International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 09 | Sept – 2025 DOI: 10.55041/ISJEM05026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Project Development and Evaluation of Project Risk Analysis Model Using Machine Learning Algorithms

K. Tulasi Krishna Kumar[1], Mrs P.T.S Priya[2]

[1] Department of CSE, SVPEC [Andhra University]
[2] Department of MCA, SVPEC [Andhra University]

Abstract
Projects are undertaken by the contemporary organizations to experience significant growth, deal with industry competition and achieve sustainability. The increase in size, complexity and duration of the projects resulted in facing several risks and this project is undertaken to develop a project risk analysis model with the help of machine learning algorithms. Identifying that machine learning is feasible for making risk analysis process autonomous, random forest and logistic regression were the two machine learning algorithms utilized to develop the model and then train it. Evaluation of the project risk analysis model showed its effectiveness in recording high number of true positives and true negatives during the detection activity of project risks. From evaluation, random forest is identified as the effective algorithm due to high prediction accuracy, robustness and avoiding overfitting problem.
Keywords: Project risks, risk analysis, supervised machine learning, random forest

1. Introduction
Projects are temporary endeavours undertaken by organisations to achieve specific objectives, often within dynamic, complex, and uncertain environments. Such conditions expose projects to numerous risks that are difficult to predict and manage. Traditional risk management—identifying, assessing, prioritising, and mitigating risks manually—demands significant time and effort while being prone to human error. As projects grow in scale and complexity, these manual methods have become less effective, prompting the need for more advanced approaches (Chattapadhyay et al., 2021; Sarkar & Maiti, 2020).
To overcome these challenges, researchers have explored automated risk assessment models using machine learning algorithms. Prior studies have developed autonomous models in industries such as construction and software (Shah, 2021; Filippetto et al., 2021), but these often rely on pre-existing or simulated datasets, limiting their applicability and accuracy. Moreover, current models fail to identify the most effective algorithms for project risk analysis and are often industry-specific. To address these gaps, this project aims to develop and evaluate a machine learning–based risk analysis model supported by a custom dataset designed to automatically identify, assess, and analyse project risks more effectively.

1.1 Overview on Risk Analysis Model
Project management is a critical activity for enhancing project performance, particularly in software projects where managers often struggle with efficiency and sustainability due to limited knowledge, technology, skills, and resources (Mahadi et al., 2021). These challenges increase the likelihood of project failure, especially as rapidly changing requirements introduce new risks that significantly influence software outcomes (Sousa et al., 2021). Effective risk planning, assessment, and management—ideally in an automated manner—have therefore become essential for improving project success (Lin et al., 2021).
Literature highlights that automated risk management should be scalable, capable of detecting and notifying various types of risks across projects. Studies show that risk management typically involves four stages: identification, assessment, analysis, and treatment (Jomthanachai et al., 2021). Some research further emphasizes risk analysis, including not only identification but also reporting the severity of risks. While human intervention in risk management remains common in industries like construction, the highly dynamic and uncertain nature of such projects underscores the need for more autonomous, technology-driven approaches (Mashrur et al., 2020; Sanni-Anibire et al., 2022).

1.2 Feasibility of Machine Learning
Machine learning adds intelligence to machines by enabling them to learn from datasets, making it a valuable tool for risk management (Sousa et al., 2021; Naseem et al., 2021). Its applications in this area involve building models, training them with datasets, evaluating their performance, and validating results (Naseem et al., 2021).

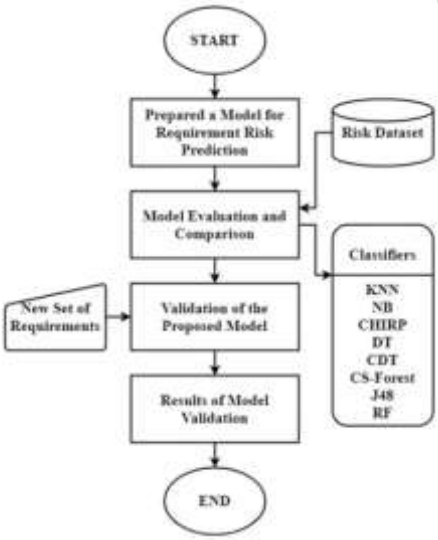


Figure 1: Showing the important job dealing (Naseem et al, 2021)

Studies show that machine learning supports the development of frameworks for risk assessment across diverse sectors, including construction, software, and financial projects (Ferreira et al., 2020; Mashru et al., 2020). By applying techniques such as artificial intelligence, deep learning, and decision tree classifiers, researchers have demonstrated the potential of machine learning to reduce human errors, minimize manual intervention, and improve the accuracy of risk detection and analysis (Malhotra, 2018; Anysz et al., 2021).

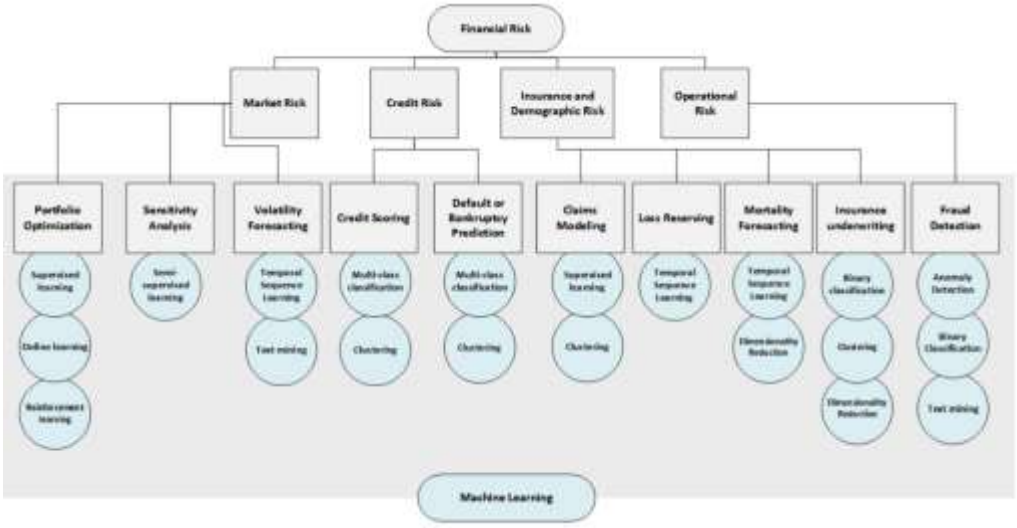


Figure 2: Risks which dedicated taxonomy (Mashru et al, 2020)

Compared to traditional risk management methods, machine learning approaches are more dynamic, cost-effective, and precise (Sousa et al., 2021). They enable automated identification, assessment, and prediction of risks, making them particularly valuable for complex and uncertain projects. In software project management, for example, machine learning algorithms have been shown to effectively analyse project risks and improve decision-making (Mahti et al., 2021; Naseem et al., 2021). Overall, literature highlights that the integration of machine learning into risk management processes provides scalability, adaptability, and intelligence, positioning it as a transformative approach to improving project outcomes.

2. Design details of project

2.1 Creation of data set

The design of this project outlines the working process of the autonomous risk analysis model. To represent the system's structure and behaviour, **UML diagrams** have been selected as the primary design tool. These diagrams, along with supporting explanations, illustrate the flow of operations and decision-making within the proposed model.In particular, the **activity diagram** demonstrates the functional workflow of the machine learning–based risk assessment model. It begins with the creation of new datasets, followed by preprocessing,

training, and evaluation of the model. Subsequent steps include the identification and classification of risk levels, along with decision points that guide appropriate mitigation actions. This diagram provides a clear visualization of how the model operates autonomously to analyse and assess project risks.

Figure 3: Activity diagram for risk analysis model (Source: Author)

2.2 Built environment creation

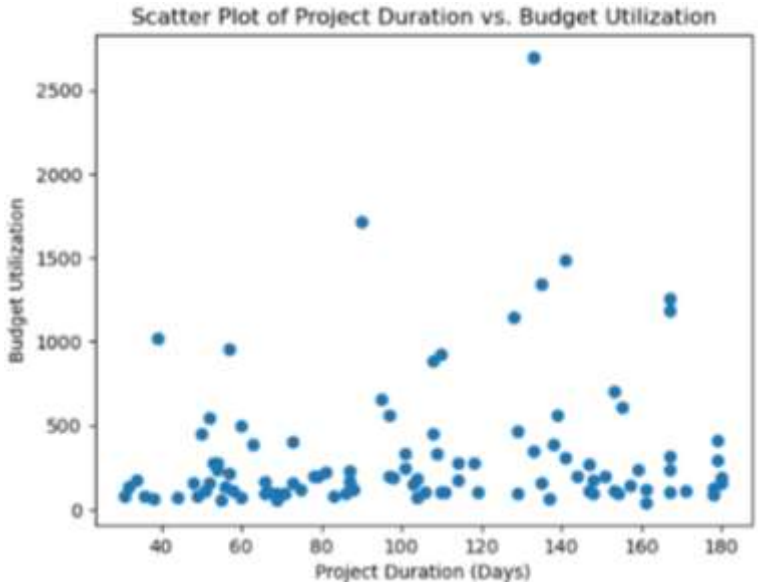
The creation of built in environment is done by getting the required tools installed in the required computer environment. The main environment to do the coding part, compile the code and then execute the code is provided in Jupiter notebook. The commands which were used to download Jupiter notebook and then install it are shown above. After the implementation of all these commands in the command prompt, installation was done successfully.

3.Preprocessing of data set

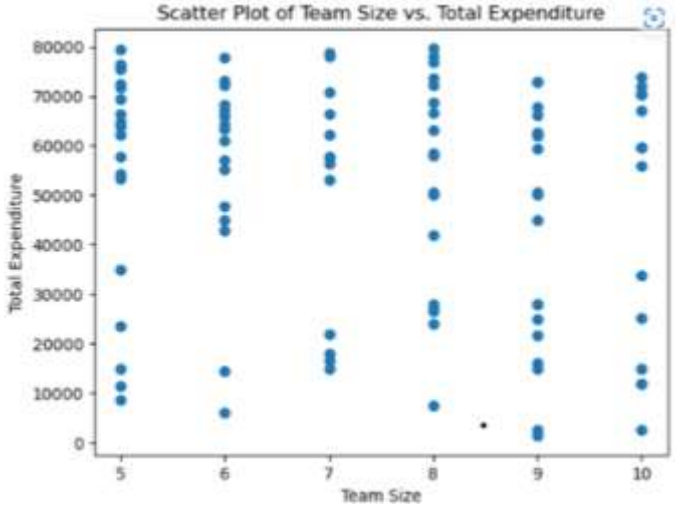
3.1 Exploratory data analysis

Exploratory data analysis is one way of doing the attribute based analysis on the data set.At mentioned by Robinson et al (2022) this exploratory data analysis is widely used to understand the way in which the attributes present in a data set are related to each other.The important features of tasks data set or project duration and budget utilisation. EDA application on this task data set helped in showing the way in which these two attributes are related to each other and mentioned in scatter plot is given below.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 09 | Sept – 2025 DOI: 10.55041/ISJEM05026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata



This scatterplot indicates that with increasing in the duration of the project, the budget utilisation has also increased. However, the increase in the project duration above 100 days resulted in increasing the amount of budget that is being utilised. Projects data file is also critically analysed to note the attributes and the relation existing among these attributes. Because of team size and total expenditure identified as the key features of this projects datafile, they both are related with each other using EDA and mentioned using a scatter plot shown below.



This scatter plot indicates the presence of large extent of outliers, thereby indicating the differences in the relation between team size and the total expenditure of the project. It is after increase in the size of the team above six members, there is a great increase in the total expenditure of the project, which is about 50,000 units, as mentioned in the above graph.

4. Testing, Evaluation

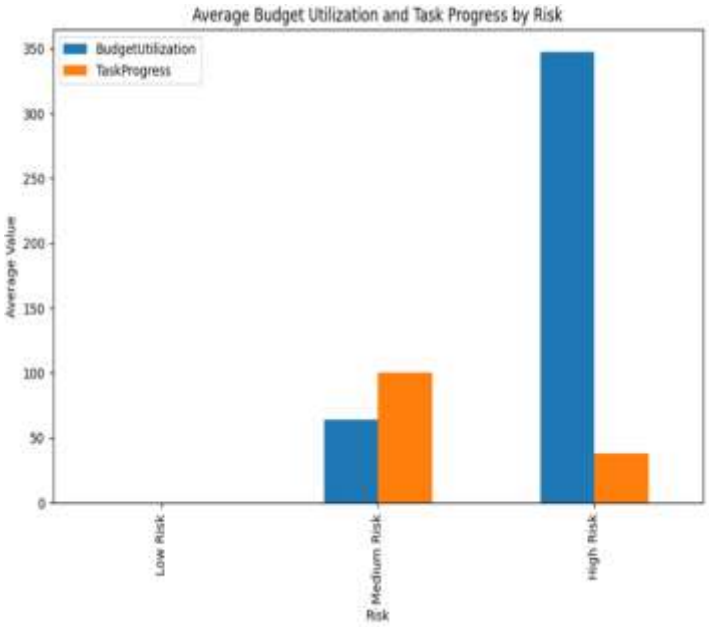
1st test case 1: Risk calculation

This is the first test case prepared with the objective of checking if the model is able to perform the calculations on budget utilisation and task progress levels. To conduct this test case, the two datasets are given in the form of input. Based on the given input details, calculation is carried out for budget utilisation and progress of the task. The result of all the calculations is displayed above.

2nd test case: Categorisation of risks

The second test case used for performing data testing activity on the developed risk analysis model is the categorization of the risks.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 09 | Sept – 2025 DOI: 10.55041/ISJEM05026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

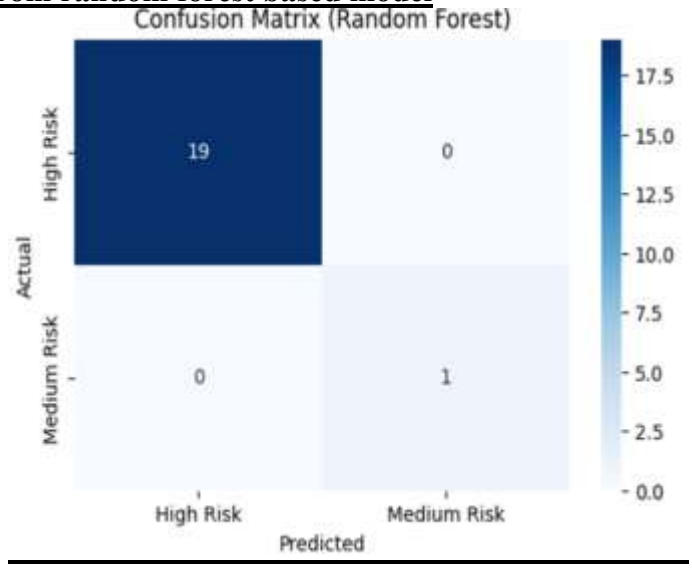


After identifying the risks, data visualisation of the output helped in doing the risk categorization among the three categories: high risk, low risk and medium risk. This graphical representation shows the majority of the data grouping under medium and high risk categories.

5.1 Evaluation activity

The purpose of evaluation in this project is to identify the machine learning algorithm which is effective to make the prediction of risk and risk level in accurate manner.

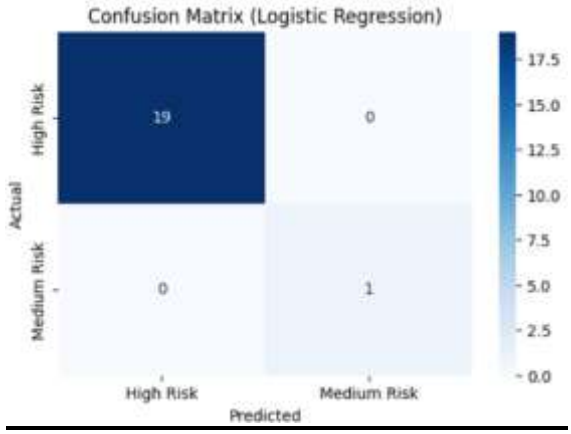
a. Confusion matrix from random forest based model



Confusion matrix is also the result of showing that model is working in making the prediction about the level of risk. As the model is trained using random forest machine learning algorithm, this confusion matrix indicate the number of true positives or higher with value of 19. Such a value indicate that it is actually high risk, this random forest based model has also predicted accurately.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 09 | Sept – 2025 DOI: 10.55041/ISJEM05026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

b. Confusion matrix from regression analysis based model



After developing the model using logistic regression, the result is given in the form of confusion matrix. This result also indicate that the model is able to provide high number of true positives because it is able to make the predictions accurately.

5.2 Execution

```
[5]: import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.model_selection import train_test_split
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import classification_report, confusion_matrix, roc_curve, auc
     import joblib

[6]: # Load data from CSV files
     projects_df = pd.read_csv('random_projects (2).csv')
     tasks_df = pd.read_csv('tasks1.csv')
     risks_df = pd.read_csv('random_risks (1).csv')

[7]: # Preprocessing for random_projects.csv
     # Example: Date cleaning and feature engineering
     projects_df['StartDate'] = pd.to_datetime(projects_df['StartDate'], format='%d/%m/%Y')
     projects_df['EndDate'] = pd.to_datetime(projects_df['EndDate'], format='%d/%m/%Y')
     projects_df['ProjectDuration'] = (projects_df['EndDate'] - projects_df['StartDate']).dt.days

[8]: # Preprocessing for tasks.csv
     # Example: Date cleaning and feature engineering
     tasks_df['CreatedAt'] = pd.to_datetime(tasks_df['CreatedAt'])
     tasks_df['DueDate'] = pd.to_datetime(tasks_df['DueDate'])
     tasks_df['TaskDuration'] = (tasks_df['DueDate'] - tasks_df['CreatedAt']).dt.days
```


International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 09 | Sept – 2025 DOI: 10.55041/ISJEM05026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

```
# Train a Random Forest Classifier
clf_rf = RandomForestClassifier(random_state=42)
clf_rf.fit(X_train, y_train)
```

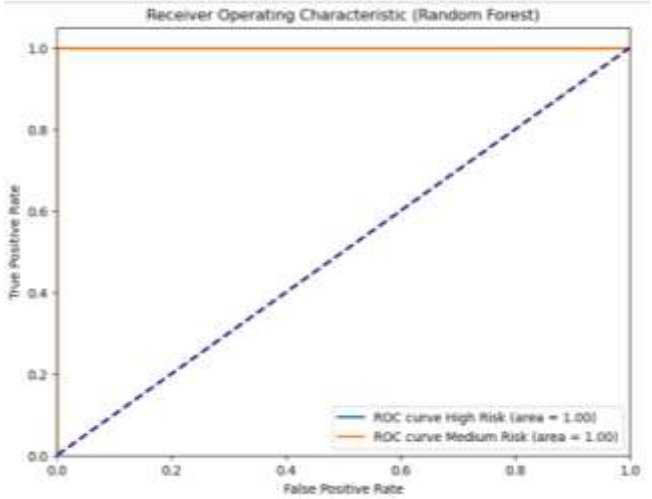
```
RandomForestClassifier(random_state=42)
```

```
# Train a Logistic Regression model for comparison
clf_lr = LogisticRegression(solver='liblinear')
clf_lr.fit(X_train, y_train)
```

```
LogisticRegression(solver='liblinear')
```

```
# Predict the risk for the test set using Random Forest
y_pred_rf = clf_rf.predict(X_test)
```

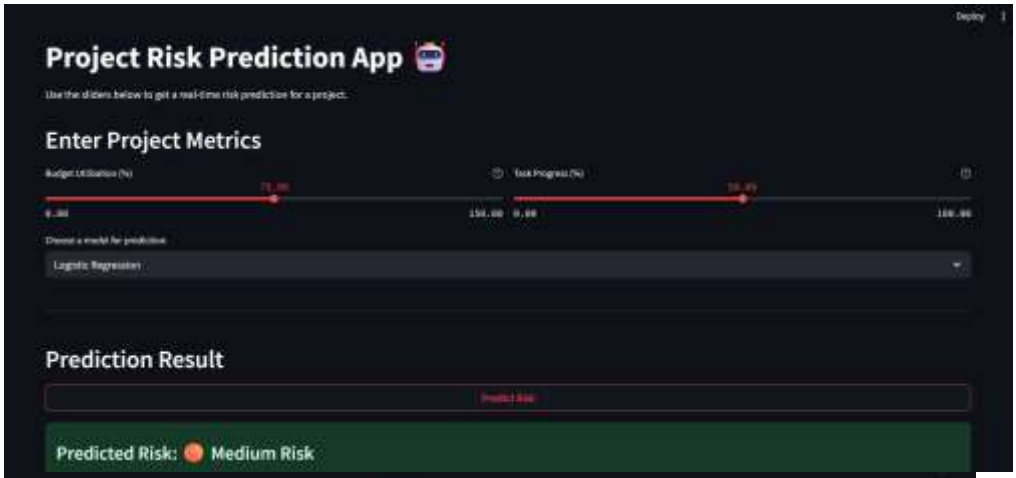
```
# Predict the risk for the test set using Logistic Regression
y_pred_lr = clf_lr.predict(X_test)
```

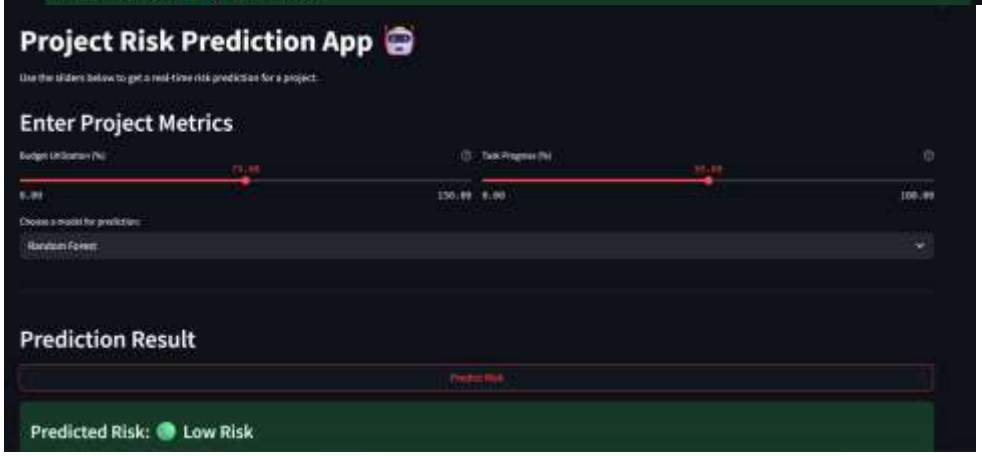



```
# Heatmap for Confusion Matrix
plt.figure(figsize=(6, 4))
cm_rf = confusion_matrix(y_test, y_pred_rf)
sns.heatmap(cm_rf, annot=True, fmt="d", cmap="Blues", xticklabels=clf_rf.classes_, yticklabels=clf_rf.classes_)
plt.title('Confusion Matrix (Random Forest)')
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

```
Text(45.72222222222221, 0.5, 'Actual')
```

ISSN: 2583-6129
DOI: 10.55041/ISJEM05026






```
[37]: # Calculating the Loss column
risk_report['Loss'] = risk_report['TotalBudget'] - risk_report['TotalExpenditure']
risk_report['Loss'] = risk_report['Loss'].apply(lambda x: 0 if x >= 0 else x)

risk_report.to_csv('risk_report.csv', index=False)
print(projects_df.head())
plt.show();
```

```
   ProjectID                  ProjectName  StartDate    EndDate  \
0          1                Software Upgrade 2023-11-10 2024-01-04
1          2         Green Energy Initiative 2023-11-10 2023-12-29
2          3          Social Media Marketing 2023-11-22 2024-05-07
3          4  E-commerce Platform Development 2023-11-09 2024-05-05
4          5                Software Upgrade 2023-11-19 2024-04-24

  ProjectManager  TotalBudget  ActualExpenditure  TotalExp  TeamSize  \
0       John Doe        38911              21647      2805         9
1    Bob Johnson        14126              11530     73987         5
2       Ava Wong        24026              25200     78993        10
3    Bob Johnson        12076              15000     63894         7
4    Bob Johnson        12872              18000     72814         7

   ProjectDuration  TotalTasks  TotalExpenditure  BudgetUtilization  \
0               55          23             21647          55.632083
1               49          24             11530          81.622540
2              167          35             25200         104.886373
3              178          22             15000         124.213316
4              157          21             18000         139.838409
```

6. Conclusion

The main aim of this project was to develop an automated risk analysis model capable of accurately predicting project risks using machine learning algorithms. A review of existing literature confirmed the feasibility of automating risk analysis, highlighting the potential of ML to enhance model intelligence. To achieve this, a new dataset was created, consisting of project and task files enriched through feature engineering with attributes such as budget utilization and task progress. The model was developed and tested using Random Forest and Logistic Regression, trained on the dataset to enable autonomous risk detection. Results indicated that 96% of identified risks fell into the high-risk category. Finally, evaluation using a confusion matrix showed a high number of true positives, confirming both algorithms to be effective in predicting risks and risk levels with strong accuracy.

References

- Ahmadi, S. (2023). Optimizing Data Warehousing Performance through Machine Learning Algorithms in the Cloud. *International Journal of Science and Research (IJSR)*, *12*(12), 1859-1867.
- Bachute, M. R., &Subhedar, J. M. (2021). Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, *6*, 100164.
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., &Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, 3-21.
- Babikian, J. (2023). Navigating Legal Frontiers: Exploring Emerging Issues in Cyber Law. *Revista Espanola de DocumentacionCientifica*, *17*(2), 95-109.
- Chandrinos, S. K., Sakkas, G., &Lagaros, N. D. (2018). AIRMS: A risk management tool using machine learning. *Expert Systems with Applications*, *105*, 34-48.
- Chattapadhyay, D., Putta, J., & Rao P, R. M. (2021). Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model. B*uildings*, *11*(4), 172.
- Fan, C. L. (2020). Defect risk assessment using a hybrid machine learning method. *Journal of Construction Engineering and Management*, *146*(9), 04020102.
- Ferreira, L., Pilastri, A., Martins, C., Santos, P., & Cortez, P. (2020, February). A scalable and automated machine learning framework to support risk management. In *International Conference on Agents and Artificial Intelligence* (pp. 291-307). Cham: Springer International Publishing.
- Filippetto, A. S., Lima, R., & Barbosa, J. L. V. (2021). A risk prediction model for software project management based on similarity analysis of context histories. *Information and Software Technology*, *131*, 106497.
- Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine learning algorithms for construction projects delay risk prediction. *Journal of Construction Engineering and Management*, *146*(1), 04019085.
- Grotov, K., Titov, S., Sotnikov, V., Golubev, Y., &Bryksin, T. (2022, May). A large-scale comparison of Python code in Jupyter notebooks and scripts. In *Proceedings of the 19th International Conference on Mining Software Repositories* (pp. 353-364).
- Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., & Tiwari, V. (2022). Comparing the performance of machine learning algorithms using estimated accuracy. *Measurement: Sensors*, *24*, 100432.
- Hegde, J., &Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment–A review. *Safety science*, *122*, 104492.
- Hill, C. (2020). *Learning scientific programming with Python*. Cambridge University Press.
- Hillson, D., & Simon, P. (2020). *Practical project risk management: The ATOM methodology*. Berrett-Koehler Publishers.
- Hubbard, D. W. (2020). *The failure of risk management: Why it's broken and how to fix it*. John Wiley & Sons.
- Jomthanachai, S., Wong, W. P., & Lim, C. P. (2021). An application of data envelopment analysis and machine learning approach to risk management. *Ieee Access*, *9*, 85978-85994.
- Lin, S. S., Shen, S. L., Zhou, A., & Xu, Y. S. (2021). Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. *Automation in Construction*, *122*, 103490.
- Ma, G., Wu, Z., Jia, J., & Shang, S. (2021). Safety risk factors comprehensive analysis for construction project: Combined cascading effect and machine learning approach. S*afety science*, *143*, 105410.
- Mahdi, M. N., MH, M. Z., Yusof, A., Cheng, L. K., Azmi, M. S. M., & Ahmad, A. R. (2020). Design and Development of Machine Learning Technique for Software Project Risk Assessment-A Review. In *2020 8th International Conference on Information Technology and Multimedia (ICIMU)* (pp. 354-362). IEEE.
- Mahdi, M. N., Mohamed Zabil, M. H., Ahmad, A. R., Ismail, R., Yusoff, Y., Cheng, L. K., ... &Happala Naidu, H. (2021). Software project management using machine learning technique—A Review. *Applied Sciences*, *11*(11), 5183.
- Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.
- Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *Ieee Access*, *8*, 203203-203223.
- Naseem, R., Shaukat, Z., Irfan, M., Shah, M. A., Ahmad, A., Muhammad, F., ... & Sulaiman, A. (2021). Empirical assessment of machine learning techniques for software requirements risk prediction. *Electronics*, *10*(2), 168.
- Nelson, M. J., & Hoover, A. K. (2020, June). Notes on using Google Colaboratory in AI education. In *Proceedings of the 2020 ACM conference on innovation and Technology in Computer Science Education* (pp. 533-534).
- Paltrinieri, N., Comfort, L., & Reniers, G. (2019). Learning about risk: Machine learning for risk assessment. S*afety science*, *118*, 475-486.
- Pozgar, G. D. (2023). *Legal and ethical issues for health professionals*. Jones & Bartlett Learning.
- Robinson, D., Ernst, N. A., Vargas, E. L., & Storey, M. A. D. (2022, May). Error identification strategies for Python Jupyter notebooks. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension* (pp. 253-263).
- Werth, A., Oliver, K. A., West, C. G., & Lewandowski, H. J. (2022, September). Engagement in collaboration and teamwork using Google Colaboratory. In *PERC Proceedings*.
- Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021). Application of deep learning algorithms in geotechnical engineering: a short critical review. *Artificial Intelligence Review*, 1-41.
- Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., ... & Zhang, H. (2021). Machine learning: new ideas and tools in environmental science and engineering. *Environmental Science & Technology*, *55*(19), 12741-12754.

Bibliography

K. Tulasi Krishna Kumar is a Training & Placement Officer with 15 years of experience in training and placing students in IT, ITES, and Core sectors. He has trained 9,700+ students and 450+ faculty members through FDPs. Author of 6 books on CRT & Computer Science, Certified Campus Recruitment Trainer (JNTUA), holds an M.Tech in CSE, and is pursuing his Ph.D. A CITD-certified Pro-E, CNC professional. He has published 65+ research papers in international journals on Databases, Software Engineering, HRM, and CRT.

Mrs. P. T. S. Priya is currently serving as the Head of the Department (HoD) for MCA at SVPPEC. She is a postgraduate in MCA and, driven by her passion for higher learning, is presently pursuing an M.Tech in Computer Science and Technology (CST). With a keen interest in the field of Machine Learning, she is actively engaged in a research project titled “Project Development and Evaluation of Project Risk Analysis Model Using Machine Learning Algorithms.” This project focuses on leveraging advanced algorithms to design an autonomous system capable of identifying and evaluating risks in project management, thereby contributing to both academic knowledge and practical applications in the field. She is undertaking this project under the guidance of Mr. K. Tulasi Krishna Kumar, whose expertise is helping shape the study towards meaningful outcomes.