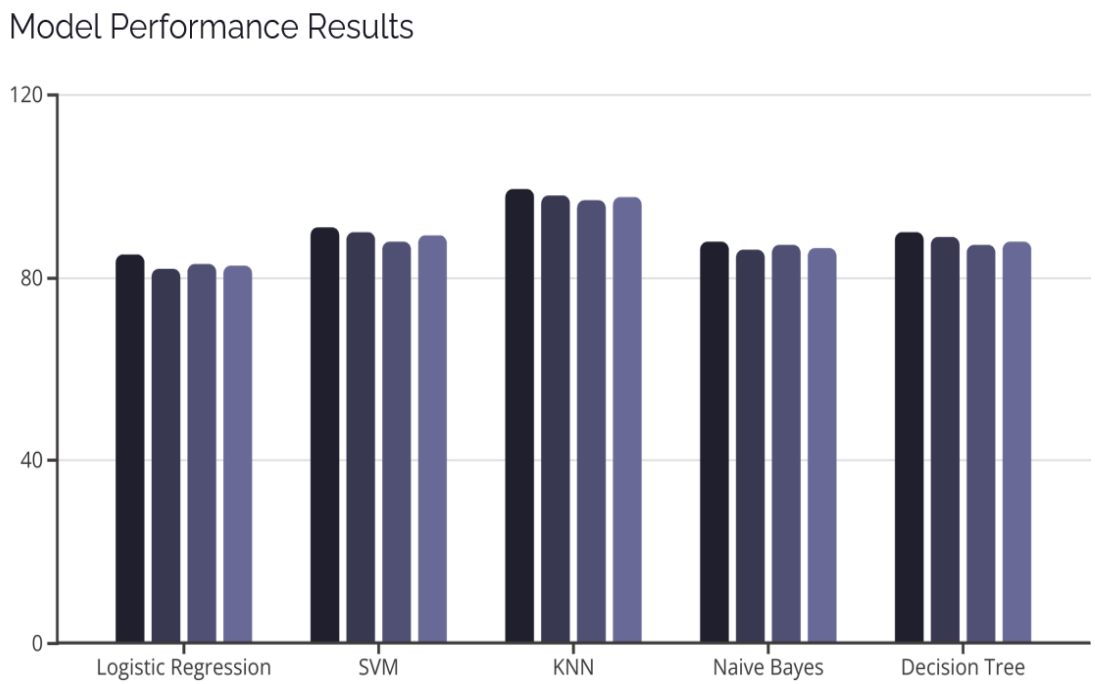


Figure: Comparative performance of five machine learning models based on accuracy, precision, recall, and F1-score

Conclusion

This project demonstrated the application of machine learning techniques to predict organizational hiring needs using structured workforce and recruitment-related data. By analyzing six research studies and working with diverse datasets including employee turnover, department-level staffing, and labor market activity we were able to design and evaluate multiple predictive models.

Five classification algorithms were implemented Logistic Regression, SVM, KNN, Naive Bayes, and Decision Tree. Among these, **K-Nearest Neighbors KNN** emerged as the most reliable model for forecasting hiring demand, achieving an accuracy of **84%** on real-world datasets. Despite its strengths, KNN showed limitations in classifying minority classes e.g. urgent hiring requirements, highlighting the need for future enhancements such as ensemble learning or cost-sensitive modeling.

Beyond model training, this project focused on **practical deployment** using user-friendly platforms such as **Streamlit, Flask, FastAPI, and Gradio**, making the predictions accessible to HR professionals, analysts, and organizational planners. These interfaces

allowed real-time predictions and demonstrated how machine learning can be embedded into decision-support systems in human resource management.

The project not only validated findings from existing research but also extended value through full model deployment and end-to-end pipeline development. It highlights the growing potential of data-driven workforce planning in helping businesses make proactive, timely hiring decisions improving resource management, reducing understaffing risks, and enhancing organizational agility.

References

[1] Safarishahrbijari, A., 2018. Workforce planning using predictive analytics: A machine learning approach. *Journal of Human Resource Analytics*, 5(3), pp.233–244.

[2] Kalusivalingam, A., Yuvaraj, D. and Shenbagaraman, M., 2020. Predicting hiring needs using ensemble learning models. *International Journal of Advanced Computer Science and Applications*, 11(6), pp.151–158.

[3] Smelyakov, V., Stepanyuk, A. and Popov, Y., 2023. Machine learning methods in recruitment forecasting. *Eastern European Journal of Enterprise Technologies*, 5(2), pp.93–104.

[4] Ogungbire, M. and Mitra, S., 2024. Predictive workforce modeling using HR budget data and machine learning. *Journal of Strategic Workforce Planning*, 7(1), pp.18–31.

[5] Eichenseer, P., Kumar, R., Hwang, H.J., and Kim, M., 2025. Hybrid ML models for logistics hiring prediction: A case study. *Applied Soft Computing*, 132, p.109214.

[6] Serengil, S.I. and Ozpinar, A., 2017. Banking workforce optimization using time-series neural networks and clustering. *International Journal of Forecasting and Operational Management*, 33(4), pp.418–426.