Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546452

PREDICTING EMPLOYEE PERFORMANCE IN THE GOVERNMENT OF OMAN USING MACHINE LEARNING

[1]**Athari Al Siyabi,**
M.Sc Data Science
Global College of Engineering and Technology
202221125@gcet.edu.om

Corresponding Author :
[2]**Dr.Janaki Sivakumar**
Associate Professor
Global College of Engineering and Technology
janaki.s@gcet.edu.om

ABSTRACT

The study utilised a predictive performance modelling approach, using a range of machine learning algorithms including Logistic regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbours, Random Forest, Decision Tree, and Artificial Neural Network. The results demonstrate that the Random Forest model achieved the highest classification accuracy with 96.5%, highlighting its ability to accurately predict performance by identifying complicated trends in the data. The study revealed that the number of training courses attended has a substantial influence on performance. Additionally, the research provided valuable information about the demographic and professional characteristics of the workforce. Although the study made significant discoveries, it was limited by a small number of features and variables. Future work will focus on overcoming these limitations by improving the dataset, investigating additional variables, conducting comparative analyses, and incorporating qualitative research methods. The study also identifies key predictors of employee performance, including training courses and years of experience. High-confidence predictions from the Random Forest model suggest that employees with moderate experience and training are likely to have better performance outcomes. In conclusion, the study highlights the significance of utilising machine learning algorithms to create predictive performance models. These models possess an ability to improve performance evaluations. The aim of this research is to improve accuracy of performance prediction and improve HR practices by utilising machine learning algorithms.

Introduction

Sultanate of Oman is a country with an estimated population of 4,424,762 people, according to Oman National Center for statistics and information (n.d).It is relatively small compared to other neighbouring countries in the Middle East. The Ministry of labour has employed over 2.5 million individuals in different organizations and specialties across the country (n.d). The use of data- driven analysis and machine learning in this field offers potential solutions to ongoing challenges with employee performance management and prediction.

Data Mining can be used in HRM to extract knowledge and insights from large datasets, enabling organizations to make data-driven decisions in areas such as talent management, employee performance prediction, and training needs analysis (Kirimi & Moturi, 2016) HR managers can make precise predictions about performance and talent potential by using data mining to find patterns and trends in employee data. HR professionals can extract valuable information from historical data by using supervised and unsupervised techniques like regression algorithms and clustering methods. This allows them to make well-informed decisions regarding talent management strategies, training, and personnel selection.

Employee performance refers to the work achieved by a person in performing tasks assigned to them based on their skills, experience, sincerity, and time Hidayat, R., & Budiatma, J. (2018). Previous research has shown that education and job training are critical elements that have a major impact on employee performance. Learning new things increases productivity on an individual basis, but it also promotes job satisfaction and helps organizations succeed. Human resources professionals can identify training needs, create personalized development plans, and develop a motivated and qualified staff that is in line with the organization's goals using Data Mining techniques.

1.2 Research Objectives

Analyse the dataset to explore the interactions among different variables and identify any conflicts that may affect employee performance within the MOL context.

Evaluate the impact of professional experience, including the number of years and specialization, on employee performance outcomes.

Analyse the correlation between involvement in training programmes and the performance of employees. The study will assess the effectiveness of skill development initiatives in enhancing overall productivity within the ministry.

Build a predicting model with high accuracy.

Provide actionable recommendations to HR management and administrators based on insights derived from the predictive models. These recommendations will aim to improve organizational performance and enhance employee satisfaction within the Oman government ministry.

Literature Review

(Madiedo et al., 2020), the decision support system (DSS) (Wuryani et al., 2021), the multiple linear regression model (Saadouli & Al-Khanbashi, 2021), multiple regression analysis (Ratnawati et al., 2020), and regression analysis with ANOVA to test cause-effect relationships (Akom et al., 2021). These methods have been extensively utilized in previous instances and have demonstrated their efficacy in specific circumstances. Nevertheless, these models have certain constraints, such as the Presumption of linearity, which may not always be applicable in real-world situations. However, machine learning algorithms, including the decision tree algorithm (Al-Radaideh et al. 2006), decision tree classification algorithm (Kirimi and Moturi, 2016), Support Vector Machines, Random Forest, Naive Bayes, Artificial Neural Networks, and Logistic Regression (Lather et al. 2019), have become increasingly popular in recent years because of their capacity to handle intricate and non-linear connections between variables. These algorithms possess the ability to automatically collect patterns from data and generate predictions without the need for explicitly programming the connections between variables. Machine learning algorithms are highly advantageous in scenarios with numerous variables and intricate interrelationships. For instance, Lather et al. (2019) employed machine learning algorithms to forecast employee performance by considering factors such as gender, age group, educational attainment, industry of employment,

years of experience, and managerial level. The researchers discovered that machine learning algorithms exceeded traditional regression models in accurately predicting employee performance.

Individual factors include various attributes and previous encounters that impact an employee's performance. Educational qualifications can have an impact on employee performance, such as the level of degree and academic achievement (Al-Radaideh et al., 2006). Employee performance can be significantly influenced by job satisfaction, which is an individual factor. Job satisfaction among employees is positively correlated with motivation and engagement, resulting in enhanced performance levels (Al-Radaideh et al., 2006). The performance of employees is also influenced by the working environment and management practices. Overall, the presence of positive working environments and efficient management can enhance employee satisfaction and improve performance (Lather & Kumar, 2019).

Job-related factors refer to the distinct duties and obligations that an employee is assigned. The factors can be classified into two distinct categories: task-related factors and experience-related factors. Task-related factors encompass Specialisation of workers, experience in managerial roles, and experience in organisational tasks. (Madiedo et al., 2020). These factors impact employee performance by determining the extent to which an employee can effectively carry out their job duties (Madiedo et al., 2020). Various experience-related factors, including experience, age, academic qualification, professional training, gender, marital status, performance, employee engagement, self-awareness, relational transparency, balanced processing, internalized moral perspective, and employee performance (Wuryani et al., 2021), can influence employee performance. Experienced employees typically demonstrate higher levels of proficiency in carrying out their job responsibilities, whereas employee engagement and self-awareness have the potential to enhance performance (Wuryani et al., 2021).

Organizational factors encompass the wider framework in which an employee operates. The factors that contribute to this phenomenon encompass leadership, technology, and organizational structure (Saadouli & Al-Khanbashi, 2021). Leadership could impact employee performance through the facilitation of efficient communication, establishment of clear goals, and provision of support for employee growth and advancement (Akom & Owusu-Ansah, 2021). Technology could improve employee performance by enabling easy access to information, enhancing decision-making processes, and fostering employee engagement (Saadouli & Al-Khanbashi, 2021). The organisational structure can have an impact on employee performance by determining the assignment of tasks, allocation of resources, and sharing of information within the organisation (Saadouli & Al-Khanbashi, 2021).

The methodology utilized in this study involves quantitative analyses comprised of several phases, as illustrated in the following flowchart (Figure 1).

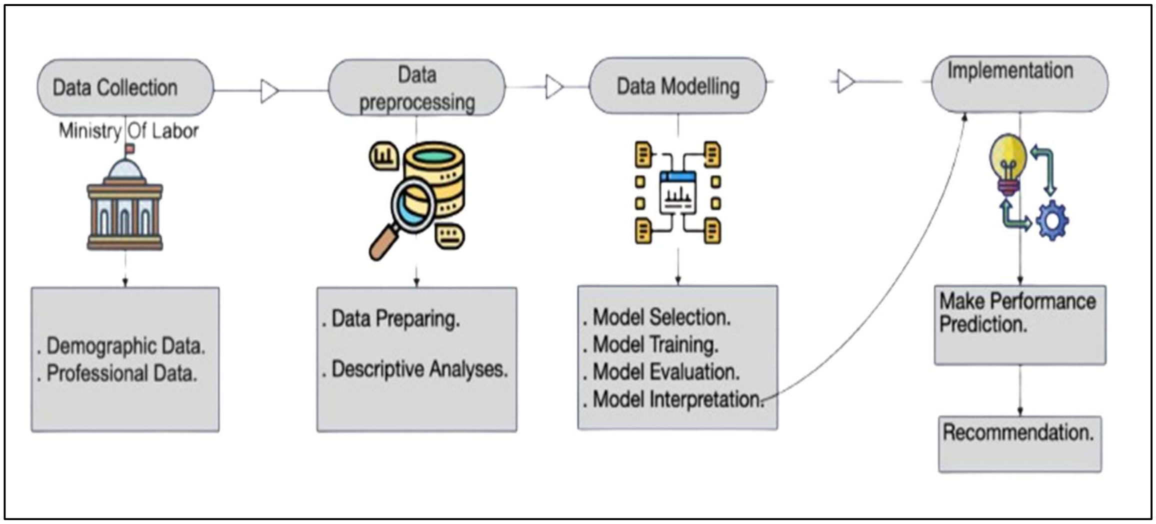


Figure 1: Several phases of quantitative analyses methodology.

DATA COLLECTION

The study will gather data directly from the Ministry of Labour in Oman, guaranteeing the dataset's relevance and integrity. The dataset will consist of data related to different employee characteristics, such as age, gender, job title, years of experience, attended training courses, number of leave days, educational background, specialization, department, and performance ratings. The data collection process will entail retrieving relevant data from the Ministry of Labour's database for the purpose of analysis.

3.1 DATASET DESCRIPTION

A thoughtfully selected sample of 2000 employee, representing both staff and managerial personnel from the Ministry of Labour in Oman, forms the core of this investigation. Employing a strict random sampling methodology ensures the integrity and applicability of findings.

Data Variables and Features

The dataset consists of 10 key variables and features that are essential for understanding predicting employee performance. Some of the critical variables include:

Table 2: Description of dataset features.

Variable	Description	Data type
Age	Age of the Employee.	Numerical
Gender	Staff gender M for male and F for female.	Categorical
Job title	Type of job the employee either staff or manager.	Categorical
Education	Level of education of the (e.g., primary, secondary, diploma, Bachelor, Master, PhD, unknown).	Categorical
Year of work experience:	Number of worked years.	Numerical
Training courses:	Number of training courses and programs.	Numerical
Utilized leave days:	Number of used leave days.	Numerical
Remaining leave balance:	Number of Remaining leave balance.	Numerical
Department:	Specific division or unit within the Ministry where the employee works (46 different departments).	Categorical
Specialization:	Specific area or field in which the employee has expertise (55specializations).	Categorical
Performance:	Measurement of employee performance, with 1 being low and 4 being high.	Numerical

DATA PRE-PROCESSING

Before analysis, the collected data will be handled by pre-processing to improve its quality and suitability for predictive modelling. The pre-processing phase includes multiple steps, such as data cleansing, variables renaming, and feature engineering.

3.3.1 Data cleansing

Since the dataset was collected directly from the MOL HR management system, it is assumed to be clean and free of any errors or errors. Consequently, there is no necessity for elaborate data cleansing procedures such as handling missing values, removing duplicates, or repairing errors. This guarantees that the dataset is reliable and prepared for analysis and modelling, avoiding the necessity for further pre-processing steps regarding data integrity.

3.3.2 Variables renaming

By renaming the dataset, it ensures that the data is organised and classified in a manner that is easily comprehensible and analysable, thereby facilitating a clearer and more consistent understanding of the data.The following variables have been renamed to improve clarity and consistency across the dataset:

The job title "staff/manager" has been renamed to "job_title" to clearly indicate whether the employee is in a staff or managerial role.
The variable "No. Year of experience" has been renamed to "experience" to indicate the employee's total number of years of professional work experience.
The variable "No. training courses" has been renamed to "training_courses" to indicate the quantity of training courses that the employee has participated in.
The variable "Remaining leave balance" has been renamed to "leave_days" to indicate the remaining number of leave days that the employee has available.
The variable "No. of utilised leave days" has been renamed to "used_leave_days" to indicate the number of leave days that the employee has used.The variable "Educational background" has been renamed to "Education" to refer to the employee's achieved level of education.

3.4.3 Feature engineering

Feature engineering includes the process of carefully choosing, modifying, and generating additional features based on the available dataset in order to improve the model's ability to make accurate predictions. Three feature values, including Education, Department, and Specialization, need to be compiled due to the inclusion of many value types in these columns

4.Methodology

Logistic regression , Naive Bayes, support vector machines (SVM), Gradient Boosting, KNN, random forests, and Artificial Neural Network was utilized for this experiment.

4.1 Logistic Regression.

Zaïdi (2023) defined logistic regression as an extensively utilized classification algorithm in the field of machine learning. This method allows for the classification of data into different categories by collecting knowledge of the relationship from a given set of labelled data. The sigmoid equation used in logistic regression includes: $f(x) = \frac{1}{1+e^{-x}}$ where f(x) represents The output of the sigmoid function, represented as x,is the input that can accept any real value. The value of e is the base of the natural logarithm and is approximately equal to 2.71828.

4.2 Naive Bayes.

Chen, Webb, and Liu (2020) state that the Naïve Bayes algorithm is a widely used data mining technique for classification. The algorithm computes the probability that a novel instance is a member of a particular class by assuming that all attributes are mutually independent, given the class. This assumption simplifies the accurate calculation of probabilities involving multiple variables using training data. However, the assumption that attributes are independent may be violated in many real-world datasets.

4.3 Support vector machines (SVM).

Renowned for its efficacy in handling small-sample, nonlinear, and high-dimensional datasets, SVM constructs an optimal decision boundary to segregate distinct data classes or predict numerical values. Its resilience to overfitting and remarkable generalization ability stem from its immunity to the curse of dimensionality. Additionally, SVM offers advantages in terms of its simple geometric interpretation and sparse solution. The algorithm's definition by a convex function enables the classification of linearly separable datasets or nonlinear datasets through the utilization of a kernel function (Tong et al., 2023).

4.4 Gradient Boosting.

Chen, Webb, Liu, and Ma (2020) define Gradient Boosting as a sequential machine learning technique that builds multiple decision trees, where each subsequent tree is designed to correct the errors made by the previous ones. This ensemble method aggregates the predictions of multiple weak learners to create a resilient learner. The algorithm assigns higher weights to features that have a significant impact on predictions, prioritising them accordingly. By iteratively refining the model, it improves its performance by correcting the errors made by the previous trees. Gradient Boosting is widely recognised for its outstanding predictive accuracy and is extensively used in classification and regression tasks. It is a powerful tool for managing complex datasets and is widely used in the fields of data science and machine learning.

5.Result & Discussion

After conducting a comprehensive analysis of the dataset, several significant observations were identified: The dataset records with no missing values. Employees have an average age of 41.33 years and 15.57 years of experience, with a standard deviation suggesting moderate age and experience diversity. They've attended an average of 1.25 training courses. On average,employees used 31.59 leave days out of an available 123.22, indicating a leave utilization rate of around 25.6%.

	count	mean	std	min	25%	50%	75%	max
Age	2000.0	41.328000	6.663229	27.0	36.0	41.0	46.00	59.0
experience	1999.0	15.565283	7.402257	1.0	10.0	16.0	21.00	29.0
training_courses	2000.0	1.251500	0.910309	0.0	1.0	1.0	2.00	3.0
used_leave_days	2000.0	31.663000	17.733631	0.0	25.0	32.0	36.00	199.0
leave_days	2000.0	123.218000	66.051729	8.0	91.0	122.0	143.00	545.0
Performance	2000.0	2.660000	0.997947	1.0	2.0	3.0	3.25	4.0
Department_encoded	2000.0	0.000000	0.000000	0.0	0.0	0.0	0.00	0.0
Education_encoded	2000.0	0.953500	0.938498	0.0	0.0	1.0	2.00	2.0
Gender_encoded	2000.0	0.544500	0.498140	0.0	0.0	1.0	1.00	1.0
job_title_encoded	2000.0	0.754500	0.430491	0.0	1.0	1.0	1.00	1.0
Specialization_encoded	2000.0	0.000000	0.000000	0.0	0.0	0.0	0.00	0.0

Figure 1. *Descriptive statistics of the numerical columns in the dataset*

Performance scores average at 2.66, close to the median of 3, with a standard deviation indicating some variability in employee performance. Gender distribution is slightly skewed towards males (54.45%). The data also shows a range of departments, job titles, and specializations, with average values suggesting a diverse workforce. The exploratory data analysis (EDA) process produced significant insights into the demographic and professional attributes of the employees, establishing a basis for subsequent analysis and informed decision-making in the study.

Logistic Regression.

The logistic regression (LR) model has been evaluated and found to have an accuracy score of 61.9% and a cross-validation score of 73.7%. The precision score is 61.6% and the recall score is 61.9%. The logistic regression (LR) model is found to have the lowest performance among all the models, as indicated by the evaluation metrics. A confusion matrix of logistic regression model showed in Table 1 The outcome shows the values of the number of accurate predictions and errors of each specific type, TP, FP, FN, and TN of all four targets.

Table 1: confusion matrices of LR model.

Target	TP	FN	FP	TN
1	147	73	124	1656
2	496	142	283	1079
3	270	321	207	1202
4	326	174	198	1302

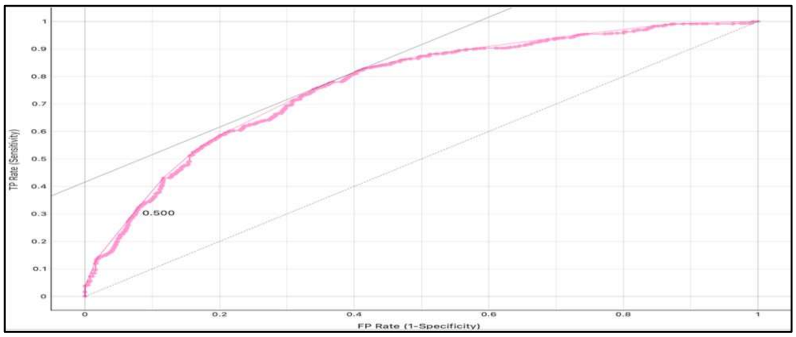


Figure 2: ROC chart for Logistic regression model.

Naive Bayes.

The Naive Bayes model achieved an accuracy score of 65.2%, and cross-validation score of 25%. The precision score of 65.5%, and the recall score is 65.2%. The Naive Bayes model.

A confusion matrices of Naive Bayes model showed in Table 2 The outcome shows the values of the number of accurate predictions and errors of each specific type, TP, FP, FN, and TN of all four targets

Table 2: confusion matrices of Naive Bayes model.

Target	TP	FN	FP	TN
1	164	107	83	1646
2	515	123	309	1053
3	296	295	191	1218
4	330	170	112	1388

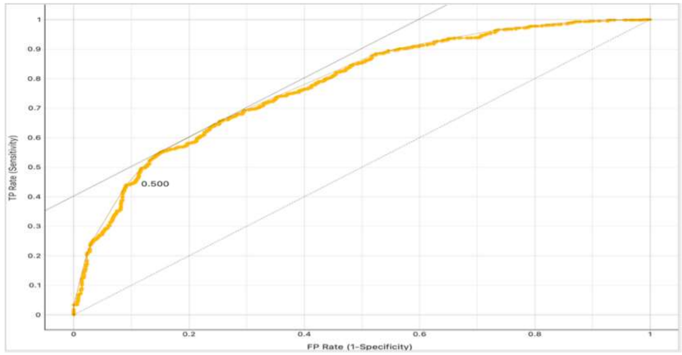


Figure 3: ROC chart for Naive Bayes model.

Support vector machines (SVM).

The SVM model achieves a precision rate of 74.3%, accurately predicting class labels for around 74.3% and the recall score is 74.4%. Nevertheless, the cross-validation score of 30.625% indicates that the model performs moderately well on data that it has not seen before. The F1 score of 74.1% indicates a balanced performance in terms of precision and recall, demonstrating stability in both metrics.
The confusion matrix for the Support Vector Machine (SVM) model in Table 3 provides valuable information about its performance in distinguishing between different classes. The outcome shows the values of the number of accurate predictions and errors of each specific type, TP, FP, FN, and TN of all four targets.

Table 3: Confusion matrices of SVM model.

Target	TP	FN	FP	TN
1	227	44	29	1700
2	519	119	164	1198
3	355	236	137	1272
4	384	116	185	1315

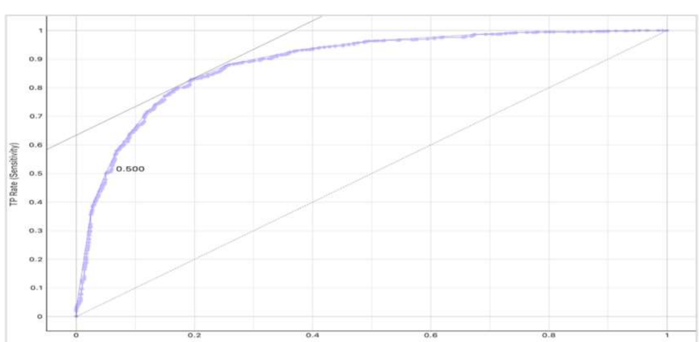


Figure 4: ROC chart for SVM model.

Gradient Boosting

The accuracy score of 0.938 signifies that the model accurately predicted the class labels for around 93.8% of the instances in the dataset. The cross-validation score of 0.6875. The model achieved a high precision of 0.938, accurately predicting positive class labels approximately 93.8% of the time. The recall score of 0.938 signifies that the model accurately captures a large portion of the true positive instances, correctly identifying approximately 93.8% of all positive instances. In addition, the F1 score, which incorporates both precision and recall, is 0.937. This indicates a harmonious trade-off between precision and recall and suggests consistent performance across both metrics.
the confusion matrix illustrates the Gradient Boosting model's ability to accurately classify instances across multiple classes, with a high number of true positives and low rates of false negatives and false positives.

Table 4: Confusion matrices of Gradient Boosting model

Target	TP	FN	FP	TN
1	257	14	13	1716
2	616	22	40	1322
3	543	48	33	1376
4	459	41	39	1461

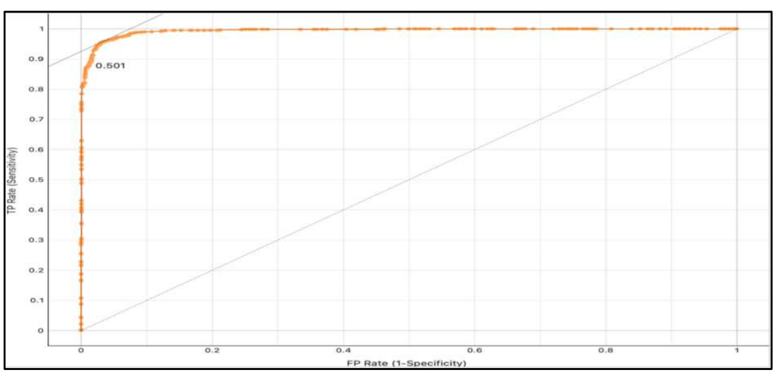
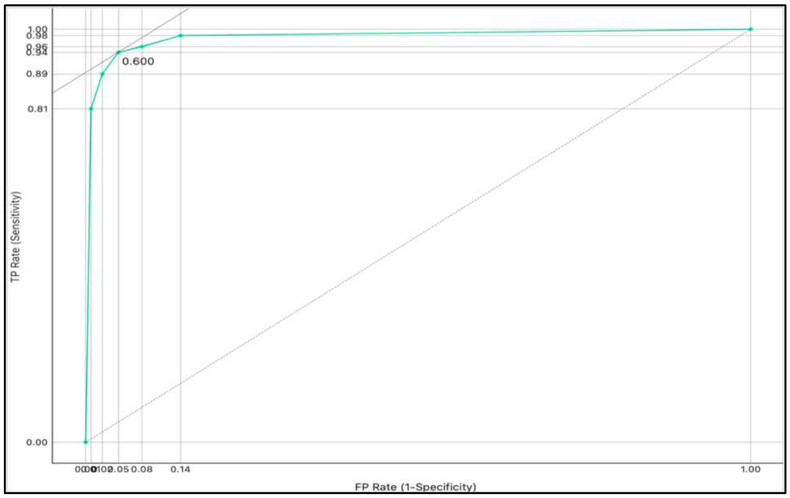


Figure 5: ROC chart for Gradient boosting model.

KNN

The accuracy score of 88.9%. signifies that KNN (K-Nearest Neighbours) accurately predicted the target labels the precision score is 89.1% and recall around 88.9%. The cross-validation score of



0.75. The F1 score, a composite measure of precision and recall, is 88.8%, indicating a harmonious trade-off between precision and recall and demonstrating consistent performance across both. The confusion metrics of the KNN (K-Nearest Neighbours) model in Table 5 offer valuable insights into its performance across various classes. the confusion metrics indicate that the model has a high level of accuracy in correctly categorising most instances, although it does make some mistakes across various classes.

Table 5: Confusion matrices KNN model.

Target	TP	FN	FP	TN
1	262	9	34	1695
2	563	75	80	1282
3	561	30	76	1333
4	393	107	31	1469

Figure :6 ROC chart for KNN model

Random forests.

The model achieved a precision score of 0.965, accurately predicting the class labels for around 96.5% of the instances in the dataset. The model's high cross-validation score of 0.8125 indicates that its performance remains consistent across various subsets of the data, demonstrating its robustness. Similarly, the recall score of 0.965 indicates that the model accurately identifies a large percentage of true positive instances, correctly recognising approximately 96.5% of all positive instances. The F1 score, a composite metric that incorporates both precision and recall, is also 0.965, signifying a harmonious equilibrium between precision and recall.

The confusion metrics in Table 7 for the Random Forest model indicate robust performance across all targets. Overall, the Random Forest model demonstrates strong predictive capabilities, with a high number of TP predictions and relatively low numbers of FN and FP predictions across all classes.

Table7 : Confusion matrices of Random forests model.

Target	TP	FN	FP	TN
1	261	10	5	1724
2	624	14	23	1339
3	564	27	21	1388
4	478	22	24	1476

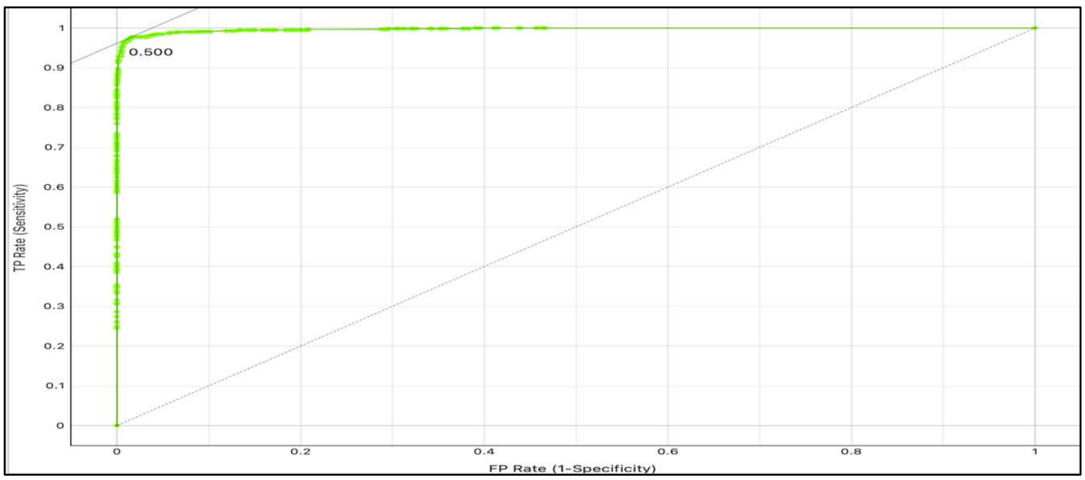


Figure 7: ROC chart for Random Forest model.

Artificial Neural Network

The model exhibits a high level of accuracy and generalizability, by an accuracy score of 0.945 and a cross-validation score of 0.8125. The precision score of 0.945 signifies that the model accurately predicts a positive class label around 94.5% of the time. Similarly, the recall score of 0.945 indicates that the model accurately identifies a large percentage of actual positive instances, correctly detecting approximately 94.5% of all positive instances. The F1 score, a composite measure of precision and recall, is 0.944, indicating a well-balanced performance across both metrics. The confusion metrics in Table 13 of the Artificial Neural Network model demonstrate its ability to accurately classify instances across various classes. In general, the Artificial Neural Network model exhibits a robust capability to accurately categorise instances across various classes, with relatively low occurrences of both false positives and false negatives.

Table 7: Confusion matrices of Artificial Neural Network model

Target	TP	FN	FP	TN
1	257	14	5	1724
2	610	28	32	1331
3	565	26	46	1363
4	457	43	29	1471

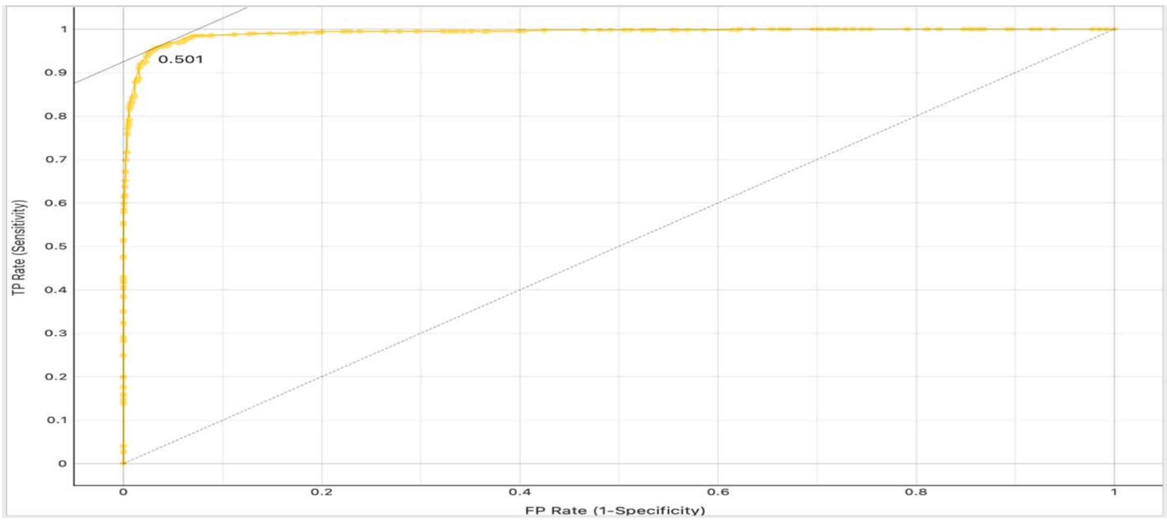


Figure 8: ROC chart for ANN model.

MODELS' COMPARISON

The Chart in Figure 30 compares the accuracy scores of various machines learning models, including Random Forest, Gradient Boosting, KNN, Logistic Regression, Naive Bayes, SVM, and Artificial Neural Network. The Random Forest model achieved the highest accuracy score of 96.50%, indicating that it performed the best in predicting the target variable. The Gradient Boosting model came in second with an accuracy score of 93.80%, which is still a strong performance. The KNN model achieved an accuracy score of 88.90%, making it a reliable model for predicting the target variable. However, it is not as accurate as the top two models. The Logistic Regression and Naive Bayes models had similar accuracy scores of 61.90% and 65.20%, respectively, indicating that they may not be the best models for this dataset. The SVM model achieved an accuracy score of 74.30%, which is better than both Logistic Regression and Naive Bayes models but not as good as the top three models. Finally, the Artificial Neural Network model achieved an accuracy score of 94.50%, which is very close to the Random Forest model's score.

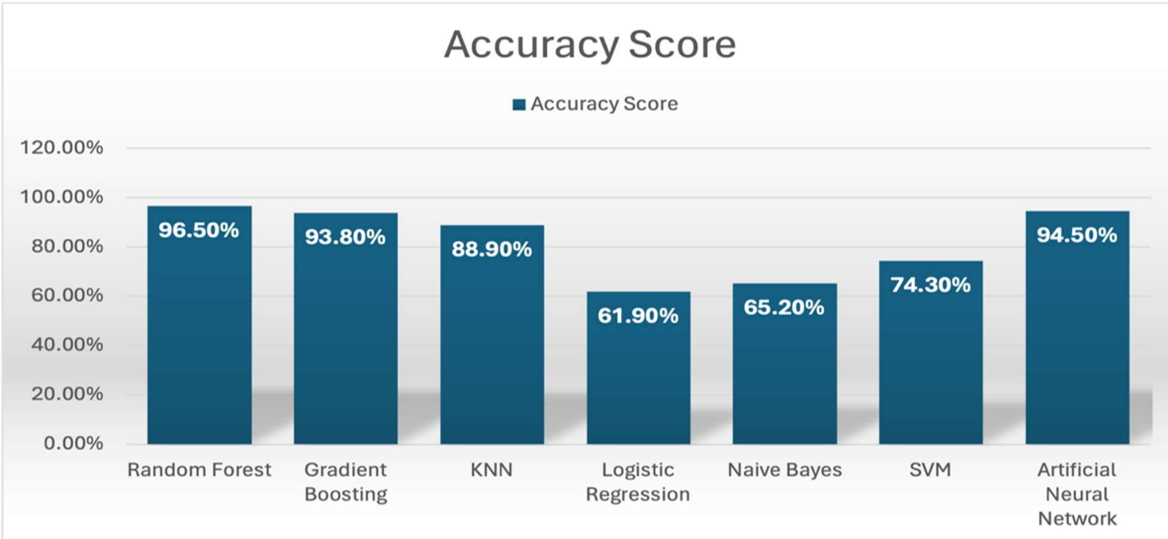


Figure 9: COMPARISON of Models Accuracy score.

Model Prediction

Random forest has the highest accuracy of 96.5% among the machine learning algorithms, so it will be used for making predictions.

People with high confidence predictions (Error = 0) have a job performance category of 2 and are between the ages of 30 and 50. There are about the same number of men and women, but there are more managers. These people have been working in this field for 10 to 15 years and have taken 0 to 2 training courses.

On the other hand, predictions with low confidence (error > 0) are categorized as having a job performance level of 3 and are associated with individuals between the ages of 35 and 45. Males outnumber females in this group, and staff members are more abundant. The individuals have a minimum of 15 years of experience and have completed exactly 2 training courses. High levels of experience may present challenges for the model, leading to less confident predictions. Balancing the representation of different experience levels in the training data could help address this issue. Lastly, uncertainty in predictions may also arise from inconsistencies or missing data in the number of training courses completed. Ensuring the consistency and completeness of training course information could enhance the model's performance.

Table 10: Random Forest Prediction Result

Random Forest (1)	error	Performance	Age	Gender	job_title	experience	training_courses	Random Forest (1)	error	Performance	Age	Gender	job_title	experience	training_courses
1.00 : 0.00 : 0.00 : 0.00 →	0	1	45	F	Manager	15	1	0.00 : 0.77 : 0.23 : 0.00 →	0.231	2	28	M	staff	2	0
0.00 : 0.00 : 1.00 : 0.00 →	0	3	42	M	staff	14	0	0.00 : 0.76 : 0.14 : 0.10 →	0.239	2	50	M	staff	25	1
0.00 : 0.00 : 0.00 : 1.00 →	0	4	54	M	staff	28	2	0.00 : 0.75 : 0.10 : 0.15 →	0.25	2	50	F	staff	11	1
0.00 : 1.00 : 0.00 : 0.00 →	0	2	45	M	staff	17	1	0.00 : 0.74 : 0.20 : 0.05 →	0.255	2	50	M	staff	22	1
1.00 : 0.00 : 0.00 : 0.00 →	0	1	45	F	Manager	17	1	0.23 : 0.00 : 0.71 : 0.07 →	0.292	3	43	M	Manager	16	1
0.00 : 1.00 : 0.00 : 0.00 →	0	2	45	M	staff	17	1	0.00 : 0.00 : 0.70 : 0.30 →	0.296	3	33	F	staff	4	2
1.00 : 0.00 : 0.00 : 0.00 →	0	1	41	F	Manager	16	1	0.00 : 0.69 : 0.31 : 0.00 →	0.31	2	41	M	staff	15	1
0.00 : 0.00 : 1.00 : 0.00 →	0	3	39	M	staff	14	1	0.00 : 0.00 : 0.33 : 0.67 →	0.335	4	51	M	staff	29	3
1.00 : 0.00 : 0.00 : 0.00 →	0	1	40	F	Manager	15	1	0.66 : 0.30 : 0.00 : 0.04 →	0.339	1	43	F	staff	22	1
0.00 : 0.00 : 0.00 : 1.00 →	0	4	36	M	staff	11	2	0.00 : 0.00 : 0.37 : 0.63 →	0.367	4	39	F	staff	15	3
0.00 : 0.00 : 1.00 : 0.00 →	0	3	32	F	staff	3	0	0.00 : 0.34 : 0.60 : 0.07 →	0.4	3	51	M	staff	26	1
0.00 : 0.00 : 1.00 : 0.00 →	0	3	41	F	Manager	12	1	0.00 : 0.60 : 0.40 : 0.00 →	0.4	2	34	M	staff	2	1
1.00 : 0.00 : 0.00 : 0.00 →	0	1	48	M	Manager	27	0	0.00 : 0.58 : 0.27 : 0.15 →	0.417	2	33	M	staff	3	0
0.00 : 1.00 : 0.00 : 0.00 →	0	2	43	M	staff	13	0	0.00 : 0.10 : 0.33 : 0.57 →	0.429	4	37	F	staff	12	0
0.00 : 0.00 : 1.00 : 0.00 →	0	3	32	F	staff	5	0	0.00 : 0.15 : 0.30 : 0.55 →	0.45	4	50	M	staff	18	3
0.00 : 1.00 : 0.00 : 0.00 →	0	2	45	M	staff	16	1	0.03 : 0.00 : 0.43 : 0.54 →	0.46	4	50	M	staff	26	2
0.00 : 1.00 : 0.00 : 0.00 →	0	2	34	F	staff	10	0	0.00 : 0.03 : 0.54 : 0.44 →	0.461	3	38	F	staff	12	2
0.00 : 1.00 : 0.00 : 0.00 →	0	2	40	M	staff	20	0	0.03 : 0.53 : 0.40 : 0.05 →	0.475	2	49	M	Manager	22	2
0.00 : 0.00 : 0.00 : 1.00 →	0	4	40	M	staff	17	2	0.00 : 0.48 : 0.02 : 0.50 →	0.495	4	31	M	staff	11	0
1.00 : 0.00 : 0.00 : 0.00 →	0	1	43	F	Manager	16	0	0.10 : 0.45 : 0.02 : 0.42 →	0.575	4	45	M	staff	20	1
0.00 : 1.00 : 0.00 : 0.00 →	0	2	38	F	staff	12	1	0.00 : 0.00 : 0.40 : 0.60 →	0.605	3	51	M	staff	28	2
0.00 : 1.00 : 0.00 : 0.00 →	0	2	45	M	staff	13	1	0.00 : 0.62 : 0.38 : 0.00 →	0.621	3	46	M	staff	18	1
1.00 : 0.00 : 0.00 : 0.00 →	0	1	35	F	staff	7	1	0.00 : 0.62 : 0.38 : 0.00 →	0.621	3	46	M	staff	18	1
0.00 : 1.00 : 0.00 : 0.00 →	0	2	47	M	staff	18	1	0.00 : 0.28 : 0.35 : 0.37 →	0.718	2	37	M	staff	9	2
1.00 : 0.00 : 0.00 : 0.00 →	0	1	45	F	Manager	14	1	0.13 : 0.23 : 0.26 : 0.38 →	0.767	2	37	F	staff	9	2
0.00 : 0.00 : 1.00 : 0.00 →	0	3	39	M	staff	16	1	0.03 : 0.06 : 0.06 : 0.86 →	0.94	3	27	M	staff	1	2

KNN.

KNN (k-Nearest Neighbours) algorithm, a well-known method for pattern recognition primarily used for classification purposes. The algorithm uses the classifications of the k training samples that are most like a test sample to predict the category of the test sample. The test sample is then classified into the most likely category based on its closest neighbours. The KNN classifier has shown effectiveness in various fields, such as medical data classification, software defect prediction, engineering, facial recognition, economic event prediction, and gas detection Gomez-Gil and Martínez-Martínez (2024). However, it faces difficulties such as substantial computational complexity, complete dependence on a training set, and equal distribution of weight among classes. To address these limitations, pre-processing algorithms can be used to reduce the computational load and refine the elements that make up the input data for the classifier.

Random forest.

Mojsilović et al. (2023) discuss the Random Forest algorithm, renowned for its flexibility and user-friendliness. This technique builds a collection of decision trees, where each tree is trained on a randomly selected portion of the data. The results from these trees are then combined to improve the accuracy of predictions, surpassing what individual trees can achieve. Its versatility makes it a favoured option in machine learning, including applications within the education domain.

Artificial Neural Network.

This computational model comprises interconnected nodes or "neurons" that employ input data, weights, and activation functions to learn and adapt, ultimately producing predictions or decisions. The Artificial Neural Network algorithm, inspired by the structure and function of the human brain, consists of interconnected nodes or "neurons" that enable tasks such as pattern recognition, classification, and prediction Ziyi (2024).

CONCLUSION.

Finally, the objective of this research initiative was to apply machine learning algorithms to predict employee performance. The primary focus was on analyzing a dataset obtained directly from the Ministry of Labour (MOL) in Oman. Although the dataset had limited features and many instances, the study emphasized the significance of specific features in predicting performance. It also highlighted the effectiveness of training initiatives in improving job performance.

The analysis produced several significant findings. The number of training courses attended was found to be the main factor affecting performance, suggesting a positive relationship between training and performance ratings. This highlights the significance of providing resources to employee development initiatives. Furthermore, the dataset provided valuable information regarding the distribution of genders, job titles, age groups, and levels of experience. This data provided insight into the demographic and professional attributes of the workforce.

The study utilized seven machine learning algorithms and determined that Random Forest achieved the highest level of classification accuracy, with Artificial Neural Network coming in a close second. These findings indicate that Random Forest is highly skilled at identifying deeper trends in the data to make accurate predictions about performance.

In addition, the study showed extensive relationships between variables such as age, experience, and performance, as well as variations in performance across various departments. The findings also emphasized demographic characteristics and educational experiences that are linked to exceptional performance, including possessing a bachelor's degree, a PhD, and specialized education in specific academic departments.

High-confidence predictions (Error = 0) were associated with job performance category 2 employees aged 30–50, with a balanced gender distribution and 10–15 years of experience. They had 1-2 training courses. Low-confidence predictions (error > 0) involved job performance category 3 employees aged 35–45, with more males and staff. These workers had 15 years of experience and 2 training courses.

To summaries, the research concluded that Random Forest is the most efficient model for predicting employee performance using the Ministry of Labour dataset. The results of this study provide valuable information for human resources management and administrators, highlighting the significance of focused training programs and the potential of machine learning algorithms to improve the accuracy of performance evaluations.

REFERENCE

Akom, M.S., Obeng-Sarpong, C., Aflakpui, F.E., & Sarpong, S.A. (2021) 'Exploring Leave Management Practices and Relationship with Performance of Administrative Staff: Evidence from a Tertiary Institution in Ghana', Universal Journal of Management, 9, pp. 71-82.

Al-Radaideh, Q. A., & Al-Nagi, E. (2012) 'Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance', International Journal of Advanced Computer Science and Applications, 3(2), pp. 144-150.

Almawali, H. A., Hafit, N. I. A., & Hassan, N. (2021) 'Motivational Factors and Job Performance: The Mediating Roles of Employee Engagement', International Journal of Human Resource Studies, 11(3), pp. 67-77.

Badshah, B. H. (2022) 'Authentic Leadership and Staff Performance Mediated by Work Engagement; Case of Marketing Firms in Oman', Journal of Management Science. Available at: https://journals.qurtuba.edu.pk/ojs/index.php/jms/article/view/533 (Accessed: 20 April 2024).

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020) 'A novel selective naïve Bayes algorithm', Knowledge-Based Systems, 192, 105361.

Elnaga, A., & Imran, A. (2013) 'The effect of training on employee performance', European Journal of Business and Management, 5(4), pp. 137-147.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020) 'Predicting Employee Attrition Using Machine Learning Techniques', Computers, 9, 86.

Gomez-Gil, F. J., Martínez-Martínez, V., Ruiz-Gonzalez, R., Martínez-Martínez, L., & Gomez-Gil, J. (2024) 'Vibration-based monitoring of agro-industrial machinery using a k-Nearest Neighbors (KNN) classifier with a Harmony Search (HS) frequency selector algorithm', Computers and Electronics in Agriculture, 217, 108556.

Hidayat, R., & Budiatma, J. (2018) 'Education and job training on employee performance', International Journal of Social Sciences and Humanities, 2(1), pp. 171-181.

Juba, B., & Le, H. S. (2019) 'Precision-Recall versus Accuracy and the Role of Large Data Sets', in The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), pp. 4040- 4046.

Kirimi, J. M., & Moturi, C. A. (2016) 'Application of data mining classification in employee performance prediction', International Journal of Computer Applications, 146(7), pp. 28-35.

Lather, A. S., Malhotra, R., Saloni, P., Singh, P., & Mittal, S. (2019, November) 'Prediction of employee performance using machine learning techniques', in Proceedings of the 1st International Conference on Advanced Information Science and System, pp. 1-6.

Madiedo, J.P., Chandrasekaran, A. and Salvador, F. (2020) 'Capturing the Benefits of Worker Specialization: Effects of Managerial and Organizational Task Experience', Prod Oper Manag, 29, pp. 973-994.

Mojsilović, M., Cvejić, R., Pepić, S., Karabašević, D., Saračević, M., & Stanujkić, D. (2023) 'Statistical evaluation of the achievements of professional students by combination of the random forest algorithm and the ANFIS method', Heliyon, 9, e21768.

Ratnawati, E., Sukidjo, S., & Efendi, R. (2020) 'The effect of work motivation and work experience on employee performance', International Journal of Multicultural and Multireligious Understanding, 7(8), pp. 109-116.

Rivaldo, Y., & Nabella, S. D. (2023) 'Employee Performance: Education, Training, Experience and Work Discipline', Calitatea, 24(193), pp. 182-188.

Saadouli, N. & Al-Khanbashi, M. (2021) 'Evaluation of factors affecting employee performance: The case of government employees in Oman', Management Science Letters, 11(5), pp. 1443-1450.

Tanni, S., Patino, C., & Ferreira, J. (2020) 'Correlation vs. regression in association studies', Jornal Brasileiro De Pneumologia, 46(1).

Tong, Z., Meng, Y., Zhang, J., Wu, Y., Li, Z., Wang, D., Li, X., & Ou, G. (2023) 'A novel coal structure identification method based on wavelet transform and particle swarm optimization- support vector machine', Fuel, 305, 122960.

Wuryani, E., Rodli, A. F., Sutarsi, S., Dewi, N. N., & Arif, D. (2021) 'Analysis of decision support system on situational leadership styles on work motivation and employee performance', Management Science Letters, 11(2021), pp. 365-372.

V. Thamilarasi, "Artificial Intelligence-Driven Smart Scenic Management: Automated Decision Making and Optimization", Int J Intell Syst Appl Eng, vol. 12, no. 4, pp. 2731 –, Jun. 2024.

Zaïdi, A. (2023) 'Predicting wildfires in Algerian forests using machine learning models', Heliyon, 9(7), e18064.

Thamilarasi V., Roselin. R., "Application of Machine Learning in chest X-Ray images", International book & River publisher series in computing and information science and Technology, pp: 52-1,2023. ISBN : 9788770228114, 9788770228107.