

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**RISK-BASED COST ESTIMATION IN CONSTRUCTION BY EMPLOYING
MACHINE LEARNING TECHNIQUES**



Ph.D. THESIS

Aynur Hürriyet TÜRKYILMAZ

Department of Civil Engineering

Structure Engineering Programme

APRIL 2025

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**RISK-BASED COST ESTIMATION IN CONSTRUCTION BY EMPLOYING
MACHINE LEARNING TECHNIQUES**

Ph.D. THESIS

**Aynur Hürriyet TÜRKYILMAZ
(501192002)**

Department of Civil Engineering

Structure Engineering Programme

Thesis Advisor: Prof. Dr. Gül POLAT TATAR

APRIL 2025

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**MAKİNE ÖĞRENİMİ TEKNİKLERİNİ KULLANARAK İNŞAATTA RİSK
TABANLI MALİYET TAHMİNİ**

DOKTORA TEZİ

**Aynur Hürriyet TÜRKYILMAZ
(501192002)**

İnşaat Mühendisliği Anabilim Dalı

Yapı Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Gül POLAT TATAR

NİSAN 2025

Aynur Hürriyet TÜRKYILMAZ, a Ph.D. student of İTÜ Graduate School student ID 501192002, successfully defended the thesis/dissertation entitled “RISK-BASED COST ESTIMATION IN CONSTRUCTION BY EMPLOYING MACHINE LEARNING TECHNIQUES”, which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Gül POLAT TATAR**

Istanbul Technical University

Jury Members : **Assoc. Prof. Dr. Atilla DAMCI**

Istanbul Technical University

Prof. Dr. Aşlı Pelin GÜRGÜN

Yildiz Technical University

Assoc. Prof. Dr. Onur Behzat TOKDEMİR

Istanbul Technical University

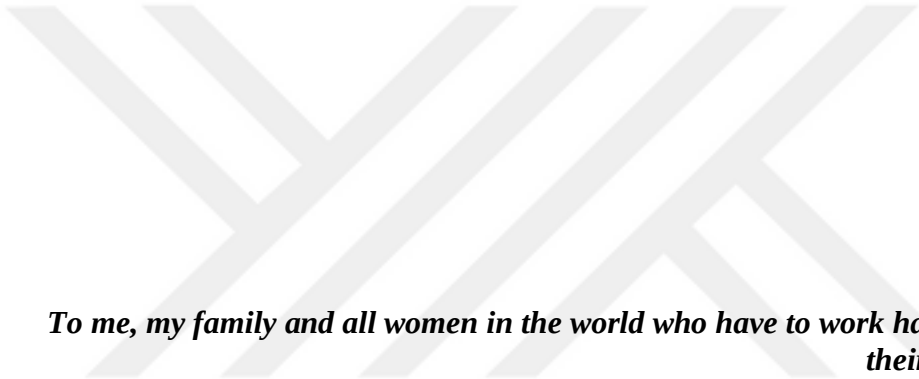
Assoc. Prof. Dr. Sevilay DEMİRKESEN ÇAKIR

Gebze Technical University

Date of Submission : 03 March 2025

Date of Defense : 17 April 2025





*To me, my family and all women in the world who have to work hard regardless of
their circumstances,*



FOREWORD

Doktora yolculuğumun ilk gününden son gününe kadar gözümdeki ışığı hep koruyan; bana özgür bir çalışma ortamı sağlayıp, bir yandan da tecrübesiyle her zaman yol gösteren çok değerli danışmanım Prof. Dr. Gül POLAT TATAR'a en içten minnetlerimi sunarım.

Destegini benden esirgemeyerek, hem iş hem akademik alanda cesaretlendiren ve yol gösteren; önerilerini can kulağıyla dinlediğim, çok kıymetli yöneticim Sayın Ayşe Feride ARMANGİL GÖKALP'e içten minnet ve saygılarımı sunarım.

Akademiye ve sektöre değerli katkılarda bulunan, her koşulda eğitimi ve gelişimi destekleyen TAV İnşaat'a; bu süreçte yanımda olan değerli yöneticilerime ve çalışma arkadaşlarıma derin teşekkürlerimi ve saygılarımı sunarım.

Her zaman doktora yapmam gerektiğine inanan, benim ve birçok öğrenci için elini taşın altına koyan, hem hocam, hem ailem olan çok kıymetli hocalarım Erhan KARAESMEN ile Engin KARAESMEN ve asistanı olduğum dönemde kendisinden çok şey öğrendiğim ilk hocam, kıymetli Prof. Dr. Jerome HAJJAR'a en içten saygı ve teşekkürlerimi sunarım. Kendisi 3 yaşındayken başladığım doktora yolculuğumda bana hep destek olan, sabah gün doğmadan kalktığında beni bilgisayar başında görünce gülümseyerek tüm hayatımıza ışık dolduran, canım oğlumuz Taylan Mete TÜRKYILMAZ'a içten sevgilerimi sunarım. Tüm bunlar yaşanırken dışa vurduğum stresimi sönümleyen, beraber büyüdüğümüz, birbirimizin hayatının en büyük tanığı olduğumuz sevgili hayat arkadaşım Okan TÜRKYILMAZ'a teşekkürlerimi sunarım. Her coğrafyada, her zaman aralığında kadın olmanın zor ve meşakkatli bir iş olurken aynı zamanda bir sanat ve denge işine dönüşeceğini bana gösteren ve her zaman çalışkanlıkları, enerjileri, neşeleri ve gülümseyen gözleriyle yoluma ışık tutan hayatımdaki en önemli iki kadın annem Sebahat ÖZTÜRK ve rahmetli anneannem Mahmure ÖZTÜRK'e içten saygı ve minnetlerimi sunarım.

İlk gününden son gününe kadar doktora konuma dair yürüttüğüm her çalışmadan çok keyif aldım. Okul, ev ve iş hayatını bir arada götürmek kolay değildi. Fiziksel ve ruhsal yorgunluk çektiğim pek çok zaman oldu. Tüm bunlara rağmen bir gün bile kendi içimde yakınlıkla çalışmadığım, yılmadığım, yorulmadığım ve keyif alarak yürüdüğüm için kendime çok teşekkür ediyorum.

Son ve en önemli olarak ulu önderimiz Mustafa Kemal ATATÜRK'e, bize her alanda ilham olduğu, lider olduğu ve bizlere sağladığı haklar için en derin saygı ve minnetlerimi sunarım.

İTÜ'nin bir parçası olmak benim için onur ve mutluluk kaynağı oldu. Hayat bir ekip işi. Her durumda bir parçası olduğum tüm ekip üyelerine, üzerimde emeği olan herkese saygı ve teşekkürlerimi sunarım.

April 2025

Aynur Hürriyet TÜRKYILMAZ
(Civil Engineer, M.Sc.)

TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	ix
TABLE OF CONTENTS	xi
ABBREVIATIONS	xiii
SYMBOLS	xv
LIST OF TABLES	xvii
LIST OF FIGURES	xix
SUMMARY	xxi
ÖZET	xxiii
1. INTRODUCTION	1
1.1 Problem Statement and Significance of Thesis	1
2. LITERATURE REVIEW	5
2.1 Risks in construction projects	5
2.2 Risk perception	8
2.3 Cost estimation methods in construction projects.....	14
2.4 Cost overrun in construction projects.....	16
2.5 Machine Learning Techniques.....	18
2.5.1 Supervised Learning	19
2.5.2 Unsupervised Learning	19
2.5.3 Reinforcement Learning	20
3. RESEARCH METHODOLOGY	21
3.1 Data gathering	21
3.1.1 Gathering risk register	21
3.1.2 Gathering estimated completion cost.....	22
3.1.3 Grouping projects and data	23
3.2 Modeling.....	24
3.2.1 Machine learning regression models	24
3.2.2 Machine learning classification models.....	25
3.3 Reporting the results.....	26
4. IMPLEMENTATION OF THE PROPOSED APPROACH: CASE STUDIES	27
4.1 Data gathering	27
4.2 Modelling.....	29
4.2.1 Risk-based completion cost estimation using machine learning prediction algorithms.....	29
4.2.1.1 Data understanding and preprocessing.....	29
4.2.1.2 Model selecting and modeling.....	31
4.2.1.3 Evaluating and validating the prediction models.....	32
4.2.2 Risk-based completion cost overrun ratio estimation using machine learning classification algorithms	41
4.2.2.1 Data understanding and preprocessing.....	41
4.2.2.2 Model selecting and modeling.....	42
4.2.2.3 Evaluating and validating the classification models	45

4.3 Results	51
5. EVALUATION AND CONCLUSION.....	53
5.1 Benefits of the proposed approaches	53
5.2 Limitations of study	54
5.3 Recommendations for future research	54
5.4 Conclusion.....	55
REFERENCES	57
APPENDICES	75
APPENDIX A	76
APPENDIX B	77
APPENDIX C	78
APPENDIX D.....	82
CURRICULUM VITAE.....	85



ABBREVIATIONS

ACC	: Accuracy
ANN	: Artificial Neural Network
BIM	: Building Information Modelling
CV	: Cross Validation
DT	: Decision Tree
DTC	: Decision Tree Classifier
FN	: False Negative
FP	: False Positive
I	: Impact
KNN	: K-Nearest Neighbors
LOOCV	: Leave One out Cross Validation
LR	: Linear Regression
LRC	: Logistic Regression for Classification
ML	: Machine Learning
MSE	: Mean Square Error
NBC	: Naïve Bayes Classifier
P	: Probability
PR2	: 2 nd Degree Polynomial Regression
PR4	: 4 th Degree Polynomial Regression
PT	: Project Type
RF	: Random Forest
RFC	: Random Forest Classifier
SVC	: Support Vector Classifier
T/A/K	: Title, Abstract, and Keyword Search
TC	: Total Cost
TN	: True Negative
TP	: True Positive
TRS	: Total Risk Score



SYMBOLS

k	: Total number of executives
m	: Total number of risk categories
n	: Total number of risks
p	: Number of data points
q	: Number of folds
s	: Standard deviation of a sample
X	: Total completion cost value in a related project
y_i	: Observed value
W_{i,e}	: Weight value assigned by the e th executive for i th risk category
P_{i,j,e}	: Probability value assigned by the e th executive for j th risk item in i th risk category
I_{i,j,e}	: Impact value assigned by the e th executive for j th risk item in i th risk category
$\hat{y}_i$	: Predicted value
$\bar{y}$	: Mean value of a sample



LIST OF TABLES

	<u>Page</u>
Table 3.1 : Project types and groups (Hendrickson and Au, 1989)	23
Table 4.1 : Risk probability scale..	27
Table 4.2 : Risk impact scale.....	28
Table 4.3 : R^2 for training group and all data group (PT.01)	33
Table 4.4 : R^2 for training group and all data group (PT.02)	34
Table 4.5 : R^2 for test group (PT.01).....	35
Table 4.6 : R^2 for test group (PT.02).....	36
Table 4.7 : R^2 for test groups in different random states (PT.01).....	37
Table 4.8 : R^2 for test groups in different random states (PT.02).....	37
Table 4.9 : $CV_{(q)}$ for 10 states in PT.01.....	38
Table 4.10 : $CV_{(q)}$ for 10 states in PT.02.....	38
Table 4.11 : Classes for cost overrun ratios	42
Table 4.12 : Accuracy percentages of PT groups according to algorithms.....	46
Table 4.13 : Overall accuracy values for 10 random states.....	49
Table 4.14 : Summary of average performance indices for classifiers from cross-validation	50
Table 4.15 : Prediction and classification algorithms ranking for PT.01 and PT.02 groups	51



LIST OF FIGURES

	<u>Page</u>
Figure 2.1 : Flowchart of research methodology for risk perception review.....	10
Figure 2.2 : Network visualization of co-occurrence of keywords	12
Figure 2.3 : Overlay network visualization of co-occurrence keywords	12
Figure 2.4 : Sankey diagram for profiles of publications about risk perception	13
Figure 2.5 : Main categories of machine learning techniques.....	18
Figure 4.1 : Standardized structure of risk register	28
Figure 4.2 : Groups depending on the project types	29
Figure 4.3 : Linear relationship between total cost and risk scores (PT.01).....	29
Figure 4.4 : Project based linear relationship between total cost and risk scores (PT.01).....	30
Figure 4.5 : Linear relationship between total cost and risk scores (PT.02).....	30
Figure 4.6 : Project based linear relationship between total cost and risk scores (PT.02).....	31
Figure 4.7 : Prediction results of regression models (PT.01).....	33
Figure 4.8 : Prediction results of regression models (PT.02).....	33
Figure 4.9 : Variability of R^2 different training and testing groups (PT.01).....	39
Figure 4.10 : Variability of MSE with different training and testing groups (PT.01)	40
Figure 4.11 : Variability of R^2 different training and testing groups (PT.02).....	40
Figure 4.12 : Variability of MSE with different training and testing groups (PT.02)	40
Figure 4.13 : Heatmap of confusion matrices for the PT.01 group	45
Figure 4.14 : Heatmap of confusion matrices for the PT.02 group	46
Figure 4.15 : Evaluation of classifier algorithms based on performance indices for the PT.01 group.....	48
Figure 4.16 : Evaluation of classifier algorithms based on performance indices for the PT.02 group.....	48
Figure 4.17 : Average accuracy score matrices of grid search.....	51
Figure A.1 : Risk breakdown structure and references.....	76
Figure A.2 : Flow chart of the research methodology	77



RISK-BASED COST ESTIMATION IN CONSTRUCTION BY EMPLOYING MACHINE LEARNING TECHNIQUES

SUMMARY

The construction industry is one of the sectors that experiences frequently cost overruns. Therefore, precise cost estimation for the completion of construction projects is essential. Several research studies focus on cost estimation, construction risk parameters, and their monetary effects. While they developed significant predictive models for project completion costs, they mainly focused on the initial construction phases. Estimating the completion cost accurately in the early phases of construction projects is critical to their success. However, cost overruns are almost inevitable due to the risks inherent in construction projects. Hence, the completion cost fluctuates throughout the execution phase and requires periodic updates. Limited studies have developed methodologies for estimating completion costs throughout the execution phase of projects. However, their models do not incorporate the implications of total risk scores. Further study is required to examine risk-based cost prediction for completion throughout the construction execution phase. There is a need for a prompt and user-friendly completion cost estimation model that accounts for fluctuating risk scores and their impact on the total cost during the execution phase. Machine learning (ML) techniques could address these requirements by providing effective methods for tackling dynamic systems.

The proposed approach includes ML prediction and classification models to estimate total completion cost and cost overrun percentage class of the project, respectively. Within the predictive models, six predictive algorithms were utilized, employing machine learning techniques. Analysis of the outputs revealed that polynomial regression yielded the most precise predictions for the supplied data from globally operating construction company. The classification approach aims to predict the cost overrun ratio classes of the completion cost according to the changes in the total risk scores at any time of the project. Six classification algorithms were utilized and validated by employing data points from a globally operating construction company. The performances of the algorithms were evaluated with validation and performance indices. The decision tree classifier surpassed other algorithms.

The main objective of this study is to provide a system that predicts the total completion cost and/or cost overrun percentage classification based on the total risk score of projects at any point of the execution phase. Although there are some research limitations, including risk perception, data gathering restrictions, and selecting proper ML algorithms upon data properties, this research improves the planning abilities of construction executives by providing completion cost and cost overrun ratio based on changing total risk scores, facilitating swift and simple assessments at any stage of a construction project's execution.



MAKİNE ÖĞRENİMİ TEKNİKLERİNİ KULLANARAK İNŞAATTA RİSK TABANLI MALİYET TAHMİNİ

ÖZET

İnşaat sektörü, sıklıkla maliyet aşımı sorunları yaşayan sektörlerden biridir. Bu nedenle, inşaat projelerinin tamamlanma maliyetinin doğru bir şekilde tahmin edilmesi büyük önem taşımaktadır. Ek olarak, pek çok sektörde olduğu gibi inşaat sektöründe de kalite standartları gereği risk analizleri yapılmakta ve bu analizler inşaat süresi boyunca güncellenerek devam etmektedir. Risk analizleri sonunca oluşma ihtimali öngörülen tüm risklerin olasılık ve etki değerleri belirlenmekte ve sonuç olarak risk skorları hesaplanmaktadır. Bu risk skorları her ne kadar bazı projelerde güvenlik tutarları olarak maliyetlere yansıtılmaya çalışılsa da genellikle inşaat tamamlanma maliyetlerine organik bir bağ ile bağlı değildir. Riskleri dikkate alarak maliyet tahmini yapmak, tüm inşaat projeleri için hayati bir konudur.

Birçok araştırma çalışması, maliyet tahmini, inşaat risk parametreleri ve bunların mali etkilerine odaklanmaktadır. Bu çalışmalar, proje tamamlanma maliyetleri için önemli tahmin modelleri geliştirmiş olmakla birlikte, genellikle inşaatın başlangıç aşamalarına ve ihale maliyeti üzerine yoğunlaşmaktadır. Ancak, projenin başlangıç aşamalarında ayrıntılı bilgilere ulaşmak zordur. Ancak, inşaat projelerinin erken aşamalarında tamamlanma maliyetini doğru tahmin etmek kritik olsa da, inşaat projelerinin doğasında bulunan riskler nedeniyle maliyet aşımı neredeyse kaçınılmazdır. Dolayısıyla, tamamlanma maliyeti, uygulama aşaması boyunca değişkenlik göstermekte ve periyodik güncellemeler gerektirmektedir.

Mevcut çalışmaların yalnızca sınırlı bir kısmı, projelerin uygulama aşaması boyunca tamamlanma maliyetini tahmin etmeye yönelik metodolojiler geliştirmiştir. Bunun temel nedenlerinden biri, proje uygulama aşamasına ilişkin yeterli verinin bulunmaması veya olan veriye ulaşmanın güçlüğüdür. Çünkü teklif aşaması maliyetleri ya da inşaat tamamlanma maliyetini paylaşmak proje ekibi tarafından daha mümkün iken, proje devam süresince periyodik olarak hesaplanan güncel tamamlanma maliyetini paylaşmak pek olası değildir. Varolan modeller de toplam risk skorlarının etkilerini içermemektedir. Bu çalışmanın önemi, gerçek yaşam verilerini kullanarak proje uygulama aşaması maliyetine odaklanması ve profesyonellere risk içeren bir maliyet tahmin modellerini makine öğrenmesi yöntemleri ile sunmasıdır.

Öte yandan, riskler proje boyunca değerlendirilmekte olup, risklerin olasılığı ve etkisi uygulama aşamasında değişim göstermekte ve tamamlanma maliyet aşım oranları bu değişen risk skorlarına bağlı olarak dalgalanmaktadır. Ayrıca, proje ilerledikçe bazı riskler ortaya çıkarken bazıları ortadan kalkmakta ve bu risklerin potansiyel maliyet aşımı etkilerini bireysel olarak öngörmek zorlaşmaktadır. Risk kayıtlarının, sözleşme maddelerinde ele alınmayan ve maliyet tahmin süreçlerinden dahil edilmeyen durumları içerdiği unutulmamalıdır.

İnşaatın uygulama aşaması boyunca risk temelli maliyet tahmini konusunun daha ayrıntılı incelenmesi başarılı bir maliyet tahmini için hem gerekli hem de faydalıdır. Uygulama aşamasında değişkenlik gösteren risk skorlarını ve bunların toplam maliyet üzerindeki etkilerini dikkate alan hızlı ve kullanıcı dostu bir tamamlanma maliyeti tahmin modeline ihtiyaç duyulmaktadır. Model, yeni biriken verileri periyodik olarak entegre edebilecek şekilde uyarlanabilir olmalıdır. Makine öğrenmesi teknikleri, dinamik sistemlere yönelik etkili çözümler sunarak bu gereksinimleri karşılayabilir.

Bu çalışmayı önceki risk bazlı maliyet tahmin çalışmalarından ayıran temel özellikler aşağıdaki şekilde özetlenebilir:

- Yapılan pek çok maliyet tahmin çalışmalarının aksine, sadece proje başlangıç ve ihale aşamasında değil, projenin yürütme aşamasından bitimine kadar uygulanabilecek bir model sunulmuştur. Bu model yine proje yürütme aşamasındayken üretilen toplam risk skoru ve toplam tamamlanma maliyeti verileri ile değerlendirilmiş ve geçerliliği bu tez kapsamında açıklanmıştır.
- Kalite standartları gereği pek çok sektörde olduğu gibi inşaat sektöründe de periyodik olarak üretilen risk değerlendirme raporlarının toplam risk skoru değerleri, sözleşmeler, ekler veya güvenlik katsayısı şeklinde dahil edilmeyen durumların özeti olarak inşaat tamamlanma maliyeti ile ilişkilendirilmiştir.
- İnşaat projelerinin yürütme aşamasının yoğunluğu ve hızı bilinmektedir. Bu dönemde yapılacak her türlü çalışmanın en kullanıcı dostu ve pratik şekilde sonuç vermesi beklenmektedir. Bu nedenle önerilen model çok az iş gücü gerektirmektedir. Bütçe planlamacıları ya da proje yöneticileri tarafından kolaylıkla uygulanabilecek şekilde tasarlanmıştır.
- Sunulan yöntem makine öğrenmesi tabanlıdır. Bunun ana sebebi istatistik yöntemleriyle elde edilebilecek ilişkisellik ve neden sonuçtan ziyade, tahmin yeteneği yüksek bir model sunma ihtiyacıdır. Makine öğrenmesi algoritmaları ile sunulan model, esnek ve dinamik olduğu gibi, proje yürütme aşamasındayken oluşan her gerçek verinin kolayca eklenmesi ile tahmin gücünü kuvvetlendirmektedir.

Bu çalışma kapsamında Önerilen yaklaşım, toplam tamamlanma maliyetini tahmin etmek için makine öğrenmesi tabanlı bir tahmin modelini ve maliyet aşımı yüzdesini sınıflandırmak için bir sınıflandırma modelini içermektedir. Tahmin modellerinde, altı farklı makine öğrenmesi algoritması kullanılmıştır. Bu algoritmalar, doğrusal regresyon 2. ve 4. derece polinom regresyonları, rassal orman, destek vektörü regresyonu ve karar ağacıdır. Sınıflandırma yaklaşımı ise, toplam risk skorlarındaki değişimlere bağlı olarak projenin herhangi bir zamanında maliyet aşımı oranı sınıfını tahmin etmeyi amaçlamaktadır ve bu aşamada da altı farklı sınıflandırma algoritması kullanılmıştır. Bu algoritmalar lojistik regresyon sınıflandırıcısı, K-en yakın komşu, destek vektörü sınıflandırıcısı, Naïve Bayes sınıflandırıcısı, karar ağacı sınıflandırıcısı ve rassal orman sınıflandırıcısıdır. Tahmin ve sınıflandırma modellerinin analiz sonuçları, küresel ölçekte faaliyet gösteren bir inşaat şirketinden elde edilen verilerle geçerleme ve performans ölçütleri kullanılarak değerlendirilmiş ve polinom regresyonun tahmin modellerinde karar ağacı sınıflandırıcısı da sınıflandırma modellerinde gerçeğe en yakın sonuçları vererek diğer algoritmalarından daha başarılı olduğu belirlenmiştir.

Bu çalışma kapsamında önerilen yöntemin bir takım avantajları ve dezavantajları bulunmaktadır. Bu yaklaşımın temel avantajı, projelerin erken aşamalarında kolaylıkla

uygulanabilmesidir. Bu durum, daha önce tamamlanmış projelerin toplam maliyet ve toplam risk skoru verilerinin kullanılmasıyla sağlanmaktadır. Ayrıca, güncellenmiş maliyet ve toplam risk skoru verileriyle, uygulama aşamasındaki projelere istenildiği zaman uyarlanabilir veya modifiye edilebilir. Önerilen yaklaşım, maliyet tahmin süreçlerinden hariç tutulan ve genellikle risk kayıtlarında değerlendirilen risklerin finansal etkilerini de kapsamaktadır.

Bununla birlikte, önerilen modelin uygulanması basittir, çünkü kırılımlar yerine toplam maliyet ve toplam risk skorları kullanılmaktadır. toplam maliyet ve toplam risk skorlarının alt bileşenler yerine kullanılmasının avantajları şu şekilde özetlenebilir:

- Kolay uygulanabilirlik: Her proje için tutulan risk kayıtlarını kullanarak kümülatif risk skorlarını hesaplamak önemli bir zorluk teşkil etmemektedir. Kırılımlar yerine toplam değerler üzerinden gitmek uygulama kolaylığı sağlamaktadır.
- Maliyet dağılımı standartlaştırma gereksiniminin ortadan kalkması: Toplam maliyetin bir faktör olarak dahil edilmesi, maliyet bileşenlerinin standartlaştırılmasını zorunlu kılmamaktadır. Farklı projelerde farklı alt kırılımlar kullanılabilen, direkt ve endirekt harcama ayrımlarını projeler arası standartlaştırma zorlaşabilmektedir. Önerilen yöntemle bu standartlaştırmaya gerek kalmadan sadece toplam tamamlanma maliyetine konsantre olmak yetmektedir.
- Sonraki projeler için sürdürülebilir sonuçlar: Toplam maliyet ile toplam risk faktörü arasındaki tutarlı bir ilişki kurulması, benzer projelere kolaylıkla uygulanabilecek daha anlamlı sonuçlar elde edilmesini mümkün kılmaktadır.

Yaklaşımın bir diğer avantajı, maliyet aşımı oranı sınıflandırmalarının, maliyet artışlarının ciddiyetini net bir şekilde anlamayı sağlamasıdır. Buna ek olarak, yöneticiler, toplam maliyetin beklenen yüzde artışına bağlı olarak bir risk azaltma eylem planı geliştirebilirler. Risk skorlarına dayalı toplam tamamlanma maliyetinin tahmin edilmesi, projenin uygulama aşamasında risk odaklı öngörülerin hızlı bir şekilde üretilmesini de sağlamaktadır.

Öte yandan önerilen yaklaşım çerçevesinde iyileştirilme potansiyeli taşıyan bazı sınırlamalar göz önünde bulundurulabilir. İlk olarak, toplam risk skorunun hesaplanmasını projeden projeye değiştirebilecek risk algısı farklılıkları bulunmaktadır. Ayrıca, maliyet aşımı oranı, projenin özelliklerine bağlı olarak farklı şekilde kategorize edilebilir. Benzer şekilde, her şirket, proje türü ve organizasyonel dinamiklerden etkilenerek risk olasılığı ve etkisini farklı ölçeklerde değerlendirebilir ve verilerinin doğasına bağlı olarak farklı makine öğrenmesi algoritmaları kullanabilir.

Bununla birlikte, her proje, risk ve maliyet verilerini güncelledikçe eşdeğer miktarda veri üretmektedir. Dolayısıyla, mevcut veri miktarı açısından bir kısıtlama söz konusu olabilir. Ayrıca, uygulama aşamasında bir inşaat projesinin risk ve maliyet güncellemelerini almak, ihale maliyeti veya proje sonunda tamamlanma maliyeti verilerini elde etmekten daha zordur. Son olarak, bu çalışmada veriler üzerinde en iyi performans gösteren algoritmalar ilgili veri özelliklerine göre iyi tahmin sonuçları vermiştir. Bu nedenle, algoritma ve parametre seçimleri, verinin özgül niteliklerine bağlı olarak belirlenmelidir ve güncel çalışmalarda algoritmaların veriyi ezberleme durumundan kaçınılmalıdır.

Analizler, en ilgili bileşenleri incelemek amacıyla maliyet ve/veya risk kırılım yapıları kullanılarak geliştirilebilir. Ayrıca, büyük veri kümelerini daha etkin bir şekilde

işleyebilen alternatif makine öğrenmesi yöntemleriyle geniş veri setleri analiz edilebilir. Bunun yanı sıra, konut ve ağır altyapı projeleri dışında farklı inşaat proje türlerinin incelenmesi faydalı olabilir. Bu sınırlamalar, farklı şirketler arasında ve alternatif makine öğrenmesi algoritmaları kullanıldığında sonuçların değişmesine neden olabilmektedir.

Sonuç olarak bu çalışmanın temel amacı, projelerin uygulama aşamasının herhangi bir noktasında toplam risk skoruna dayalı olarak toplam tamamlanma maliyetini ve/veya maliyet aşımı yüzdesini tahmin eden bir sistem geliştirmektir. Risk algısı, veri toplama kısıtları ve veri özelliklerine uygun makine öğrenmesi algoritmalarının seçimi gibi bazı araştırma sınırlamaları bulunmasına rağmen, bu çalışma, değişen toplam risk skorlarına dayalı olarak tamamlanma maliyetini ve maliyet aşımı oranını tahmin ederek inşaat yöneticilerinin planlama yeteneklerini geliştirmekte ve projelerin uygulama sürecinin herhangi bir aşamasında hızlı ve kolay değerlendirmeler yapmalarına ve risk azaltma eylem planı geliştirmelerine olanak tanımaktadır.



1. INTRODUCTION

1.1 Problem Statement and Significance of Thesis

The main constraints of a project are time, quality and cost. According to Baratta (2006), most of the projects are completed out of schedule, with cost overruns and having a tendency under the scope requirements, and this is an inconsistency with triple constraints of the project. It is obvious that these constraints are also valid for construction projects and most of the construction projects experience cost overruns with various reasons (Aziz, 2013; Doloi, 2013; Flyvbjerg et al., 2018; Johnson and Babu, 2020; Odeck, 2004; Shane et al., 2009). Cost performance of a construction project is one of the most important indicators of its success (Ahiaga-Dagbui and Smith, 2014). The cost overrun is defined as the exceeded final cost when compared to initial estimate (Avotos, 1983; Aziz, 2013). If the project team initially projected a cost and subsequently exceeded that estimate, a cost overrun occurs. It is obvious that, Cost estimating accuracy is crucial for the success of a construction project (Kim et al., 2005).

Cost estimation is one of the most comprehensively investigated topics in literature. Numerous methods exist in literature for estimating construction costs. One of the popular methods is regression model (Chan and Park, 2005; Lowe et al., 2006; Ofori-Boadu, 2015; Thomas and Thomas, 2016; Türkel and Ergen, 2016). Another mostly used model is neural network (Hegazy and Ayed, 1998; Kim et al., 2005; Kim et al., 2004; Matel et al., 2022; Petroutsatou et al., 2012). Cased based reasoning is an alternative used method to reach the cost to complete of the project (Ji et al., 2011; Kim et al., 2004; Kim and Kim, 2010; Leśniak and Zima, 2018). Previous tender cost data of similar projects were used in regression, neural networks and other methods. However, it is known that tender price is generally the lowest price to award the bid (Lowe et al., 2006). It is needed to use execution phase data to determine the real cost of the project. Additionally, most of the estimation methods present single value cost of the construction, however, single value estimation is insufficient to understand and

evaluate potential risks (Hull, 1990; Rezaie et al., 2007). In the frame of this thesis study, range estimation is going to be formed to include the effects of risks with classification methodology.

Furthermore, cash flow always draws all the attention due to giving information about the near future more exactly in terms of liquidity. In the frame of this research accrual cost is considered instead of cash flow since the main concentration is cost to complete of the project and updated value of this through execution phase. Accrual cost gives profitability and the revision depending on the changes throughout the project.

Moreover, cost estimation is a significant indicator of the success of construction projects (Kim et al., 2004). However, estimating the cost of construction is a risky aspect of the process due to the dynamic nature of the construction industry. Most construction projects exceed their estimated cost for various reasons, as demonstrated by multiple studies (Alhammadi et al., 2024; Alshihri et al., 2022; Chadee et al., 2023; Flyvbjerg et al., 2018; Shane et al., 2009).

Certainly, in the construction industry there are a lot of situations which involve unknown, unexpected, unpredictable factors that are directly affecting the project's main objectives, cost, schedule and quality (Akintoye and MacLeod, 1997). It is possible that these unforeseeable occurrences will have either a beneficial or negative impact on the goal of a project. The risk analysis process and risk management are vital issues for the construction industry and risk management is also essential to construction activities in minimizing losses and enhancing profitability (Fong, 1987). Although risk analysis significance is indicated in the literature, these techniques are infrequently used in the construction sector (Akintoye and MacLeod, 1997). However, the ISO 9001:2015 Quality Management System requires implementing a "Risk management" approach in all business processes (2015). All companies that receive a quality certificate in accordance with this standard have to evaluate risks and opportunities within their business processes based on total quality management principles.

Cost estimation considering risks is a vital issue for all construction projects. Studies performed about cost estimation are mainly on tender cost (i.e. early stage, conceptual cost). However, in the initial stages of the project, it is difficult to reach detailed information and subjective experience of the construction authorities, which do not

have scientific basis, determine the cost of the project (Chou, 2011). Execution phase cost estimation generally is not concentrated in studies. The possible main reason of not studying on execution phase cost of the project is not having data on this phase. The significance of this study is to concentrate on the execution cost of the project by using real life data and proposing a risk embedded cost estimation model to the professionals. Since the estimated completion cost of the project incorporates the anticipated risks associated with the construction activities (Gurcanli et al., 2015; Liu and Napier, 2010; Sohrabi and Noorzai, 2024). On the other hand, risks are evaluated throughout the projects, and the probability and impact of risks evolve during the project execution phase, and the completion cost overrun ratios also fluctuate based on these changing risk scores. Furthermore, as the project progresses, certain risks arise while others dissipate, and it is challenging to foresee the potential cost overrun effects of these risks individually. It is essential to recognize that risk registers include unexpected cases not addressed in contract clauses and excluded from cost estimation procedures. Consequently, it is essential and beneficial to develop an approach for estimating risk-based cost models throughout the execution phase of construction projects to achieve successful cost estimation. The proposed approach should be practical and capable of producing results quickly due to the high workflow speed in the execution period of a construction project. Since the estimations are mainly conducted monthly and within a limited duration, the model must be adaptive and integrate new data that accumulates periodically. This user-friendly approach can ensure executives evaluate the potential completion costs at any stage of the project. Moreover, they can develop a mitigation action plan depending on these estimations.

Additionally, although extensive research on cost estimation during the tender or preliminary stages is highly beneficial, there is a need for an approach applicable to the execution phase that can be simply and promptly updated at any moment. Furthermore, assessing the completion cost and cost overrun ratio based on the total cost and risk scores, rather than breakdown components, facilitates the estimation process in the execution phase of a construction project.

Finally, the data for the case study of this research is going to be used from real-world projects. The tender prices or completion cost of the projects are reachable from some databases of firms. However, the change through project, especially in execution

phase, is a black box of the firms. Using real-world project cost values through the execution phase of the project is another strong aspect of this research.

The main objective of this research is to provide a systematic approach for determining the construction cost and cost overrun ratio class by considering the assessment of the total risk score throughout the execution phase. In this context, a proposed ML-based prediction and classification approaches were implemented on construction projects obtained from a globally operating construction company for practical application.

The thesis framework is outlined according to the flow as below:

- Literature review: This section of the thesis provides a comprehensive overview of literature about risks in construction projects, risk perception, cost estimation methods in construction projects, cost overrun in construction projects, and machine learning techniques.
- Research methodology: This section includes data gathering procedures for risk register, cost estimations of the project, and grouping of data. Subsequently, the methodologies for machine learning regression and classification modeling were clarified, and the final findings were presented.
- Implementation of the proposed approach: This section details the execution of the proposed strategy through a case study based on data provided by a globally operating construction firm.
- Discussion: This section summarizes the benefits of the proposed approach, limitations of study, and recommendations for future research.
- Conclusion: The results of the study were summarized in this section.

2. LITERATURE REVIEW

Within the framework of this thesis, literature review is provided in four main sections: (1) risks in construction projects, (2) risk perception, and (3) cost estimation methods in construction projects, and (4) cost overrun in construction projects.

2.1 Risks in construction projects

The risk is defined as “An uncertain event or condition, that if it occurs, has a positive or negative effect on a project's objective.” in PMBOK Guide (2017). Also defined as “A convenient way of describing an unknown state that may occur in the future” by Weaver (2008). Bufaied (1987) describes risk in connection to construction as a factor in the execution of a construction project whose unpredictability results in uncertainty about the project's eventual cost, quality and duration. Moreover, risk is defined in ISO 31000:2018 as “effect of uncertainty on objectives” (2018). To sum up, risk could be defined as an event or condition which is uncertain and with known probability, if it occurs, has progressive or destructive effect on project objectives (i.e. time, cost, quality etc.

Managing risk is an important issue for all relevant groups in construction and it is crucial throughout project execution period. As expected, the uncertainties are high in the beginning of the project therefore the risks are at the highest level (Smith et al., 2014). In the beginning of the project main objectives (budget, schedule and quality targets) are determined and evaluating risks in tendering period provides a crucial information about the project. However, it should not be forgotten that the construction sector includes many risks through execution phase, too. Therefore, risk management should be an ongoing and essential part of the whole period of construction.

In literature there are numerous studies related to construction risks. Depending on the academic databases, by searching “construction risks” and selecting top ranking papers on Scopus, it seemed that researchers grouped the construction risks in various ways.

Tummala and Burchett (1999) developed a risk management model for transmission line projects with risks related to cost items of the project. Long et al. (2004) highlighted the most common and general problems in construction. They identified 62 problems and grouped into 7 main titles as problems related to financier, owner, contractor, consultant, project attributes, coordination and environment. Depending on the questionnaire results, non-correct time estimation, clearance of the site slowly and severe change orders are the most influenced problems to the construction. Oztas and Okmen (2004) reviewed the literature and studied a model for schedule and cost risks in design-build projects by using Monte Carlo Analysis. They used 14 risk items and by using them probabilistic cost and schedule data of the project in tender phase is formed. Yildiz et al. (2014) grouped 49 construction risks under country risks, complexity of projects, uncertainty regarding geological issues, rigid requirements, contractual problems, inadequate engineering, incompetence of the client, undesirable site conditions, inexperienced contractor, lack of resources by contractor, not proper management by contractor. Chan and Au (2007) investigate the construction risk in terms of weather. Abdul-Rahman et al. (2013) applied a fuzzy synthetic model to assess the impact of construction risk factors. Zhao et al. (2016) used the fuzzy synthetic evaluation approach to create a model for risk assessment for green buildings. They categorized 28 risk items into 11 main risk titles by means of a questionnaire with professionals in the construction sector. Non-correct cost estimation and cost overrun risks are the most crucial risk items in the study. Risks in construction are active, complex, and interconnected systems (Etemadinia and Tavakolan, 2021), encountered at every level of the process (Ismail et al., 2021) and arises from the ephemeral nature of the industry (Do and Lam, 2019). Shaktawat and Vathera (2021) reviewed risk management on hydropower projects. They concentrated on various risk assessment methodologies and suggested sensitivity analysis for evaluating major risk factors. They categorized risks under social, environmental, political & legal, economic & financial, technological, construction related and safety. Furthermore, safety concerns in construction were analyzed using stochastic process modeling (Zhang et al., 2021).

There are several key influential studies about construction risks. The articles are summarized in order of their citation rates, beginning with the highest. It is noticed that the studies on determining and modelling risk factors on construction projects

have received significant attention. Zou et al. (2007) investigated the main risk variables in construction projects in China. They identify the primary risk factors and rank them based on their impact on project objectives. A questionnaire was conducted, and based on the findings, 25 critical risks for the success of construction projects were identified. The risks were compared with a similar study conducted in Australia. Furthermore, they explained strategies of risk management for the project's stakeholders. Another questionnaire was conducted by El-Sayegh (2008) in the United Arab Emirates among construction experts about substantial risks in constructions. The findings indicate that economic, labor, material, scheduling, and design risks are crucial for construction success, whereas cultural, social, and political risks are comparatively insignificant. Aminbakhsh et al. (2013) developed a comprehensive approach for prioritizing safety risks. Their objective was to implement a logical budget and planning process in construction, including proper safety considerations. Through their investigation, the construction specialists were able to formulate investment plans for appropriate safety measures. Dikmen et al. (2007) proposed a fuzzy risk assessment model to evaluate cost overrun risks in international construction projects. They developed a tool for the tender stage to assess risks in international construction projects, enabling the evaluation of cost overrun risks, and the model was validated on an actual project. The tool includes company objectives and can be adapted to different organizations based on their own goals. Taroun (2014) conducted an examination of construction risk modeling and assessment by analyzing relevant publications from 1983 to 2012. The results indicated a shift in risk perception towards project attributes, and a cost-based analytical tool was recommended to connect theory and practice in risk assessment. Shen et al. (2006) focused on construction risks and their management in public-private partnership projects in Hong Kong. The primary risk factors for the public sector are identified as site risks, legal risks, and political risks, but for the private sector, they encompass design, construction, operational, and industrial risks, according to the literature. Luu et al. (2009) analyzed scheduling risks in construction projects employing Bayesian belief networks. They conducted a questionnaire and interview to ascertain risk factors and their interrelations. The model was validated through two case studies. The findings indicate that financial risks, client and contractor risks, and material risks are the primary factors contributing to delays in construction projects in Vietnam. There are a number of studies concentrated on safety risks in construction projects. Antwi-Afari et al. (2017) and Wang et al. (2015)

worked on ergonomic and musculoskeletal disorders and Sousa et al. (2014) studied risk management for health research and occupational safety. Moreover, Ding et al. (2016) studied construction risk knowledge management in a BIM environment utilizing ontology and semantic web technology. A prototype system was developed for this aim and validated using a case application. A questionnaire was administered to assess both the advantages and disadvantages of this approach. The findings indicate that the suggested approach may improve risk and safety management in construction projects. The review studies also receive attention. Siraj and Fayek (2019) performed a comprehensive review of common construction risks. They aggregated risks into ten categories based on the analyzed literature. Kuo and Lu (2013) investigated the improvement of risk assessment by the application of a fuzzy multiple criteria decision-making approach in urban construction projects. They found 20 risk variables from the literature and stated that their approach is advantageous for identifying substantial risks and evaluating the overall project risks. Xia et al. (2018) conducted a review on risk and stakeholder management in the construction industry to strengthen the connection between these fields and mitigate stakeholder management risks.

Certain studies concentrated on mathematical models for risk assessment or specifically addressed certain risks, such as HSE issues (Aminbakhsh et al., 2013; Antwi-Afari et al., 2017; Carpio-de Los Pinos et al., 2021; Rowlinson et al., 2014), risk perception (Trillo-Cabello et al., 2021), collusion risk (Shan et al., 2018), communication risks (Ceric, 2014), and BIM models to visualize construction risks (Kang et al., 2013). On the other hand, most of the studies address more than one construction risk. These studies were summarized in Figure A.1 in Appendix A with a proposed risk breakdown structure with references.

2.2 Risk perception

Construction projects are characterized by numerous uncertainties that arise at various stages of their development. These variables provide potential hazards, and risk is a perceived phenomenon. Risk perception, as defined by Douglas and Wildavsky (1990), refers to the capacity to predict and explain which specific individuals will consider certain possible risks to be the most frightening. There are numerous studies in the literature concerning risk perception in construction. In a questionnaire survey, Thomas et al. (2003) examined the perception of risk concerning the management,

allocation, and mitigation of construction hazards for highway infrastructure projects in India. They resulted in different risk perceptions between parties in risk management. Hallowell (2010) examined safety risk perception by comparing the risk perception of employees and supervisors. Tixier et al. (2014) investigated the relationship between risk perception, psychological factors, and risk-taking behavior among individuals experiencing various emotional states. Based on the outcomes, the negative group exhibited a greater perception of hazards than the positive group. Namian et al. (2016) emphasized the need to enhance the accurate perception of safety risks and recognition of hazards. The results indicated that those with a higher level of training exhibited a more remarkable ability to perceive safety risks and recognize hazards. Wang et al. (2016) researched the big five personality dimensions and risk perception of construction project managers in China. Chan et al. (2021) examined the risk perception of safety climate among workers and managers and identified the specific categories for both groups. Pooladvand and Hasanzadeh (2023) focused on construction workers' physiological and cognitive reactions to risks. They found that high-stress levels had a negative impact on the workers' ability to perceive risks accurately. Although numerous studies examined risk perception in construction from various perspectives and in diverse contexts, limited studies have specifically focused on risk perception related phenomena in the construction industry. This study performs a bibliometric literature review to examine the publishing trend of risk perception in the construction industry.

There are various literature review strategies for gathering current knowledge on a specific topic. The systematic literature review is a method used to analyze and summarize data reproducibly (Linnenluecke et al., 2020). The research employs systematic mapping and bibliometric analysis to arrange and examine risk perception in the construction industry.

A systematic review involves a thorough search of specific databases. Scopus is a comprehensive database and one of the most esteemed academic search engines (Calik et al., 2023; Linnenluecke et al., 2020).

The review methodology for risk perception in construction management is illustrated in Figure 2.1.

The research subject pertains to the role of risk perception within the construction management literature. Construction risks encompass not only quantitative models but also incorporate social factors. This study considers the latest research topics and trends related to and researched the “risk perception” concept in the construction industry.

Step 1. Research Question

Review "risk perception" in construction industry



Step 2. Data Gathering

Determination of database : Scopus
Keywords (T/A/K : “risk perception” AND construction)
2874 publications



Step 3. Data Elimination

Abstract include "risk perception" (548 publications)
Language : English (500 publications)
Document Type : Article, Conference Paper, Book Chapter (470 publications)
Eliminating the non-civil engineering sources (96 publications)



Step 4. Mapping Process

Using extracting data for visualization and systematic mapping
Software : Excel, VOSviewer & SankeyMATIC

Figure 2.1 : Flowchart of research methodology for risk perception review.

Scopus is extensively utilized for literature reviews due to its high dependability and efficiency. T/A/K (title, abstract, and keyword search) was employed to obtain all associated documents from the Scopus database. The keywords for the study were “risk perception” AND “construction.” A total of 2874 publications were retrieved.

Document elimination was performed in order to select related literature. Publications containing any of the specified features were excluded from this study: (1) not including the “risk perception” concept in the abstract part, (2) original documents that were not in English, (3) document types were out of an article, conference paper, and book chapter, (4) not being directly related to the construction industry. After elimination, 96 publications remained, including 60 journal articles, 34 conference papers, and two book chapters.

VOSviewer is a software application designed to create and display bibliometric networks by analyzing relationships such as bibliographic coupling, co-authorship, co-occurrence, citations, or co-citation (Van Eck and Waltman, 2014). In order to identify the trends related to the research topic, keyword co-occurrence analyses were conducted in VOSviewer (ver. 1.6.20). The SankeyMATIC tool was utilized to visually represent the trends in publications.

The density visualization of the co-occurrence of keywords from the bibliographic portfolio was illustrated in Figure 2.2. The text contains significant mentions of terminology such as "risk perception" with "risk management," "project management," and "risk assessment," along with their conjunction with the terms "occupational risks," "hazards," and "safety engineering," with four co-occurring keyword clusters.

The red cluster was characterized as "Construction Risk Management with Risk Perception." The group has 21 keywords, primarily focused on risk management, evaluation, and analysis in construction projects and building design. The Green cluster was identified as "Project Safety and Risk Perception in the Construction Industry," consisting of 14 keywords. It primarily focuses on safety engineering, safety management, safety culture, health risks, and human resource management. The blue cluster consists of 13 elements, focusing on occupational risks, hazard identification and recognition, and personnel and safety training for construction workers. This cluster is specifically named "Construction Workers and Risk Perception." The last cluster is yellow and comprises five keywords, with the primary ones being accident and disaster prevention and risk-taking behaviors. The yellow cluster is categorized as "Prevention Action and Risk Perception in Construction".

The distribution of trend study subjects in construction with risk perception is depicted in Figure 2.3. Based on the data, it appears that the focus has shifted from "Construction Risk Management with Risk Perception" to primarily "Construction Workers and Risk Perception" and "Project Safety and Risk Perception in the Construction Industry". These phenomena encompass identifying and recognizing hazards and occupational risks, ensuring safety in construction activities, and managing human resources in the construction industry.

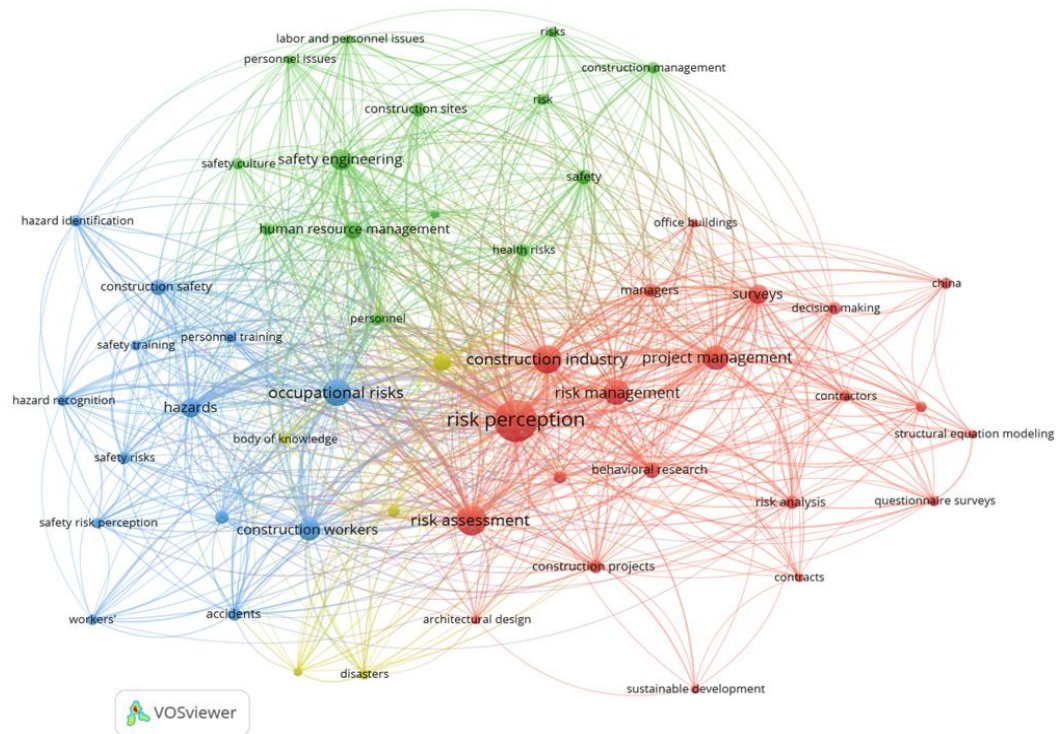


Figure 2.2 : Network visualization of co-occurrence of keywords.

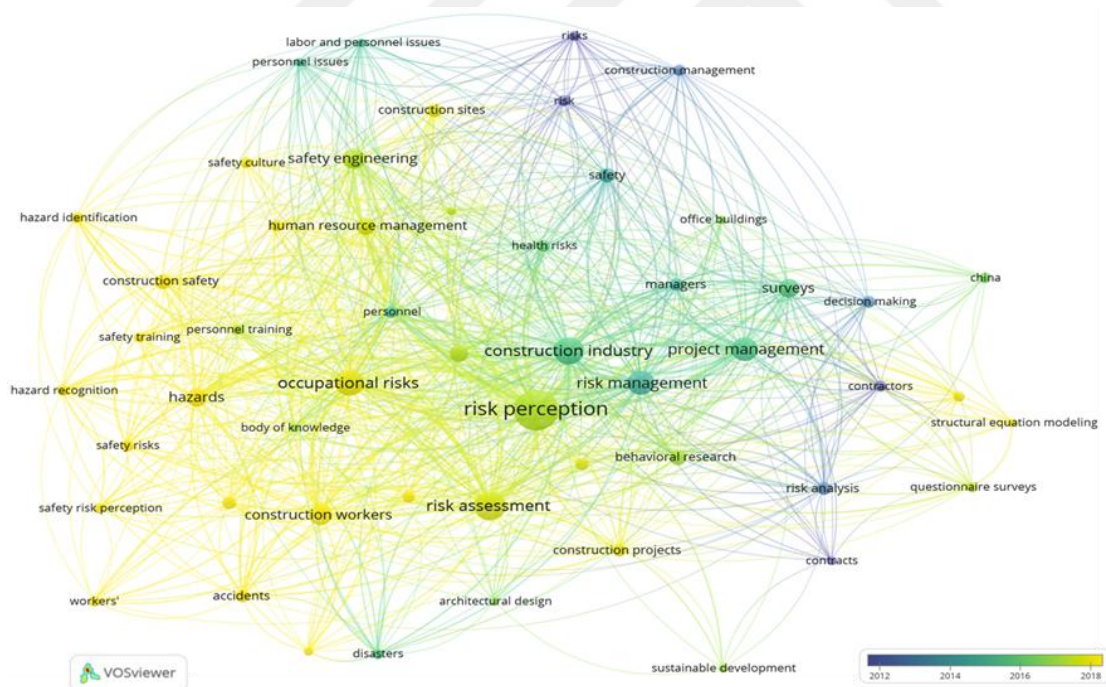


Figure 2.3 : Overlay network visualization of co-occurrence keywords.

In order to identify the profile of the publications, a Sankey diagram is employed, which is a visualization tool that simultaneously represents numerous properties (Linnenluecke et al., 2020). Figure 2.4 depicts the distribution of publication sources and the years of the studies using a Sankey diagram. 60 articles published in 19

journals, 34 conference papers in 8 conferences and 2 book chapters were analyzed. Figure 2.4 indicates that most articles discussing the notion of risk perception in the construction industry are published in Journal of Construction Engineering and Management, Construction Management and Economics, International Journal of Project Management and Sustainability. Additionally, Construction Research Congress, International Structural Engineering and Construction Conference, and Association of Researchers in Construction Management Conference are the preeminent conferences concerning this subject matter. Furthermore, there has been a growing inclination towards researching risk perception in the construction industry in recent years.

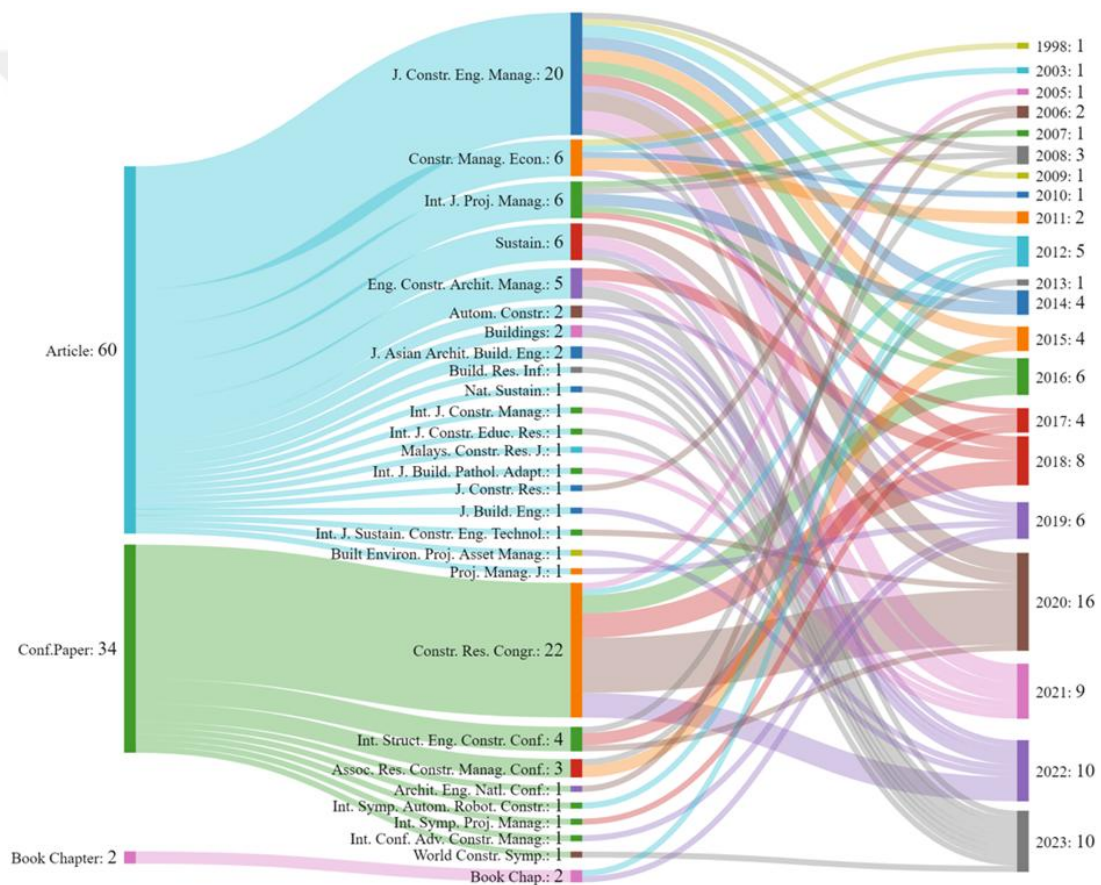


Figure 2.4 : Sankey diagram for profiles of publications about risk perception.

To sum up, risk perception is an essential phenomenon in the construction industry. The bibliometric literature review study investigated the related research topic and publication trends for risk perception in construction. 96 publications were reviewed, and their co-occurrence and bibliometric analysis were performed. Four primary research clusters were identified, which mainly included risk management, safety, construction workers, and prevention actions concepts with risk perception in the

construction research topic. Furthermore, research on risk perception in the construction industry has shown that occupational and project safety, construction workers, and human resources are current trending research areas of investigation. Finally, this study's conclusions determine the growing interest in researching risk perception within the construction industry. The findings of this study can offer researchers a research direction and new areas for further investigation. To enhance the proposed study, it is recommended to conduct searches in additional databases with Scopus. Furthermore, it would be beneficial to include publications in languages other than English. Additionally, forming distinct clusters will aid in identifying and developing research subjects related to risk perception in the construction industry.

2.3 Cost estimation methods in construction projects

Numerous methodologies exist in the literature for estimating construction costs. Certain studies focus on particular construction costs rather than the total completion cost. Chang and Shinozuka (1996) calculated the life-cycle cost of bridge construction by aggregating both anticipated and potential unforeseen expenses by users. Adeli and Wu (1998) predicted the costs associated with reinforced concrete pavements in highway projects using a neural network. Ma et al. (2013) and Lee et al. (2014) utilized building information modeling (BIM) for cost estimation to determine the concrete costs of a six-story building and the tiling costs in international construction projects, respectively. Gurcanli et al. (2015) assessed the safety costs of residential buildings during the initial phases of the project utilizing activity-based costing.

Moreover, numerous methods exist to estimate the total cost of the construction projects. The researchers mainly used neural networks (Alshboul et al., 2022; Antoniou et al., 2023; Das et al., 2024; Elhegazy et al., 2022; Hegazy and Ayed, 1998; Kantianis, 2022; Kim et al., 2005; Kim et al., 2004; Matel et al., 2022; Petroutsatou et al., 2012; Rafiei and Adeli, 2018; R. Wang et al., 2022), different types of regression models (Chan and Park, 2005; Kantianis, 2022; Kim et al., 2004; Lowe et al., 2006; Ofori-Boadu, 2015; Shoar et al., 2022; Thomas and Thomas, 2016; Türkel and Ergen, 2016), case-based reasoning (Ji et al., 2011, 2012; Kim et al., 2004; Kim and Kim, 2010; Leśniak and Zima, 2018), BIM-based cost estimation (Fazeli et al., 2021; Forgues et al., 2012; Park and Yun, 2023; Yang et al., 2022), support vector machine and random forest (Cheng and Hoang, 2014; Juszczuk, 2020; Park et al., 2022; Shoar

et al., 2022), genetic algorithm (Hegazy and Ayed, 1998; Kim et al., 2005; Kim and Kim, 2010), and Monte Carlo simulation (Chou, 2011; Ökmen and Öztaş, 2010) to estimate the total construction cost of a project, mainly in early phases. Recent studies indicate that ML approaches have been progressively employed in the construction industry for cost estimation. Alshboul et al. (2022) employed ML algorithms, including extreme gradient boosting, deep neural network, and random forest, to estimate green building costs. In the research, extreme gradient boosting gave the highest accuracy, followed by deep neural network and random forest. Park et al. (2022) employed a combination of algorithms, including random forest, support vector machine, and CatBoosting, to estimate the total cost of building construction projects in the early phases. Their suggested model demonstrated superior predictive performance compared with the individual models. Park and Yun (2023) studied the total cost of a building project in the design phase, utilizing a BIM model and ML. They determined that precise construction cost estimation is achievable by integrating the general characteristics of the building with BIM attributes. Furthermore, Das et al. (2024) employed ML methods, including artificial neural networks, natural gradient boosting, and linear regression, to forecast construction costs, with natural gradient boosting exhibiting superior performance in predicting the total construction cost.

Additionally, it is noteworthy that most of the research focused on the early phases of the construction process. Neural networks (Hegazy and Ayed, 1998; Kantianis, 2022; Kim et al., 2005; Kim et al., 2004; Mhady et al., 2024; Rafiei and Adeli, 2018), regression models (Kantianis, 2022; Kim et al., 2004; Ofori-Boadu, 2015; Shoar et al., 2022; Türkel and Ergen, 2016), and BIM-based cost estimation (Fazeli et al., 2021; Park and Yun, 2023) were predominantly employed methods mainly in residential structures and high-rise projects to estimate total costs during preliminary stages. Although the significant value of these studies in guiding future research, there is a need for estimating models that can be updated efficiently and swiftly at any moment in the execution phase of a construction project. Estimating the completion cost of a construction project is crucial not just in the design, tender, or commencement phase but also during the execution phase. While most of the project's risks identified in contract conditions were initially included in the completion cost, unexpected risks are regularly assessed in risk registers during the project's execution phase and should be reflected in the total completion cost. There is limited research on estimating total

completion cost through the execution phase of a construction project. Cheng and Hoang (2014) researched the integration of ML methods to achieve precise cost estimation and measurement of uncertainty. Their model could be applicable in the execution phase and showed higher performance in estimating compared with conventional methods, with a high level of reliability and usability in construction management. In addition, Mhady et al. (2024) conducted a study on fuzzy inference systems and artificial neural networks, utilizing data from residential projects to accurately estimate costs during the early and late phases of construction. There is a requirement for more approaches that allow contractors to foresee the final cost of completion promptly while executing the projects. ML is highly valuable in this area due to its demonstrated effectiveness and integration of many approaches. Within the scope of this study, ML classification algorithms were employed to efficiently and rapidly update the cost overrun ratios of total cost based on the total risk scores at any point during project execution.

2.4 Cost overrun in construction projects

Cost overruns frequently occur in construction projects. Several research studies have focused on the causes of cost overruns. Mansfield et al. (1994) conducted a questionnaire on Nigerian construction projects. They determined that finance, contract management, fluctuating site conditions, and material shortages were the primary factors driving cost overruns. Another questionnaire was performed in Indonesia about cost overruns on high-rise projects. Results indicate that increases in material costs, project complexity, and poor estimation were the primary factors contributing to cost overruns (Kaming et al., 1997). Frimpong et al. (2003) used a questionnaire to investigate groundwater construction projects in Ghana in terms of delays and cost overruns. The results reveal that payment difficulties and poor contractor management were critical issues for these concerns. Bin Seddeeq et al. (2019) focused on cost overruns in the Oil and Gas Construction Industry in Saudi Arabia. The survey results indicate that design and scope changes, inadequate planning, errors in design, unclear scope comprehension, and underestimation of costs and schedules are the primary causes of time and cost overruns. Alshihri et al. (2022) conducted a comparable study on building projects in the same region. They determined that change orders and selecting the lowest bidder are the main factors

contributing to cost overruns. Xie et al. (2022) performed a questionnaire about the same topic in China and found that market price and currency fluctuations, insufficient risk and contract management, design modifications, and constraints related to project locations were revealed as the most significant factors.

Moreover, numerous studies have developed a strategy for managing cost overruns in construction projects. Dikmen et al. (2007) developed a fuzzy risk assessment methodology to identify cost overrun risks in international construction projects. They identified and modeled risks and utilized fuzzy rules to produce a final risk rating in their methodology. Using random forest analysis, Shoar et al. (2022) assessed the cost overrun for high-rise residential structures. Based on the comparative analysis of prediction performance, random forests outperformed support vector regression and multiple linear regression. Chadee et al. (2023) employed a sequential mixed-methods strategy, using reference class forecasting to improve the precision of cost estimation in public sector social housing projects, and their model effectively enhances accuracy in forecasting cost overruns. Alhammadi et al. (2024) utilized the structural equation model to investigate the correlations between the causes of cost overruns and their mitigation strategies. Inadequate contract management and inaccurate cost estimation were the primary factors contributing to cost overruns. They recommended prioritizing mitigation strategies to reduce cost overruns. Hussain et al. (2024) conducted a study on reducing cost overruns in rail projects by investigating the 5D-BIM framework. Results indicate that implementing the 5D-BIM framework mitigates cost overruns by delivering accurate cost estimations and dynamic financial control throughout the project lifecycle. Finally, Sohrabi and Noorzai (2024) examined cost overrun estimation for water-related projects utilizing case-based reasoning based on varied risks and uncertainties in ML models. Their model has demonstrated a favorable correlation with ML algorithms for predicting cost overruns.

According to the reviewed literature, contract management, financial concerns, and risks associated with construction sites and materials are significant factors contributing to cost overruns in construction projects. Although these factors are crucial concerning cost overrun, there is a need to implement risks not covered in contract clauses or not included in cost estimation processes. Changes in the scope of work or risks arising related to site conditions are mainly included in cost estimation processes when they are identified. However, other risks were periodically reported

into risk registers from all departments to project executives as required by ISO 9001:2015, and the effect of this risk score is mainly not reflected in the estimated completion cost (2015). Therefore, in the framework of this study, a completion cost overrun ratio estimation approach is proposed based on the risk score reported in risk registers.

2.5 Machine learning techniques

Machine Learning (ML) algorithms can be categorized into three primary classifications: supervised learning, unsupervised learning, and reinforcement learning as illustrated in Figure 2.5.

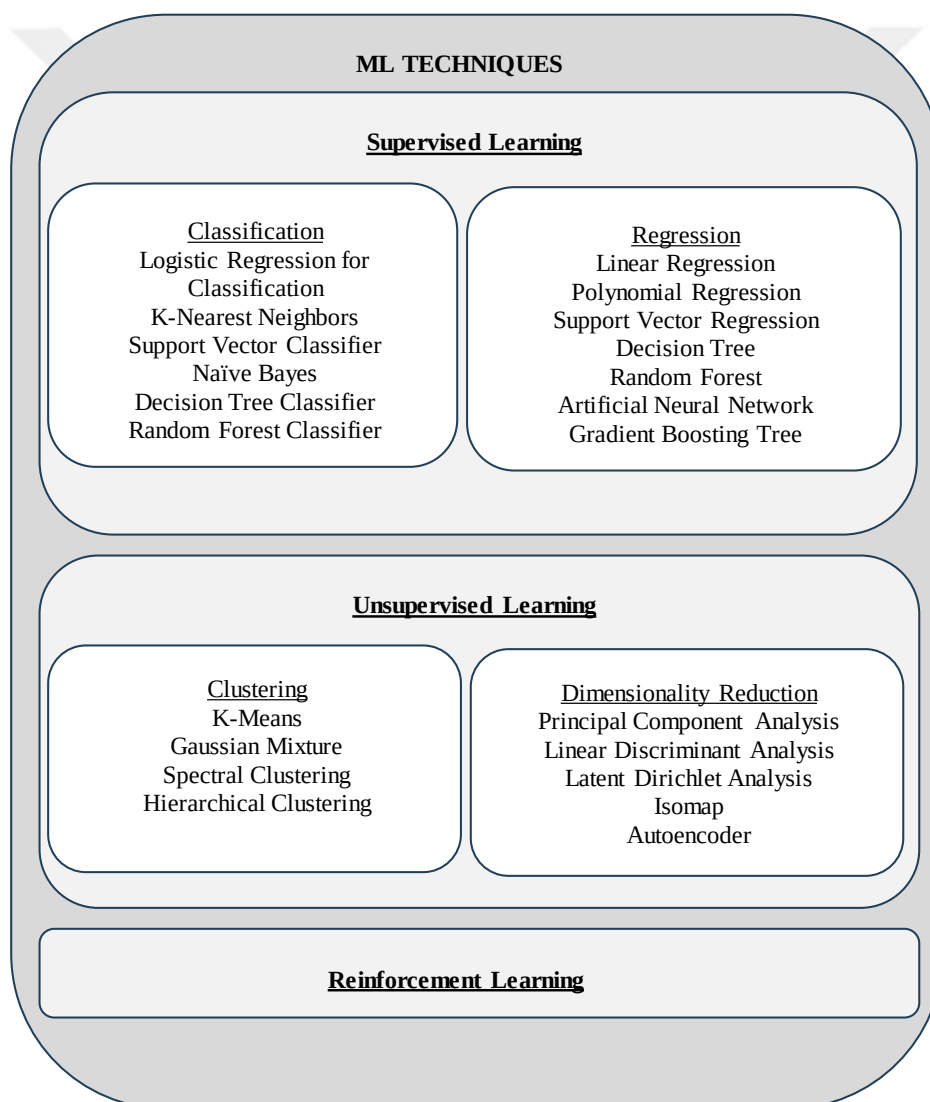


Figure 2.5 : Main categories of machine learning techniques.

Supervised and unsupervised categories are the most common categories used for machine learning problems.

2.5.1 Supervised learning

Supervised learning uses the existing information of a labeled dataset to learn a function that accurately approximates the connection between input and labeled output in the data (Xie et al., 2020). In the area of supervised learning, each observation is paired with a predictor observation and an associated response measurement. By means of observed data, a suitable model can be used to predict future observations or infer the relationship between the response and predictors (James et al., 2023).

Supervised problems can be primarily categorized into two groups: classification and regression. Classification is employed to forecast future categories based on observed data when there are categorical outcomes. Alternatively, when the data consists of quantities rather than categories, regression methods are employed to forecast future outcomes and infer correlations.

Classification is mostly employed to ascertain the classes or groups, such as gender, football teams, client segments, and medical diagnoses. Several classification techniques commonly employed include Logistic Regression for Classification, K-Nearest Neighbors, Support Vector Classifier, Naïve Bayes, Decision Tree Classifier, Random Forest Classifier, and Artificial Neural Network.

Regression is primarily employed for output variables that are both continuous and discrete. They are utilized for purposes such as weather prediction, assessment of examination results, market sales projections, and so forth. Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree, Random Forest, Artificial Neural Network, and Gradient Boosting Trees are examples of regression techniques.

2.5.2 Unsupervised learning

The unsupervised learning strategy is used when the dataset lacks labels and predicts outcomes by identifying patterns underlying data. In unsupervised learning, each observation is associated with a vector response instead of a specific reaction. This presents a more complex situation compared to supervised learning (James et al., 2023).

Cluster analysis and dimensionality reduction are techniques used in unsupervised learning. Clustering is mostly employed for consumer segmentation and the classification of data with unlabeled classes, such as market targets. The algorithms employed in clustering are K-means, Density-based Spatial Clustering of Applications with Noise, Gaussian Mixture, Spectral Clustering, and Hierarchical Clustering.

Dimensionality reduction is mostly employed in the context of big data to generate meaningful structures. The algorithms used for dimensionality reduction include Principal Component Analysis, Linear Discriminant Analysis, Latent Dirichlet Analysis, Isomap, and Autoencoder.

2.5.3 Reinforcement learning

The two essential components of reinforced learning are an interacting environment and a reward-punishment mechanism. The final objective is specified to the machine, which can then autonomously learn the objective through a system of rewards and punishments. This technique is employed in the domains of gaming, such as chess and go, robotics, and autonomous vehicles.

ML approaches offer significant benefits by facilitating data processing, generating comprehensible outcomes, and elucidating the importance of various features (Bilal and Oyedele, 2020; Choi et al., 2020; Das et al., 2024; Gondia et al., 2020; Zhou et al., 2018). Moreover, these techniques are widely used in construction management to forecast performance (Wong, 2004), productivity (Desai and Joshi, 2010), safety issues (Drozd, 2017; Koc et al., 2021; Koc et al., 2022; Mistikoglu et al., 2015), and cost estimation (Scott and Bommu, 2024). In this thesis study, the emphasis was not placed on causality, which statistical approaches might elucidate. Instead, the main focus was prediction, leading to the preference for ML algorithms. These algorithms have inherent advantages since they can surpass the limitations of explicit model structures and suppositions about distributions of data frequencies, continuously enhancing their performance via experience and maximizing their predictive capabilities (Gondia et al., 2020).

3. RESEARCH METHODOLOGY

The methodology of the study is composed of three main sections: (1) data gathering, (2) risk score-based cost prediction and classification models for groups, and (3) reporting the results. The methodology is illustrated in FigureA.2 in Appendix B. Flow Chart of the Research Methodology; details are explained in the following sections.

3.1 Data gathering

3.1.1 Gathering risk register

A risk register is a valuable instrument for recording and demonstrating all potential risks at the preliminary tendering phase and directing risk management throughout the construction process. Hillson (2003) asserts that a comprehensive understanding of the risks confronting a project and organization is essential for successful and effective risk management. Laryea and Hughes (2008) assert that contractors employ the risk register method to assess and ascertain the financial implications of risks during the tender phase. The process begins with idea generation, followed by data identification, analysis, and evaluation utilizing a spreadsheet. Risk registers typically comprise parts that outline definitions, probability of risks, consequences, and mitigation techniques. Diverse approaches may be utilized in generating risk register reports for construction projects based on the requirements of the project team and the client. A consistent framework is advantageous for ensuring uniformity in the compilation of these documents.

Moreover, the ISO 9001:2015 Quality Management System requires implementing a “Risk management” approach in all business processes (2015). Generally, in the early stages of the project, the executives commenced identifying and assessing potential risks. These risks are defined and evaluated as the project’s risk register (Patterson and Neailey, 2002). These registers delineate the potential risks and their associated likelihood and impact estimates as determined by the executives (Rani et al., 2024).

Moreover, they are the primary tool of risk management utilized to detect potential risks within a project or organization, facilitating methodical knowledge management.

The total risk score of the project is calculated depending on the probability and impact of the risk items. The evaluation of these factors might differ depending on the projects' assessment methods. Based on the project characteristics, some subcategories of risks may exist, and the executives could assign different weights to these categories. The total risk score computation is presented in equation (3.1).

$$Total\ Risk\ Score = \frac{1}{k} \sum_{e=1}^k \left(\sum_{i=1}^m (W_{i,e} \frac{1}{n_i} \sum_{j=1}^{n_i} (P_{i,j,e} \times I_{i,j,e})) \right) \quad (3.1)$$

where $P_{i,j,e}$ is the probability value assigned by the e^{th} executive for j^{th} risk item in i^{th} risk category, $I_{i,j,e}$ is the impact value assigned by the e^{th} executive for j^{th} risk item in i^{th} risk category, n is the total number of risk items, $W_{i,e}$ is the weight value assigned by the e^{th} executive for i^{th} risk category, m is the total number of categories, and k is the total number of executives.

The overall risk score is derived from the values of probability and impact. The values typically range from 1 to 5, yielding a total risk score between 1 and 25. equation (3.2) is suggested to be adjusted for various potential total risk score computations while evaluating many projects concurrently.

$$TRS_s = \frac{(TRS_o - min_o)}{(max_o - min_o)} \times (max_s - min_s) + min_s \quad (3.2)$$

where TRS_s is the scaled total risk score, TRS_o is the original total risk score, min_o and max_o are the minimum and maximum values of the original scale, and min_s and max_s are the minimum and maximum values of the new scale.

3.1.2 Gathering estimated completion cost

The cost estimation process in construction is crucial for maintaining better project management quality and technical performance (Adeli and Wu, 1998; Derakhshanalavijeh and Teixeira, 2017). The construction project's tender cost is modified post-bid acceptance and contract signing, contingent upon newly acquired information subsequent to contract initiation. The tender cost is revised following receipt of comprehensive project information, and the projected completion cost for the project studies phase is commenced. Due to the significant variability in

construction projects, it is advantageous to frequently reassess estimated costs. The project's completion cost is mostly determined by the aggregate of direct and indirect expenses. Direct costs encompass estimations of quantities and unit pricing, whereas indirect cost assessments mostly depend on the anticipated project duration and monthly expenditures for indirect products. This paper advocates for the systematic collecting of periodically forecasted completion cost data for construction projects. For clarity, the overall expense of the project should be prioritized over the cost breakdown structure. The primary focus is on the project's completion cost and the revised cost value during the execution phase.

3.1.3 Grouping projects and data

The risk registers and expected completion cost data of the construction projects are collected. Following the project's initiation, regular risk assessment and estimates of the completion cost of the project period begin. Data regarding these files throughout the same timeframe are gathered. To accomplish the goal of making this study easier to apply, it is suggested to gather both the total cost and total risk score of the project instead of breakdowns.

Depending on the specific project types, different approaches can be employed to provide risk register reports and completion cost assessments for construction projects. Hence, it is recommended to assess projects using a system based on their characteristics. The recommended project types, which includes residential housing projects, heavy civil projects and infrastructure, commercial and institutional projects and specialized industrial projects are listed in Table 3.1.

Table 3.1 : Project types and groups (Hendrickson and Au, 1989).

Code	Project Type (PT)
PT.01	Residential Housing Projects
PT.02	Heavy Civil Projects and Infrastructure
PT.03	Commercial and Institutional Projects
PT.04	Specialized Industrial Projects

3.2 Modeling

3.2.1 Machine learning regression models

The proposed approach to predict the completion cost in construction projects based on the total risk score employing ML consists of three primary stages. (1) data preprocessing, (2) model selecting and modeling, and (3) evaluating and validating the model.

In the first stage, standardized total cost is calculated as formulated in equation (3.3).

$$\text{Standardized Total Cost} = \frac{(X - \bar{y})}{s} \quad (3.3)$$

where X is the total completion cost value in a related project, $\bar{y}$ is mean of the sample and s is the standard deviation of the sample.

Subsequently, the input dataset is randomly split into training and testing datasets, with two-thirds allocated to the training dataset and one-third allocated to the testing dataset. The testing set was randomly selected and set aside to evaluate the model's performance on unseen data, while the training set was utilized to calibrate the model.

In the second stage, algorithms are selected for regression. When choosing a machine learning method for a problem, the sample size and the properties of the data serve as criteria for identifying the most appropriate model (Moreno, 2020). Nonetheless, no conclusive machine learning technique is optimal for resolving certain issues (Mahmoodzadeh et al., 2022). Consequently, it is essential to assess the efficacy of the employed ML algorithms.

In the third stage, the employed algorithms are evaluated and validated. The coefficient of determination (R^2) indicates a more robust correlation between the observed and predicted values. The accuracy of the R^2 score model is evaluated by comparing the model's output to the target value. R^2 value approaching 1 signifies a more robust relationship between observed and expected values. This study proposes establishing the R^2 threshold at 0.75 to assess the model's accuracy performance. The assessment of R^2 is performed in three groups: the training group, the complete dataset, and the validation group within the context of this study. The R^2 value is computed using the data points from the regression model determined by equation (3.4).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.4)$$

where y_i is observed i^{th} value, $\hat{y}_i$ is predicted value of y_i and $\bar{y}$ is mean value of a sample.

3.2.2 Machine learning classification models

The employed methodology of classification is composed of three main stages: (1) data preprocessing, (2) model selecting and modeling, and (3) evaluating and validating the model.

In the first stage, data is preprocessed. The collected risk records and completion cost data of the projects are consolidated. Although the currency and magnitude of the project cost might vary depending on the size and location of the project, using the cost overrun ratio presented in equation (3.5) rather than the total cost value in calculations, eliminates the differences.

$$TC_i\% = \frac{(TC_i - TC_{i-1})}{TC_{i-1}} \times 100 \quad (3.5)$$

where TC_i is the total cost of i^{th} prediction period of the project.

The suggested approach involves classification algorithms to categorize the data into several classes. These classifications are contingent upon the project's requirements and characteristics. It is recommended that guidance be sought from the project executives. These classes should be set out to facilitate the presentation for executives or project managers about project total cost escalation depending on total risk scores and offer enhanced managerial understanding and advantages.

Afterward, the input dataset is randomly divided into training and testing datasets, with two-thirds assigned to the training dataset and one-third assigned to the testing dataset. The training set is used to improve the model. On the other hand, the testing set is chosen randomly and kept separate to assess the model's performance on data that it has not been exposed to before.

In the second stage, algorithms are selected for classification and subsequent modeling. The primary objective of the suggested approach is to categorize the cost overrun ratio of project completion cost based on total risk scores, depending on the determined classes. It is recommended to utilize supervised learning. Supervised

learning entails linking each observation with a predictor observation and a corresponding response measurement. A suitable model should be able to anticipate future observations or infer the correlation between the response and predictor variables using observed data (James et al., 2023). The choice of algorithm is determined by both the sample size and the characteristics of the data. Furthermore, it is acknowledged that no definitive ML technique is generally optimal for resolving specific problems (Mahmoodzadeh et al., 2022).

In the third stage of the methodology, the employed algorithms are evaluated and validated by comparing the expected and actual class outcomes throughout training and testing. The models are assessed by calculating the confusion matrix and accuracy. Once the algorithms with an accuracy level exceeding 70% are chosen, performance indices are computed to overcome the accuracy paradox. The algorithms are ranked based on their performance metrics. Subsequently, resampling validation methods are employed to examine the testing performance of models, and the accuracy is computed for each subsampling. Performance indices are produced for these algorithms if the average accuracy exceeds 80%. The algorithm with the highest accuracy and performance indices is chosen from these options, and its parameters are optimized.

3.3 Reporting the results

This study utilizes regression and classification machine learning algorithms to estimate completion costs and cost overrun classifications for construction projects, respectively. The concluding section of the approach recommends presenting the performance outcomes of the top three algorithms for regression and classification models. Alongside providing a concise summary, it also presents the methods that yield optimal results based on the analyzed data for the users.

4. IMPLEMENTATION OF THE PROPOSED APPROACH: CASE STUDIES

4.1 Data gathering

Information about overall expenses and evaluations of potential risks were gathered from a general contractor that predominantly undertakes projects on a global scale and specializes mainly in constructing airports and high-rise buildings. A case study was conducted using this data to illustrate how the proposed approach can be applied in practice.

The risk register and completion cost of projects were collected from the analyzed construction company. Various risk registers were created to meet the specific requirements of the client and project. All risk registers were standardized by integrating them into an identical format. To mitigate data limitations, a comprehensive review of all registers was conducted. To minimize risk perception and standardize the assessment of risk scores uniformly inside an organization, the probability and impact scales presented in Tables 4.1 and Table 4.2 were employed. The executives of the data-gathered company recommended these scales for the provided data.

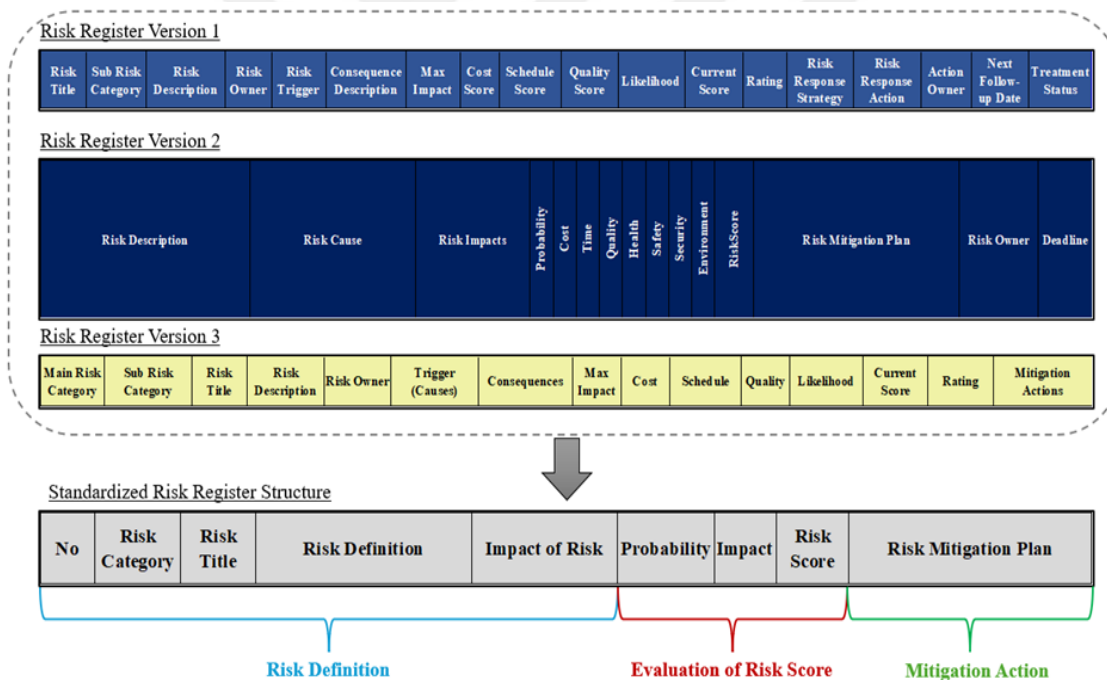
Table 4.1 : Risk probability scale.

Score	Probability Definition
5: Very High Probability	>70%
4: High Probability	51–70%
3: Probable	31–50%
2: Low Probability	10–30%
1: Very Low Probability	<10%

Table 4.2 : Risk impact scale.

Score	Expected Financial Impact	Execution Impact
5: Very High Impact	>3% contract value	Suspension of execution and/or termination
4: High Impact	1.8–3.0% contract value	Major interruption on execution
3: Medium Impact	1.2–1.8% contract value	Serious interruption on execution
2: Low Impact	0.6–1.2% contract value	Minor interruption on execution
1: Very Low Impact	<0.6% contract value	Insignificant interruption on execution

Various formats were employed in the generation of risk register reports and completion cost assessments for construction projects, contingent upon the project's features and the priorities of the project management. Figure 4.1 displays the various risk register formats obtained from projects along with their standardized structure.

**Figure 4.1 : Standardized structure of risk register.**

Then, the gathered risk registers were matched with the estimated project completion cost within the same timeframe. Finally, the projects' total risk scores and expected completion costs were classified based on their various project types, as presented in Figure 4.2. There are four projects classified as PT.01 and seven projects for PT.02.

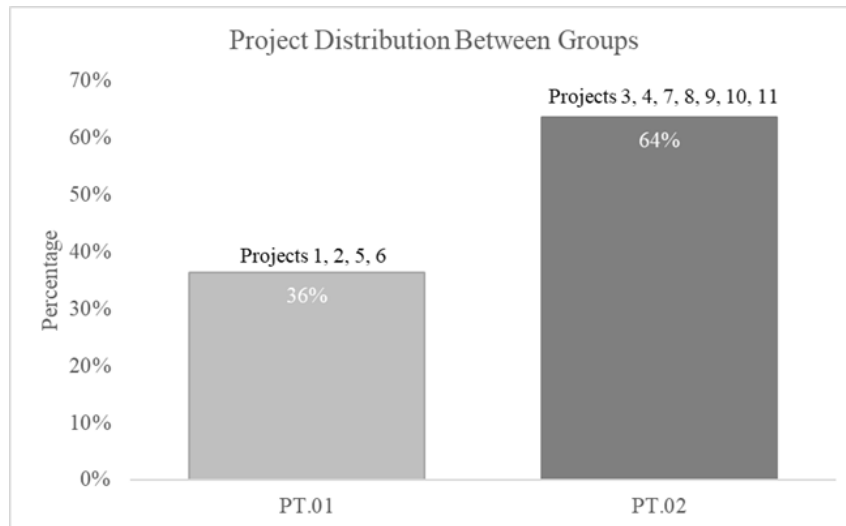


Figure 4.2 : Groups depending on the project types.

4.2 Modelling

4.2.1 Risk-based completion cost estimation using machine learning prediction algorithms

4.2.1.1 Data understanding and preprocessing

The PT.01 group consists of four projects, with a total of 53 data points. Linear prediction is implemented in Excel to comprehend the correlation between total cost and total risk score associations. There exists a significant correlation between the total cost and total risk within the PT.01 category depending on the R^2 value, which is 0.8589. The relationship is depicted in Figure 4.3, and the data from each project is illustrated in Figure 4.4.

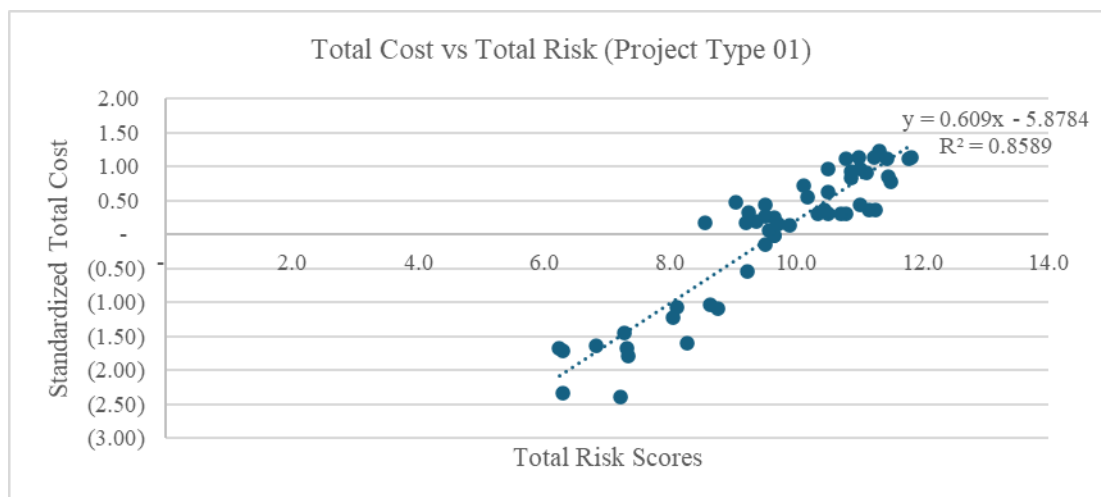


Figure 4.3 : Linear relationship between total cost and risk scores (PT.01).

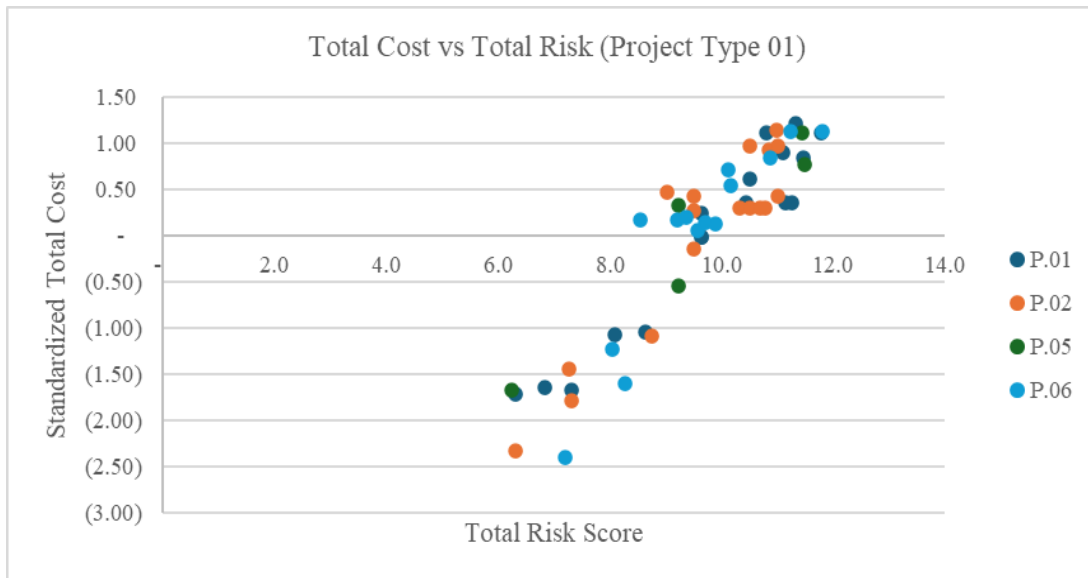


Figure 4.4 : Project based linear relationship between total cost and risk scores (PT.01).

The PT.02 group consists of seven projects, with a total of 66 data points. Linear prediction is implemented in Excel to comprehend the correlation between total cost and total risk score associations depending on R^2 value, which is 0.7526. The linear relationship is presented in Figure 4.5, and the data from each project is illustrated in Figure 4.6.



Figure 4.5 : Linear relationship between total cost and risk scores (PT.02).

The data is divided into sections that are proportional to one-third to two-thirds for training and testing purposes. A total of two-thirds of the dataset for related groups was allocated for training purposes, while the remaining one-third was placed for validation purposes.

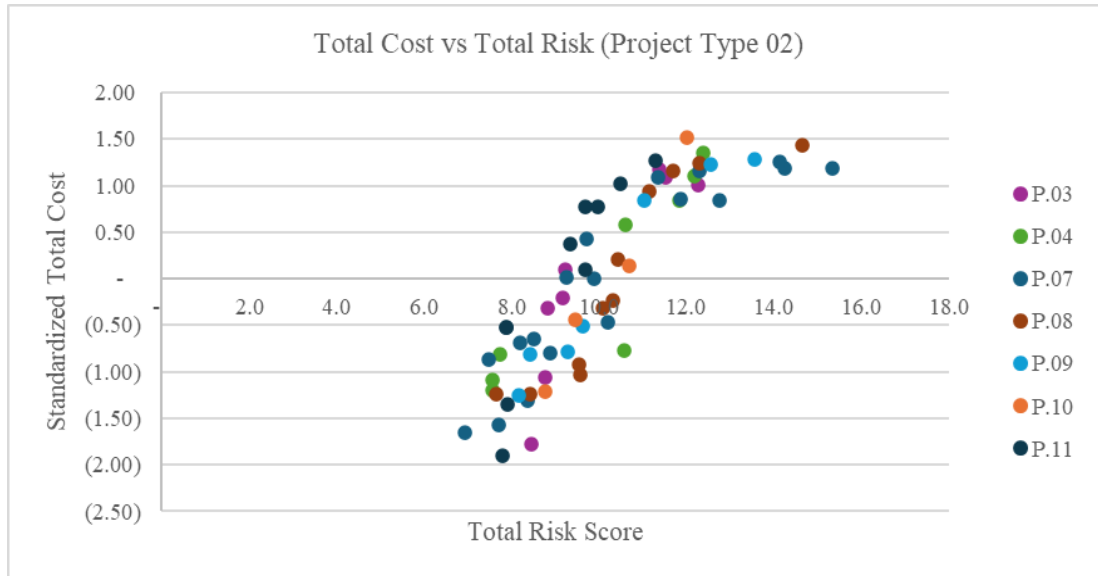


Figure 4.6 : Project based linear relationship between total cost and risk scores (PT.02).

4.2.1.2 Model selecting and modeling

The methodology section indicates that supervised learning regression techniques were employed to forecast project completion costs. The following algorithms were employed due to the dataset containing fewer than one thousand data points. These algorithms are extensively utilized and user-friendly. Furthermore, they can deliver adequate performance in small-sized data sets (Aggarwal, 2015; Gondia et al., 2020):

- **Linear Regression (LR):** In the early 19th century, Legendre and Gauss published articles that discussed the method of least squares, which is considered the earliest form of linear regression (James et al., 2013). Linear regression is a fundamental technique employed in supervised learning, known for its inherent constraint of only producing linear functions.
- **Polynomial Regression (PR2 and PR4):** Polynomial regression enhances the linear model's capacity to incorporate non-linear relationships using polynomial functions. This enhancement is accomplished by utilizing non-linear correlation.
- **Support Vector Regression (SVR):** Support Vector Regression (SVR) is a versatile technique that may be used for both classification and regression applications. The classification characteristic is converted into a regression attribute by the minimization of a certain sort of loss function. The loss function in question only takes into account residuals that exceed a positive constant in

absolute value, as mentioned by James et al. (2013). The radial basis function is employed as the default kernel function.

- **Decision Tree (DT):** The Decision Tree methodology is frequently utilized in the field of data analysis for the objectives of classifying and predicting outcomes. The methodology employed in this study entails the identification of distinct essential regions inside the predictor space through categorization or segmentation.
- **Random Forest (RF):** The Random Forest algorithm has a higher degree of flexibility in comparison to the Linear Regression model. Random forests provide an improvement over bagged trees by incorporating a slight modification that promotes decorrelation among the individual trees. In a manner similar to the bagging technique, the construction of numerous decision trees is facilitated by utilizing bootstrapped training samples, as outlined by James et al. (2013).

The utilization of machine learning methods is prevalent across various industries; however, it has yet to be extensively applied in construction estimating, as per the findings of Park et al. (2022). The Python programming language is widely employed to manage classification and prediction tasks. Regression approaches within the field of machine learning techniques are employed to determine the optimal inference model. In this study, various regression models including linear regression (LR), second-degree polynomial regression (PR2), fourth-degree polynomial regression (PR4), support vector regression (SVR), decision tree (DT), and random forest for regression (RF) were implemented using the Python programming language. The codes are provided in Appendix C. Python codes machine learning regression algorithms.

4.2.1.3 Evaluating and validating the prediction models

By employing the selected models, two-thirds of the dataset is utilized for training purposes. Subsequently, predictions are obtained for one-third of the dataset. Figure 4.7 and Figure 4.8 illustrate the standardized cost prediction outcomes derived from the testing set of groups alongside the matching actual values of these points for PT.01 and PT.02 groups, respectively.

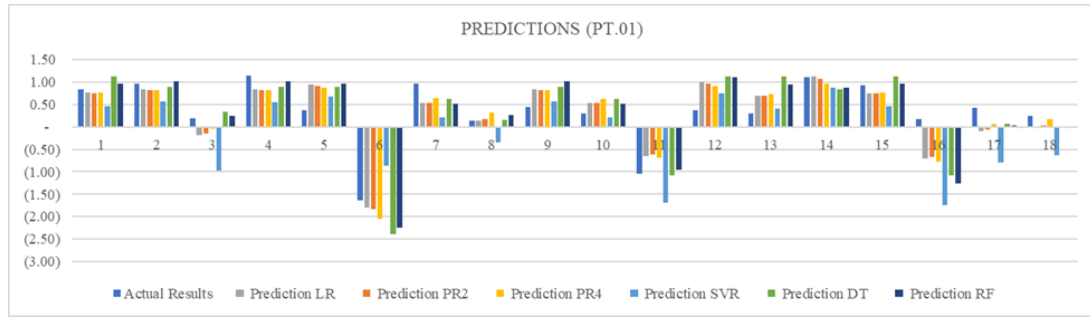


Figure 4.7 : Prediction results of regression models (PT.01).

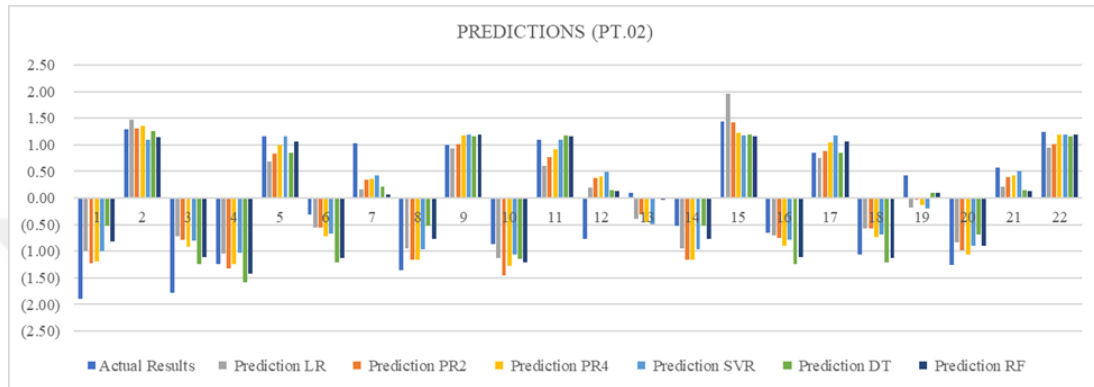


Figure 4.8 : Prediction results of regression models (PT.02).

Upon analyzing the third point in the PT.01 group, it was determined that RF and DT produced the most precise outcomes, but SVR resulted in the least precise outcome. Conversely, in the PT.02 group, SVR and RF yielded the most precise predictions for the fifth point, whilst LR exhibited the least performance.

R^2 for training group is calculated by using the training data points in accordance with the regression model derived from training data group. Furthermore, R^2 is employed to assess the degree of accuracy of fit for all data points and the regression model derived from the entire dataset. The values are displayed in Table 4.3 and Table 4.4.

Table 4.3 : R^2 for training group and all data group (PT.01).

Methods	R^2 All Data	Rank for All Data	R^2 Training Group	Rank for Training Group
LR	0.8589	5	0.8872	6
PR2	0.8648	4	0.8888	5
PR4	0.8898	3	0.9200	3
SVR	0.8404	6	0.9148	4
DT	0.9077	1	0.9933	1
RF	0.8945	2	0.9764	2

In PT.01 group DT, RF and PR4 are the best R^2 values for training data and all data groups.

Table 4.4 : R^2 for training group and all data group (PT.02).

Methods	R^2 All Data	Rank for All Data	R^2 Training Group	Rank for Training Group
LR	0.7526	6	0.7433	6
PR2	0.8000	5	0.7961	5
PR4	0.8198	3	0.8234	4
SVR	0.8124	4	0.8235	3
DT	0.9005	1	0.9998	1
RF	0.8951	2	0.9622	2

In PT.02 group DT and RF algorithms have the highest R^2 values when applied to both the training data and all data groups. SVR algorithm ranks third in terms of its R^2 performance for the training group, while the PR4 algorithm ranks fourth for all data.

All models fulfill the R^2 criteria for the PT.01 group, however the LR in PT.02 failed to exceed the threshold of 0.75. The DT model surpasses the other models for the entire dataset and the training group in PT.01 and PT.02.

After calculating R^2 for all data and training group, it should be calculated for test group. The concept of variance is to the quantification of dispersion or variability within a given amount of data points. When the variance is high, the model function is greatly influenced by changes in the training set. On the other hand, bias refers to the discrepancy between a simple estimate and the actual values inside a given dataset (James et al., 2023). Additionally, it is common for the training error rate to exhibit notable disparities with the test error rate, the test error refers to the mean error that arises when employing a statistical learning technique to forecast the outcome of a new observations (James et al., 2023). Therefore, R^2 for test data is calculated and shown in Table 4.5 and Table 4.6 for PT.01 and PT.02 groups, respectively.

Within the analyzed group, PR4 and SVR provide superior performance in PT.01, whereas LR achieves enhanced outcomes in PT.02 for R^2 in the test dataset.

The process of resampling methods entails iteratively extracting samples from a training set and reconfiguring a model of interest on each sample to acquire supplementary insights about the fitted model (James et al., 2023). Two common resampling methods are bootstrap and cross-validation. The bootstrap method is frequently used in various situations, mostly to evaluate the precision of parameter estimation or the effectiveness of a statistical learning system.

Table 4.5 : R^2 for test group (PT.01).

Actual Values	Average Diff.	Prediction LR	LR (Residual2)	Prediction PR2	PR2 (Residual2)	Prediction PR4	PR4 (Residual2)	Prediction SVR	SVR (Residual2)	Prediction DT	DT (Residual2)	Prediction RF	RF (Residual2)
0.93	1.05	0.72	0.05	0.71	0.05	0.65	0.08	0.85	0.01	0.84	0.01	0.87	0.00
-1.63	2.38	-1.64	0.00	-1.70	0.00	-2.07	0.19	-1.82	0.03	-1.67	0.00	-1.73	0.01
0.77	0.74	1.09	0.11	1.01	0.06	0.96	0.04	1.01	0.06	0.84	0.01	0.91	0.02
0.97	1.13	0.80	0.03	0.78	0.04	0.70	0.07	0.93	0.00	0.44	0.29	0.76	0.04
0.97	1.13	0.51	0.21	0.53	0.19	0.55	0.18	0.64	0.11	0.62	0.12	0.56	0.17
1.13	1.49	0.94	0.04	0.89	0.06	0.80	0.11	1.00	0.01	0.36	0.58	0.61	0.27
-2.40	5.31	-1.42	0.95	-1.44	0.92	-1.90	0.24	-1.77	0.40	-1.67	0.53	-1.59	0.65
0.17	0.07	-0.25	0.18	-0.18	0.12	-0.04	0.05	0.02	0.02	0.33	0.03	0.40	0.05
0.06	0.02	-0.04	0.01	0.03	0.00	0.19	0.01	0.18	0.01	0.15	0.01	0.12	0.00
-1.44	1.82	-1.38	0.00	-1.39	0.00	-1.85	0.17	-1.74	0.09	-1.67	0.05	-1.59	0.02
0.36	0.21	0.95	0.34	0.90	0.29	0.81	0.20	1.01	0.41	1.22	0.73	1.04	0.45
-1.07	0.95	-0.90	0.03	-0.85	0.05	-1.07	0.00	-1.05	0.00	-1.23	0.03	-1.06	0.00
-1.09	0.99	-0.52	0.32	-0.45	0.41	-0.42	0.44	-0.33	0.57	-1.04	0.00	-0.58	0.25
-0.54	0.20	-0.24	0.09	-0.16	0.14	-0.03	0.26	0.03	0.32	0.33	0.75	0.40	0.88
0.20	0.08	-0.15	0.12	-0.08	0.08	0.07	0.02	0.10	0.01	0.33	0.02	0.23	0.00
0.13	0.05	0.15	0.00	0.21	0.01	0.35	0.05	0.31	0.03	0.15	0.00	0.24	0.01
0.27	0.13	-0.07	0.12	0.00	0.07	0.16	0.01	0.16	0.01	0.15	0.02	0.12	0.02
0.54	0.41	0.32	0.05	0.36	0.03	0.45	0.01	0.44	0.01	0.71	0.03	0.55	0.00
Ave.:	-0.09												
Sum:	18.16		2.64		2.51		2.12		2.12		3.19		2.86
R^2 :			0.85		0.86		0.88*		0.88*		0.82		0.84

* Superior R^2 results in testing groups within applied ML algorithms.

Cross-validation is employed to assess the performance of a statistical learning method by utilizing the corresponding test error. Training error and test error rates are different from each other. Training error can be calculated on the bootstrap sample. The test error refers to the mean error that arises when employing a statistical learning technique to forecast the outcome of an entirely new observation. Two types of cross-validation techniques are Leave-One-Out Cross-Validation (LOOCV) and k-fold cross-validation. In LOOCV validation set is single observation. On the other hand, randomly divided groups of observations are available in k-fold cross-validation. K-fold cross-validation partitions observations into k groups (folds) in a random manner, with the first fold used for validation and the remaining $k - 1$ folds used for fitting.

Table 4.6 : R^2 for test group (PT.02).

Actual Values	Average Diff.	Prediction LR	LR (Residual2)	Prediction PR2	PR2 (Residual2)	Prediction PR4	PR4 (Residual2)	Prediction SVR	SVR (Residual2)	Prediction DT	DT (Residual2)	Prediction RF	RF (Residual2)
-1.35	1.75	-1.06	0.09	-1.29	0.00	-1.33	0.00	-1.30	0.00	-1.90	0.30	-1.57	0.05
-1.04	1.02	-0.33	0.51	-0.15	0.79	-0.26	0.61	0.10	1.30	-0.92	0.01	-0.75	0.08
-0.53	0.25	-1.07	0.29	-1.31	0.61	-1.34	0.66	-1.31	0.60	-1.90	1.88	-1.57	1.09
-0.65	0.38	-0.80	0.02	-0.85	0.04	-1.02	0.14	-1.05	0.16	-1.78	1.28	-1.63	0.96
-1.09	1.12	-1.20	0.01	-1.55	0.21	-1.43	0.11	-1.26	0.03	-1.20	0.01	-1.20	0.01
-1.23	1.45	-1.17	0.00	-1.48	0.06	-1.41	0.03	-1.29	0.00	-1.58	0.12	-1.35	0.01
1.52	2.40	0.76	0.58	0.96	0.31	1.11	0.16	1.19	0.11	0.86	0.44	0.91	0.37
0.85	0.77	0.33	0.27	0.60	0.06	0.73	0.01	1.12	0.07	0.94	0.01	1.09	0.06
0.77	0.65	-0.14	0.83	0.09	0.47	0.06	0.51	0.57	0.04	0.00	0.60	-0.22	0.98
0.58	0.37	0.13	0.20	0.40	0.03	0.48	0.01	1.01	0.19	0.14	0.19	0.35	0.05
-0.53	0.25	-1.06	0.29	-1.30	0.59	-1.33	0.65	-1.31	0.60	-1.90	1.88	-1.57	1.09
-0.77	0.55	0.12	0.80	0.39	1.34	0.46	1.51	1.00	3.13	1.02	3.23	0.53	1.70
1.10	1.28	0.83	0.08	1.01	0.01	1.15	0.00	1.21	0.01	1.00	0.01	1.04	0.00
0.10	0.02	-0.48	0.33	-0.36	0.21	-0.51	0.37	-0.31	0.17	0.01	0.01	-0.07	0.03
1.23	1.59	1.00	0.05	1.11	0.01	1.22	0.00	1.25	0.00	1.35	0.02	1.15	0.01
-0.23	0.04	0.01	0.06	0.26	0.24	0.29	0.27	0.85	1.18	0.21	0.20	0.02	0.06
0.77	0.65	-0.28	1.10	-0.09	0.74	-0.17	0.89	0.24	0.29	0.10	0.46	-0.12	0.80
-0.68	0.43	-0.93	0.06	-1.07	0.15	-1.20	0.26	-1.23	0.30	-1.26	0.33	-1.21	0.28
1.17	1.44	0.48	0.48	0.74	0.19	0.89	0.08	1.14	0.00	1.09	0.01	1.11	0.00
-0.78	0.56	-0.45	0.11	-0.32	0.21	-0.47	0.10	-0.24	0.29	0.01	0.63	-0.04	0.55
0.43	0.21	-0.26	0.47	-0.07	0.24	-0.14	0.33	0.28	0.02	0.10	0.11	-0.12	0.30
-0.32	0.08	-0.09	0.05	0.14	0.21	0.13	0.20	0.67	0.97	-0.47	0.02	-0.19	0.02
Ave.:	-0.03												
Sum:	17.25		6.68		6.74		6.92		9.49		11.74		8.51
R^2 :			0.61*		0.6		0.6		0.45		0.32		0.51

* Superior R^2 results in testing groups within applied ML algorithms.

All procedure repeated k times. LOOCV is a specific sort of k-fold cross-validation where the value of k is equal to the number of observations. K-fold cross-validation is a widely used method for validating models by examining their testing performance. There is no precise rule about the fold size (Kuhn and Johnson, 2013). In this research, random states were employed with holdout methods k times. Employing the random state function in Python, ten states were randomly chosen for each group, and all computations were executed for each state. Table 4.7 and Table 4.8 presents the R^2 values for testing groups across various random states in PT 01 and PT.02 groups, respectively.

Table 4.7 : R² for test groups in different random states (PT.01).

Random State vs R ²	LR	PR2	PR4	SVR	DT	RF
1	0.855	0.862	0.883*	0.883*	0.824	0.843
2	0.833	0.771	0.632	0.898*	0.751	0.770
3	0.874	0.875	0.914	0.919*	0.720	0.853
4	0.776	0.769	0.791*	0.651	0.689	0.705
5	0.866*	0.865*	0.849	0.397	0.728	0.769
6	0.874	0.882	0.899*	0.858	0.783	0.878
7	0.851	0.857	0.891*	0.870	0.771	0.846
8	0.840	0.844	0.883*	0.693	0.715	0.767
9	0.860	0.849	0.872*	0.849	0.807	0.839
10	0.790	0.800	0.842*	0.644	0.768	0.785
Average R ²	0.842	0.837	0.846*	0.766	0.756	0.805
Rank	2	3	1	5	6	4

* Superior R² results for each random state.

Table 4.8 : R² for test groups in different random states (PT.02).

Random State vs R ²	LR	PR2	PR4	SVR	DT	RF
1	0.613*	0.609*	0.599	0.450	0.319	0.507
2	0.707	0.756	0.779*	0.763	0.566	0.656
3	0.785	0.787	0.802	0.812*	0.689	0.792
4	0.709	0.795	0.811*	0.439	0.684	0.725
5	0.674	0.758	0.742	0.766*	0.577	0.671
6	0.725	0.735*	0.738*	0.725	0.693	0.698
7	0.683	0.763*	0.756*	0.497	0.635	0.655
8	0.672	0.818	0.839	0.847*	0.794	0.850
9	0.794	0.776	0.819*	0.767	0.528	0.631
10	0.708	0.781	0.782	0.799*	0.653	0.673
Average R ²	0.707	0.758	0.767*	0.687	0.614	0.686
Rank	3	2	1	4	6	5

* Superior R² results for each random state.

The results indicate that PR4, LR, PR2, RF, and SVR in the PT.01 group meet the R² criterion (>0.75), whereas PR4 and PR2 in the PT.02 group also fulfill the R² standard. PR4 exhibits superior predictive accuracy for both groups in estimating standardized cost outcomes based on total risk score, as assessed by the R² criterion.

The mean square error (MSE) is a commonly employed metric for assessing the quality of regression model fits by measuring the divergence between actual values and the values predicted by the model, as illustrated in equation (4.1).

$$MSE = \frac{1}{p} \sum_{i=1}^p (y_i - \hat{y}_i)^2 \quad (4.1)$$

where p is number of data points, y_i is observed value and $\hat{y}_i$ is predicted value.

The cross-validation (CV) is derived by averaging the MSE across 10 states, as specified in equation (4.2).

$$CV_{(q)} = \frac{1}{q} \sum_{i=1}^q MSE_i \quad (4.2)$$

where q is the number of folds.

The cross-validation results for 10 random states in PT.01 and PT.02 were presented in Table 4.9 and Table 4.10, respectively.

Table 4.9 : CV(q) for 10 states in PT.01.

Random State vs MSE	LR	PR2	PR4	SVR	DT	RF
1	0.147	0.140	0.118*	0.118*	0.177	0.159
2	0.182	0.250	0.402	0.111*	0.272	0.251
3	0.136	0.136	0.093	0.088*	0.303	0.159
4	0.234	0.241	0.219*	0.365	0.325	0.309
5	0.175*	0.176	0.197	0.787	0.355	0.301
6	0.105	0.099	0.084*	0.118	0.181	0.102
7	0.203	0.195	0.150*	0.177	0.312	0.211
8	0.144	0.140	0.105*	0.276	0.256	0.210
9	0.155	0.167	0.142*	0.167	0.214	0.179
10	0.223	0.212	0.168*	0.379	0.247	0.229
CV _(q)	0.171	0.176	0.168*	0.259	0.264	0.211
Rank	2	3	1	5	6	4

* Superior results for each random state.

Table 4.10 : CV(q) for 10 states in PT.02.

Random State vs MSE	LR	PR2	PR4	SVR	DT	RF
1	0.304*	0.306	0.315	0.431	0.534	0.387
2	0.292	0.243	0.220*	0.236	0.432	0.342
3	0.208	0.207	0.192	0.182*	0.302	0.202
4	0.280	0.197	0.182*	0.539	0.304	0.265
5	0.252	0.187	0.199	0.181*	0.326	0.254
6	0.269	0.259	0.257*	0.269	0.301	0.296
7	0.333	0.249*	0.257	0.528	0.384	0.363
8	0.293	0.162	0.144	0.136	0.183	0.134*
9	0.189	0.205	0.166*	0.214	0.433	0.338
10	0.319	0.240	0.239	0.220*	0.379	0.358
CV _(q)	0.274	0.226	0.217*	0.294	0.358	0.294
Rank	3	2	1	4	6	5

* Superior results for each random state.

The findings demonstrate that PR4, LR, PR2, and RF in the PT.01 group fulfill the cross-validation criterion (<0.25), whereas only PR4 and PR2 in the PT.02 group comply. PR4 yields enhanced outcomes for both groups in forecasting standardized cost results based on total risk score concerning cross-validation criteria.

The results suggest that the PR4 algorithm is an exceptionally effective regression method for the PT.01 and PT.02 groups, as indicated by the R^2 , MSE, and cross-validation metrics. At this point, it is crucial to ascertain the ideal degree of the polynomial regression model that precisely represents the relationship between total risk and cost. It is expected that an increase in the polynomial degree will result in a decrease in training error. The function maintains sensitivity to all data points, resulting in significant fluctuation. The result is a notable rise in the mean square error detected in the test data. Conversely, a low degree in a polynomial function yields a basic inference formula, leading to bias.

Polynomial features were utilized in Python programming to systematically compute R^2 and MSE values from the 1st to the 10th degree across various states. The objective of this approach was to identify the ideal degree of polynomial flexibility that yields the highest R^2 value and the lowest MSE. The comprehensive results of this technique are illustrated in Figure 4.9 and Figure 4.10 for PT.01 and PT.02 groups, respectively.

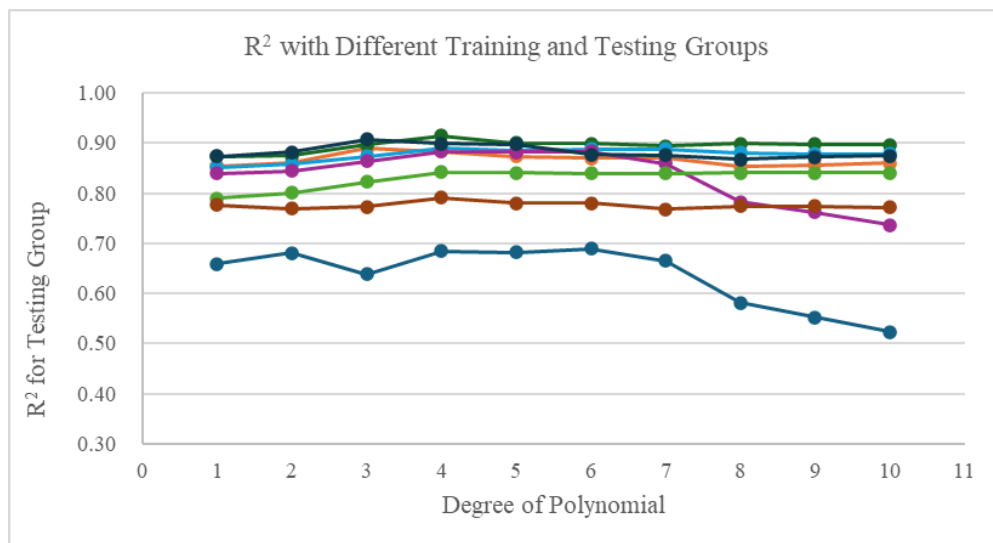


Figure 4.9 : Variability of R^2 different training and testing groups (PT.01).

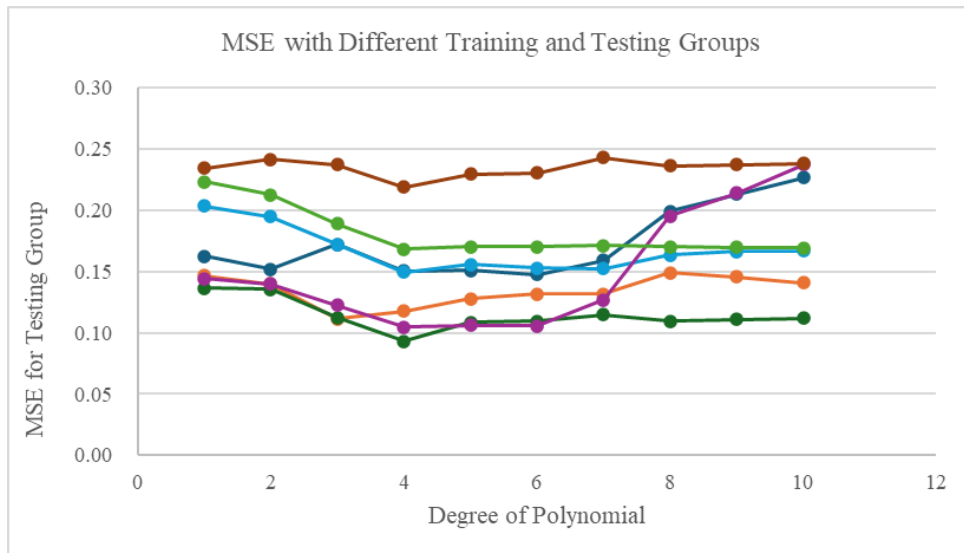


Figure 4.10 : Variability of MSE with different training and testing groups (PT.01).

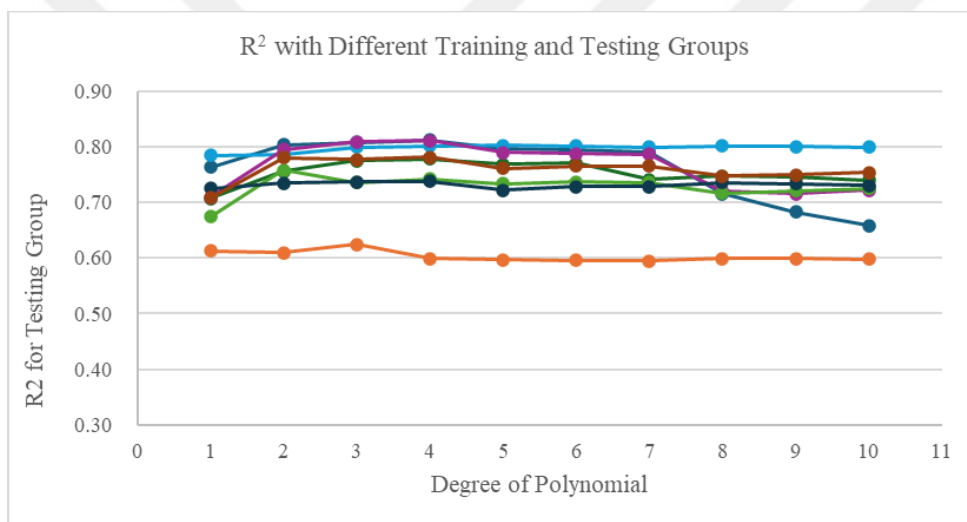


Figure 4.11 : Variability of R² different training and testing groups (PT.02).

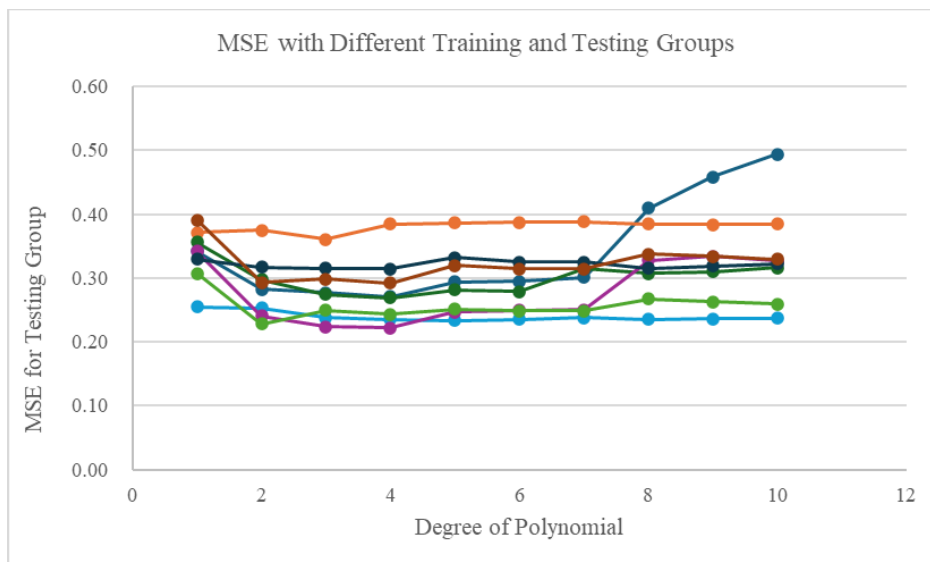


Figure 4.12 : Variability of MSE with different training and testing groups (PT.02).

According to the PT.01 group's findings, PR4 was identified as the best effective prediction method for forecasting the standardized total cost value, based on the average total risk score. In the PT.02 group, third-degree and fourth-degree polynomial regressions demonstrated the minimal MSE and maximal R^2 values. The third-degree polynomial regression (PR3) yielded marginally superior results compared to the fourth-degree polynomial regression (PR4). This section seeks to validate the precision of the models in forecasting results and to evaluate and compare their efficacy. The preceding study of both groups in the training and test datasets demonstrates that polynomial regression produces superior results. Consequently, PR4 was designated for PT.01, whereas PR3 was assigned to the PT.02 group based on their adeptness in forecasting standardized total cost utilizing the total risk score.

4.2.2 Risk-based completion cost overrun ratio estimation using machine learning classification algorithms

4.2.2.1 Data understanding and preprocessing

The PT.01 group consists of four residential projects, which have a total of 50 data points. On the other hand, the PT.02 group comprises seven heavy civil projects, primarily focused on airport-related projects, and has a total of 60 data points. This section utilizes the cost overrun ratio, which is computed based on prior cost data. Consequently, three data points from the PT.01 group and six data points from the PT.02 group were extracted, as they represent the initial data of the projects, and the cost overrun ratio for these points cannot be computed. Considering the cost data varies based on the projects, these data points were scaled. Subsequently, the data points were classified into four class labels, as recommended in Table 4.11. In the PT.01 group, 51% of the data points corresponded to Class 0, 35% to Class 1, 10% to Class 2, and 4% to Class 3. In addition, the PT.02 group had 47% of the data points in Class 0, 36% in Class 1, 10% in Class 2, and 7% in Class 3. While the distribution of data points between classes may not be uniform, these classes offer valuable perspectives on management. For a few projects, the estimated completion cost is expected to exceed the previous estimation by 10%. Consequently, there is an anticipation that the frequency of data will decline as the total cost overrun ratio rises. The development of classification algorithms and the acquisition of findings were achieved using Python 3.11 programming, a robust computational tool capable of effectively implementing

many ML approaches. Following the interpretation, data preparation, and library loading in Python, the data frame was converted into an array. Afterward, the dataset was divided into training and test sets. Training purposes were assigned two-thirds of the dataset, while one-third was set aside for validation.

As proposed in the methodology, the cost overrun ratios of each estimation in projects were calculated. Depending on the guidance of executives of the company from which data were gathered, four classes were used for cost overrun ratios, as illustrated in Table 4.11. Class 0 represents the non-changed completion cost in the estimation. Therefore, this circumstance might be considered negligible, requiring no action. Class 1 comprises minor cost increases depending on total risk scores, while Class 2 indicates significant increases, and Class 3 signifies severe increases in the completion cost depending on total risk scores and demands mitigation actions depending on reported risks.

Table 4.11 : Classes for cost overrun ratios.

Total Cost (TC) Increase %	Class
$TC\% \leq 0\%$	0
$0\% < TC\% < 5\%$	1
$5\% \leq TC\% < 10\%$	2
$10\% \leq TC\%$	3

4.2.2.2 Model selecting and modeling

The main objective of this study is to forecast the cost overrun ratio class of the project's total cost depending on total risk scores. ML categorization techniques were utilized due to the nature of the problem, and supervised learning was implemented depending on the features of the data. This technique entails utilizing existing data to acquire knowledge on how to allocate new data records to certain output classes, and it is highly efficient in predicting analytics of data (Gondia et al., 2020). Additionally, classification strategies can handle intricate factors and their efficacy in managing categorical variables (Aggarwal, 2015).

Several studies in construction management have employed ML categorization algorithms to forecast performance, productivity, and safety issues. Wong (2004) employed logistic regression for classification by examining 68 construction projects in England to forecast contractor performance using tender evaluation criteria of clients. Two categories indicated contractors' performance, and the logistic regression

classifier model accurately predicted contractor performance in 75% of instances. Desai and Joshi (2010) employed the decision tree algorithm to examine labor productivity in construction projects. The findings revealed that the decision tree algorithm yielded greater accuracy than conventional labor productivity approaches. Moreover, Mistikoglu et al. (2015) employed the decision tree algorithm to examine fall incidents. They applied a categorization approach with distinct classes, each representing a cause of fall accidents, and they concluded that the decision tree classifier achieved a high level of accuracy in classifying incidents of roofers falling. Further research has been conducted on safety concerns utilizing ML categorization approaches. Drozd (2017) utilized logistic regression to forecast the binary classification of accidents, determining whether they occurred. This was performed by considering a range of independent factors associated with building site safety measures and conditions. A total of 339 observations were analyzed, and the logistic regression model identified significant predictors of accidents. Choi et al. (2020) applied logistic regression, decision tree, random forest, and AdaBoost algorithms to design a prediction model for fatal accidents in the construction industry. The study aimed to forecast the likelihood of deadly accidents by categorizing them into “general accidents” and “fatal accidents”. Based on the findings, the random forest model had the highest level of accuracy in predicting outcomes for classes. Zhang et al. (2020) employed the random forest classifier to discover significant factors contributing to work delays and affecting decision-making. Moreover, Scott and Bommu (2024) utilized decision tree and random forest algorithms to ascertain the presence of cost overruns in construction projects. These ML models exhibit a high degree of accuracy in predicting cost overruns. It is known that there is no ML algorithm that has been proven to be the most effective for addressing specific problems (2022). Hence, it is essential to assess the efficacy of the algorithms using various approaches. The following ML classification algorithms were utilized due to their frequent usage, user-friendly nature, and ability to evaluate performance:

- **Logistic Regression for Classification (LRC):** The LRC algorithm is a classification approach that utilizes logistic functions to categorize data points based on their numerical values. The sigmoid function is employed to provide a clear distinction between different groups.

- **K-Nearest Neighbors (KNN):** The KNN methodology is a simple and resilient computing tool. This algorithm is a highly efficient technique for dividing non-linear data into distinct segments. This classifier employs multiple metric methods. The research utilized the default “Minkowski” approach with a 2nd degree, which gives Euclidian distance. For the closest neighbor number in the function, there are different results due to the operational logic of the function. Within the scope of this study, the value is set to 1 in order to simplify the classification process by choosing the closest neighboring data point. This option facilitates a more straightforward decision-making process to enhance precision while also potentially decreasing computing complexity and improving the speed of categorization (Saadatfar et al., 2020).
- **Support Vector Classifier (SVC):** The objective of this function is to identify the line that provides the maximum margin between groups. SVC enables the differentiation between binary and multi-class classification on a dataset using a Kernel function, allowing for both linear and non-linear separation (James et al., 2013).
- **Naïve Bayes Classifier (NBC):** The Naïve Bayes classifier operates by utilizing the principle of conditional probability. NBC minimizes test error by assigning each observation to the class that has the highest probability, determined by its predictor values (James et al., 2013).
- **Decision Tree Classifier (DTC):** DTC works based on the majority decision principle. This algorithm focuses on how independent factors divide data records into class labels by analyzing the training data (Gondia et al., 2020). Subsequently, the algorithm generates a classifier represented as a hierarchical structure, like a tree, composed of nodes, branches, and leaves. The root node symbolizes the complete dataset. Afterward, the data are divided according to a specific characteristic, resulting in the creation of subnodes. These nodes are linked by branches, which correspond to the associated variables. Leaves, on the other hand, reflect the final determined class labels. DTC has the benefit of comprehending and explaining complicated problems in an understandable hierarchical structure (Choi et al., 2020).

- **Random Forest Classifier (RFC):** The RFC method is a widely used classification algorithm that utilizes majority voting and integrates many decision tree principles. The act of creating an ensemble of trees and allowing them to vote for the most prevalent class has led to notable enhancements in classification accuracy (Breiman, 2001).

The codes are provided in Appendix D. Python codes machine learning classification algorithms.

4.2.2.3 Evaluating and validating the classification models

The algorithms were trained using two-thirds of the data, employing LRC, KNN, SVC, NBC, DTC, and RFC. After obtaining the models based on the training group, predictions were generated for the remaining one-third of the data. The total number of data sets is 110, including 50 from the PT.01 group and 60 from the PT.02 group. Sixteen data points from the PT.01 group and twenty data points from the PT.02 group were employed for validation. This section aims to verify the accuracy of the models' predictive capabilities and evaluate their efficiency. The Confusion Matrix is a widely used tool in the ML field for evaluating classification algorithms' performance. The rows of the confusion matrix show the number of actual classifications, and the columns reflect the number of predicted classifications. The diagonal of the table shows the correct classified data, namely accurate predictions. The confusion matrices in heatmap format for the six classifiers are shown in Figures 4.13 and Figure 4.14 for the PT.01 and PT.02 groups, respectively. The intensity of the color corresponds to the concentration of data within the cell, and these tables match the suggested classes described in Table 4.11.

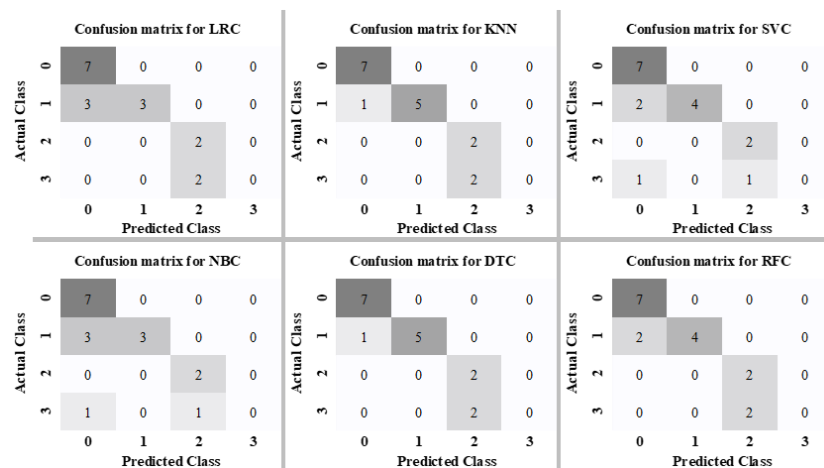


Figure 4.13 : Heatmap of confusion matrices for the PT.01 group.

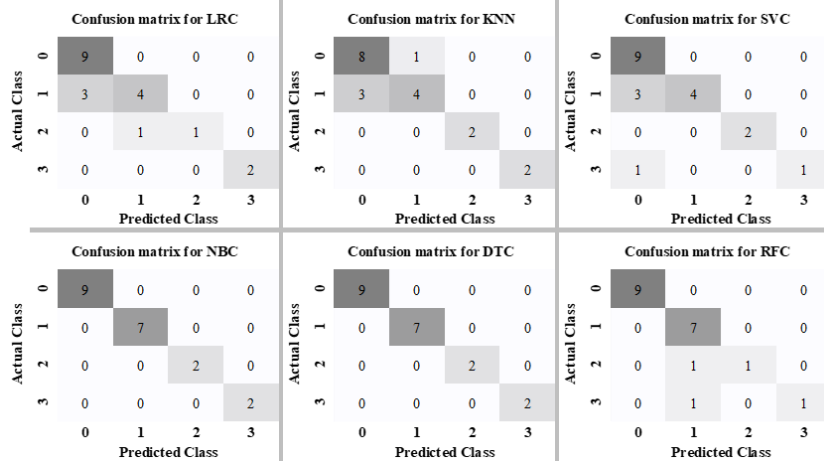


Figure 4.14 : Heatmap of confusion matrices for the PT.02 group.

Subsequently, the classification accuracies (ACC) of each algorithm within the groups were computed using equation (4.3) where CC is correct classifications and TS is the total samples, and the outcomes are displayed in Table 4.12.

$$ACC = \frac{CC}{TS} \quad (4.3)$$

Table 4.12 : Accuracy percentages of PT groups according to algorithms.

Algorithms	PT.01	PT.02
LRC	0.71	0.80
KNN	0.82	0.80
SVC	0.76	0.80
NBC	0.71	1.00
DTC	0.82	1.00
RFC	0.76	0.90

All of the algorithms achieved an accuracy level beyond 70.00%. The accuracy level of all algorithms exceeded 70% limit, only the performance of LRC and NBC in PT.01 group demonstrates a degree of accuracy that is nearly sufficient.

Nevertheless, it is important to note that larger accuracy percentages may not always reflect a high level of classifier performance, as indicated by the accuracy paradox (Zhu and Davidson, 2007). The accuracy computation is derived from the summation of the diagonal elements of the confusion matrix across all data points. However, if the algorithm categorizes all the data into a single class and if the data are imbalanced, meaning there is a large number of data points in one class compared with others, the

accuracy ratio of this approach will be high. Nevertheless, this is not an indication that the performance of the algorithm is high. Therefore, other widely used performance indicators were calculated. The accuracy of classification can be assessed by calculating the number of correctly identified values belonging to a class (true positives: TP), the number of correctly identified values not belonging to the class (true negatives: TN), and values that were either wrongly assigned to the class (false positives: FP) or not recognized as class instances (false negatives: FN) (Sokolova and Lapalme 2009). The values are derived from the binary confusion matrix. In the context of multi-class classification, the confusion matrix can be simplified to a binary form using the one vs. rest method, and TP, TN, FP, and FN values in this matrix can then be computed accordingly. equations (4.4), (4.5), (4.6) and (4.7) were employed to calculate Recall, Specificity, Precision, and F1 Score based on these data.

$$Recall = \frac{TP}{TP + FN} \quad (4.4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4.5)$$

$$Precision = \frac{TP}{TP + FP} \quad (4.6)$$

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (4.7)$$

Recall, referred to as sensitivity or True Positive Rate, is a measure used to determine the proportion of true positive results in a classification algorithm. A high recall score indicates that the classifier accurately recognizes a majority of the positive cases. Specificity refers to the ratio of TN to all actual negatives. A high specificity score indicates that the classifier correctly detects negative cases. Precision is the proportion of TP out of the predicted positives. A high precision score indicates that the classifier accurately predicts positive instances. The F1 Score is a commonly used performance metric that effectively combines precision and recall. The performance metrics of the algorithms for each of the four classes and both groups are reported in Figures 4.15 and Figure 4.16 to demonstrate their testing capabilities.

PT.01										
LRC						KNN				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	0.50	1.00	-	0.63	Recall	1.00	0.83	1.00	-
Specificity	0.70	1.00	0.87	1.00	0.89	Specificity	0.90	1.00	0.87	1.00
Precision	0.70	1.00	0.50	-	0.55	Precision	0.88	1.00	0.50	-
F1 Score	0.82	0.67	0.67	-	0.54	F1 Score	0.93	0.91	0.67	-

SVC						NBC				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	0.67	1.00	-	0.67	Recall	1.00	0.50	1.00	-
Specificity	0.70	1.00	0.93	1.00	0.91	Specificity	0.60	1.00	0.93	1.00
Precision	0.70	1.00	0.67	-	0.59	Precision	0.64	1.00	0.67	-
F1 Score	0.82	0.80	0.80	-	0.61	F1 Score	0.78	0.67	0.80	-

DTC						RFC				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	0.83	1.00	-	0.71	Recall	1.00	0.67	1.00	-
Specificity	0.90	1.00	0.87	1.00	0.94	Specificity	0.80	1.00	0.87	1.00
Precision	0.88	1.00	0.50	-	0.59	Precision	0.78	1.00	0.50	-
F1 Score	0.93	0.91	0.67	-	0.63	F1 Score	0.88	0.80	0.67	-

Figure 4.15 : Evaluation of classifier algorithms based on performance indices for the PT.01 group.

PT.02										
LRC						KNN				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	0.57	0.50	1.00	0.77	Recall	0.89	0.57	1.00	1.00
Specificity	0.73	0.92	1.00	1.00	0.91	Specificity	0.73	0.92	1.00	1.00
Precision	0.75	0.80	1.00	1.00	0.89	Precision	0.73	0.80	1.00	1.00
F1 Score	0.86	0.67	0.67	-	0.55	F1 Score	0.80	0.67	1.00	-

SVC						NBC				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	0.57	1.00	0.50	0.77	Recall	1.00	1.00	1.00	1.00
Specificity	0.64	1.00	1.00	1.00	0.91	Specificity	1.00	1.00	1.00	1.00
Precision	0.69	1.00	1.00	1.00	0.92	Precision	1.00	1.00	1.00	1.00
F1 Score	0.82	0.73	1.00	-	0.64	F1 Score	1.00	1.00	1.00	-

DTC						RFC				
Performance Index	Classes					Performance Index	Classes			
	0	1	2	3	Ave.		0	1	2	3
Recall	1.00	1.00	1.00	1.00	1.00	Recall	1.00	1.00	0.50	0.50
Specificity	1.00	1.00	1.00	1.00	1.00	Specificity	1.00	0.85	1.00	1.00
Precision	1.00	1.00	1.00	1.00	1.00	Precision	1.00	0.78	1.00	1.00
F1 Score	1.00	1.00	1.00	-	0.75	F1 Score	1.00	0.88	0.67	-

Figure 4.16 : Evaluation of classifier algorithms based on performance indices for the PT.02 group.

The results indicate that the models functioned adequately. Based on the average performance scores, DTC and KNN performed most effectively in PT.01. They were

followed by SVC, RFC, NBC, and LRC in that order. In the PT.02 group, DTC and NBC achieved the highest results, followed by RFC, SVC, KNN, and LRC.

Subsequently, the models were evaluated utilizing validation techniques to obtain comprehensive performance indices, as determined by confusion matrices. Resampling is a widely used method in ML for evaluating and selecting models. Various resampling methods include bootstrapping, k-fold cross-validation, leave-one-out cross-validation, and repeated holdout validation. This study employed repeated random subsampling validation, also called “Monte Carlo cross-validation” or “shuffle-split cross-validation”. This approach employs the holdout method iteratively k times, with different random seeds. The objective is to provide a more dependable performance estimate less influenced by how the data are divided into training and test sets (Raschka, 2020). Then, the average performance throughout these k iterations was computed. This method is beneficial for relatively small datasets. Compared with conventional k-fold cross-validation, the training and validation sets can be modified in size, and performance assessment is not excessively influenced by a certain division of data. Repeated random subsampling validation and k-fold cross-validation exhibit equivalent validation performance (Li et al., 2010).

Table 4.13 presents the overall accuracy of the cross-validation. Confusion matrices for randomly selecting states were calculated by repeating the process ten times for the PT.01 and PT.02 groups.

Table 4.13 : Overall accuracy values for 10 random states.

PT.01												
Algorithms	ACC										Ave.	Rank
	RS=1	RS=2	RS=3	RS=4	RS=5	RS=6	RS=7	RS=8	RS=9	RS=10		
LRC	0.71	0.82	0.65	0.53	0.76	0.71	0.88	0.65	0.71	0.82	0.72	6
KNN	0.82	0.82	1.00	0.82	0.94	1.00	1.00	0.82	0.88	0.88	0.90	2
SVC	0.76	0.76	0.76	0.71	0.88	0.71	0.88	0.71	0.82	0.88	0.79	5
NBC	0.71	1.00	0.82	0.71	0.76	0.94	0.76	0.88	0.76	0.94	0.83	4
DTC	0.82	1.00	1.00	0.82	0.94	1.00	1.00	1.00	0.88	0.94	0.94	1
RFC	0.76	1.00	1.00	0.71	0.82	1.00	1.00	0.94	0.76	0.88	0.89	3
PT.02												
Algorithms	ACC										Ave.	Rank
	RS=1	RS=2	RS=3	RS=4	RS=5	RS=6	RS=7	RS=8	RS=9	RS=10		
LRC	0.80	0.90	0.65	0.55	0.65	0.75	0.85	0.85	0.75	0.85	0.76	5
KNN	0.80	0.95	0.80	0.75	0.85	0.75	0.60	0.90	0.80	0.90	0.81	4
SVC	0.80	0.95	0.65	0.45	0.55	0.65	0.55	0.85	0.75	0.80	0.70	6
NBC	1.00	0.85	0.80	0.90	0.90	0.85	0.95	0.85	0.85	0.95	0.89	3
DTC	1.00	1.00	0.80	1.00	0.95	1.00	0.95	1.00	0.95	0.95	0.96	1
RFC	0.90	1.00	0.85	0.85	0.90	0.95	0.95	0.80	0.90	0.90	0.90	2

Based on the validation groups, LRC and SVC did not achieve an accuracy of 0.80 for PT.01 and PT.02. However, DTC demonstrated higher accuracy performance for both groups, followed by KNN and RFC in PT.01, and RFC and NBC in PT.02.

Subsequently, the cross-validation performance indices were computed by repeating the procedure ten times to generate confusion matrices and performance indices for each class within the PT.01 and PT.02 groups. The average scores of performance indicators for four classes are computed and presented in Table 4.14 for each classifier.

Table 4.14 : Summary of average performance indices for classifiers from cross-validation.

PT.01					
Algorithms	Recall (avg)	Specificity (avg)	Precision (avg)	F1 Score (avg)	Ave.
KNN	0.66	0.96	0.65	0.64	0.73
NBC	0.61	0.94	0.61	0.58	0.68
DTC	0.71	0.98	0.70	0.69	0.77
RFC	0.63	0.96	0.61	0.60	0.70
PT.02					
Algorithms	Recall (avg)	Specificity (avg)	Precision (avg)	F1 Score (avg)	Ave.
KNN	0.80	0.91	0.82	0.80	0.83
NBC	0.78	0.96	0.72	0.72	0.79
DTC	0.91	0.99	0.92	0.91	0.93
RFC	0.72	0.95	0.77	0.72	0.79

The DTC demonstrates outstanding predictive performance in both groups when evaluating performance indices. In PT.01, KNN and RFC both show strong performance, while in PT.02, KNN, NBC, and RFC exhibit notable performance.

The DTC method demonstrates superior results for both groups. To enhance the algorithm's performance, the parameters of the algorithm were adjusted. Grid Search is an optimization technique that involves systematically testing different parameter values for an algorithm to fine-tune its performance for a given data set. It is known for its efficiency in terms of computational time and resource usage (Kotu and Deshpande, 2014; Mahmoodzadeh et al., 2022; Wauters and Vanhoucke, 2014). The algorithm parameters of this classifier consist of "criterion" and "splitter". The "criterion" refers to the function used to assess the quality of a split, while "splitter" denotes the approach employed to select the split at each node. The options for "criterion" and "splitter" are {"gini", "entropy", "log_loss"} and {"best", "random"}, respectively. All combinations for ten random states were obtained by ten times run, and the average accuracy scores of these states for PT.01 and PT.02 are presented in Figure 4.17.

PT.01		splitter		PT.02		splitter	
criterion		best	random	criterion		best	random
	gini	0.9208	0.8842		gini	0.9708	0.9283
	entropy	0.9233	0.9000		entropy	0.9708	0.9083
	log_loss	0.9150	0.8783		log_loss	0.9708	0.9342

*The bold numbers indicate the superior results.

Figure 4.17 : Average accuracy score matrices of grid search.

0.9233 in PT.01 and 0.9708 in PT.02 group are the superior score outcomes of the parameter combinations. The optimal combination for PT.01 was “criterion” = “entropy” and “splitter” = “best”, but for the PT.02 group, “splitter” = “best” gave superior results with all criterion alternatives.

4.3 Results

Regression and classification are utilized to predict the cost of a construction project throughout the execution phase, using the overall risk scores as a basis. The regression section employs many regression techniques, including LR, PR2, PR4, SVR, DT, and RF. Based on the R2 and cross-validation results, PR4 is chosen as the optimal representation of the relationship between the total risk score and standardized total cost value for the PT.01 and PT.02 groups. The level of functionality is assessed using cross-validation estimation and mean squared error. The PT.01 group benefits the most from a fourth-degree function, whereas a third-degree function is sufficient for the PT.02 group. It is followed by LR and PT2.

In the classification part, four cost overrun ratio classes were determined. Then, these classes were analyzed using LRC, KNN, SVC, NB, DTC, and RFC based on the total risk score. According to the confusion matrix, cross-validation, and performance indices (including recall, specificity, precision, and F1 score), the DTC is the most effective classification method for the PT.01 and PT.02 groups. It is followed by KNN for the PT.01 group and the RFC for the PT.02 group. The findings are briefly presented in Table 4.15.

Table 4.15 : Prediction and classification algorithms ranking for PT.01 and PT.02 groups.

	Rank	PT.01	PT.02		Rank	PT.01	PT.02
Regression (R ²)	1	PR4	PR3	Classification (Accuracy)	1	DTC	DTC
	2	LR	PR2		2	KNN	RFC
	3	PR2	LR		3	RFC	GNB



5. EVALUATION AND CONCLUSION

5.1 Benefits of the proposed approaches

The primary benefit of this approach is its ease of application in the early stages of projects. This is achieved by utilizing data on the total cost and total risk score of previously completed projects. Furthermore, it can be deployed or modified to projects during the execution phase at any time by utilizing the updated cost and total risk score data. Moreover, the suggested approach covered the financial implications of risks, which are excluded from cost estimations and mainly assessed in risk registers. Additionally, the proposed approach is simple to use because it employs total cost and risk scores rather than their breakdowns. The advantages of using total cost and risk scores instead of breakdowns are summarized as follows:

- **The Ease of Application:** Calculating cumulative risk scores utilizing the risk registers maintained for each project does not provide a significant challenge.
- **Cost Breakdown Standardization:** The inclusion of total cost as a factor renders cost breakdown standardization beyond the scope of consideration. Certain cost elements, such as machinery equipment and consultation expenditures, have the potential to be categorized into various cost groups.
- **More Applicable Results:** Establishing a coherent correlation between the overall cost and the total risk factors could potentially yield more relevant outcomes that can be readily applied to comparable projects.

Another advantage of this approach is that the cost overrun ratio classes offer a clear comprehension of the gravity of cost escalations. Also, executives may develop a mitigating action plan based on the expected percentage increase in total cost. Furthermore, forecasting total completion costs based on risk scores expedites the generation of risk-informed projections during the project's execution phase.

5.2 Limitations of study

Certain limitations might be considered that have potential for improvement within the framework of the proposed approach. Firstly, there are risk perception characteristics that might influence the calculation of total risk score from one project to another. Furthermore, the cost overrun ratio may be categorized differently based on the project's features. Similarly, each company can assess the risk probability and impact scales variably, influenced by project type and organizational dynamics, and may use different ML algorithms based on the nature of their data. Furthermore, every project generates an equivalent amount of data as it updates its risks and costs. Hence, there might be a constraint on the quantity of data available. Moreover, it is known that receiving updates on the risk and cost of a construction project during the execution phase is more challenging compared with obtaining data regarding the tender cost or completion cost at the end of the project. Ultimately, the DTC algorithm, which is the best-performing algorithm on the data in this study, has a tendency to be biased towards dominant features, which might result in overfitting the training dataset. Therefore, the selection of algorithms and parameters should be contingent on the specific characteristics of the data.

The analysis can be improved by utilizing cost and/or risk breakdown structures to examine the most closely related components. Furthermore, extensive datasets can be analyzed with alternative machine learning methods that are more adept at managing substantial volumes of data. Moreover, investigating other construction project types beyond residential and heavy civil projects may prove beneficial. These limitations may diverge the results among different companies and alternative ML algorithms.

5.3 Recommendations for future research

This study forecasted overall completion costs and cost overrun percentages using machine learning prediction and classification models based on the total risk score. This method not only streamlines the estimation process during the project's execution phase but may also inspire further research on this subject. Initially, risk perception could be incorporated into the models as a quantitative metric. The project managers may assess themselves, or their team may evaluate them through various tests available in literature, and these values can be utilized in predictive and categorization models.

Furthermore, the cost and risk breakdown structure may be utilized. Ultimately, assessing the correlation between cost and risk factors is valuable based on their breakdowns.

5.4 Conclusion

Cost estimation plays a crucial role in the construction industry. Over the years, numerous researchers have focused on this subject and assessed the total completion cost or cost overrun ratio using various methodologies. However, there is a need for cost estimation models during the execution phase of construction projects. During the execution phase, the intense workload limits the time available for cost estimation compared with earlier phases. Consequently, it is essential to deliver estimations promptly and efficiently throughout this period. Furthermore, contract terms or changes in the scope of work and associated risks due to site conditions are primarily incorporated into cost estimation processes after identification. However, it is challenging to periodically integrate all detected risks into the completion cost. Utilizing risk scores reported in monthly risk registers presents an opportunity to apply these scores to the completion cost as a cost overrun ratio. A swift and user-friendly approach was proposed to address this gap. This approach suggests utilizing ML prediction and classification algorithms to forecast the completion cost and class of cost overrun ratios based on total risk scores.

In the regression models of machine learning, LR, PR2, PR4, SVR, DT, and FR were utilized to forecast the total completion cost of the project based on the total risk score. Polynomial regression surpassed previous algorithms in forecasting project completion costs by analyzing the disparity between projected and actual outcomes while accounting for resampling techniques and bias-variance trade-offs in the collected data. Upon system optimization, it was determined that the 3rd degree produced greater outcomes for residential projects, but the 4th degree offered enhanced predictive results for heavy civil construction projects, according to the data gathered for this study.

In classification models LRC, KNN, SVC, NBC, DTC, and RFC were employed owing to their widespread application, ease of use, and capacity to evaluate performance. The efficiencies of the algorithms were evaluated using accuracy and additional performance metrics such as recall, specificity, precision, F1 Score, and

resampling techniques. Data from eleven projects were segregated into two groups, and six proposed ML classification algorithms were employed. DTC, KNN, and RFC in the PT.01 group and DTC, KNN, NBC, and RFC in the PT.02 group demonstrated superior performance based on performance indices. After evaluating models with repeated random sub-sampling validation, the DTC algorithm yielded the most favorable outcomes in terms of overall accuracy and performance indices for both groups. Grid Search fine-tuned the parameters of this algorithm. The optimal combinations in DTC were determined as entropy-best in PT.01 and gini/entropy/log_loss-best in PT.02.

Although the proposed approach has some limitations, including risk perception about evaluating risks, data shortages or restrictions to gather, and selection of the proper ML algorithm depending on data characteristics, the suggested approach offers benefits in terms of its user-friendly implementation and the prompt and meaningful outcomes it provides to project executives at any stage of a construction project. Additionally, this study facilitates future investigations and analyses in the construction field.

REFERENCES

- A Kassem, M., Khoiry, M. A. and Hamzah, N.** (2021). Theoretical review on critical risk factors in oil and gas construction projects in Yemen, *Engineering, Construction and Architectural Management*, 28(4): 934–968.
- Abdollahzadeh, G. and Rastgoo, S.** (2015). Risk Assessment in Bridge Construction Projects Using Fault Tree and Event Tree Analysis Methods Based on Fuzzy Logic, *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part B: Mechanical Engineering*, 1(3): 031006.
- Abdul-Rahman, H., Wang, C. and Lee, Y. L.** (2013). Design And Pilot Run of Fuzzy Synthetic Model (FSM) for Risk Evaluation in Civil Engineering, *Journal of Civil Engineering and Management*, 19(2): 217–238.
- Adabre, M. A., Chan, A. P. C., Edwards, D. J. and Adinyira, E.** (2021). Assessing critical risk factors (CRFs) to sustainable housing: The perspective of a sub-Saharan African country, *Journal of Building Engineering*, 41: 102385.
- Adedokun, O. and Egbelakin, T.** (2024). Structural equation modelling of risk factors influencing the success of building projects, *Journal of Facilities Management*, 22(1): 64–90.
- Adeleke, A. Q., Bahaudin, A. Y. and Kamaruddeen, A. M.** (2018a). Organizational Internal Factors and Construction Risk Management among Nigerian Construction Companies, *Global Business Review*, 19(4): 921–938.
- Adeleke, A. Q., Bahaudin, A. Y., Kamaruddeen, A. M., Bamgbade, J. A., Salimon, M. G., Khan, M. W. A. and Sorooshian, S.** (2018b). The influence of organizational external factors on construction risk management among Nigerian construction companies, *Safety and health at work*, 9(1): 115–124.
- Adeli, H. and Wu, M.** (1998). Regularization neural network for construction cost estimation, *Journal of construction engineering and management*, 124(1): 18–24.
- Aggarwal, C. C.** (2015). *Data Mining: The Textbook*, Cham: Springer International Publishing. Retrieved from <https://link.springer.com/10.1007/978-3-319-14142-8>
- Ahiaga-Dagbui, D. D. and Smith, S. D.** (2014). Rethinking construction cost overruns: cognition, learning and estimation, *Journal of financial management of property and construction*, 19(1): 38–54.
- Akintoye, A. S. and MacLeod, M. J.** (1997). Risk analysis and management in construction, *International journal of project management*, 15(1): 31–38.

- Aladağ, H. and Işık, Z.** (2019). Design and construction risks in BOT type mega transportation projects, *Engineering, Construction and Architectural Management*, 26(10): 2223–2242.
- Albogamy, A. and Dawood, N.** (2015). Development of a client-based risk management methodology for the early design stage of construction processes: Applied to the KSA, *Engineering, Construction and Architectural Management*, 22(5): 493–515.
- Alhammadi, Y., Al-Mohammad, M. S. and Rahman, R. A.** (2024). Modeling the Causes and Mitigation Measures for Cost Overruns in Building Construction: The Case of Higher Education Projects, *Buildings*, 14(2): 487.
- Al-Sabah, R., Menassa, C. C. and Hanna, A.** (2014). Evaluating impact of construction risks in the Arabian Gulf Region from perspective of multinational architecture, engineering and construction firms, *Construction Management and Economics*, 32(4): 382–402.
- Alshboul, O., Shehadeh, A., Almasabha, G. and Almuflih, A. S.** (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction, *Sustainability*, 14(11): 6651.
- Alshihri, S., Al-Gahtani, K. and Almohsen, A.** (2022). Risk factors that lead to time and cost overruns of building projects in Saudi Arabia, *Buildings*, 12(7): 902.
- Alvand, A., Mirhosseini, S. M., Ehsanifar, M., Zeighami, E. and Mohammadi, A.** (2023). Identification and assessment of risk in construction projects using the integrated FMEA-SWARA-WASPAS model under fuzzy environment: a case study of a construction project in Iran, *International Journal of Construction Management*, 23(3): 392–404.
- Al-Zubaidi, E. A., Naji, H. I. and Ali, R. H.** (2017). Descriptive Classification of Cost Risks in Construction Projects, *ARPN Journal of Engineering and Applied Sciences*, 12(15): 4383-4389.
- Aminbakhsh, S., Gunduz, M. and Sonmez, R.** (2013). Safety risk assessment using analytic hierarchy process (AHP) during planning and budgeting of construction projects, *Journal of Safety Research*, 46: 99–105.
- Andi, A.** (2005). Navigational Measures for Managing Defective Designs, *Journal of Management in Engineering*, 21(1): 10–16.
- Andi, A.** (2006). The importance and allocation of risks in Indonesian construction projects, *Construction Management and Economics*, 24(1): 69–80.
- Antoniou, F., Aretoulis, G., Giannoulakis, D. and Konstantinidis, D.** (2023). Cost and material quantities prediction models for the construction of underground metro stations, *Buildings*, 13(2): 382.
- Antwi-Afari, M. F., Li, H., Edwards, D. J., Pärn, E. A., Seo, J. and Wong, A. Y. L.** (2017). Biomechanical analysis of risk factors for work-related musculoskeletal disorders during repetitive lifting task in construction workers, *Automation in Construction*, 83: 41–47.

- Artan Ilter, D. and Bakioglu, G.** (2018). Modeling the Relationship between Risk and Dispute in Subcontractor Contracts, *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 10(1): 04517022.
- Avotos, I.** (1983). Cost-relevance analysis for overrun control, *International Journal of Project Management*, 1(3): 142–148.
- Aziz, R. F.** (2013). Factors causing cost variation for constructing wastewater projects in Egypt, *Alexandria Engineering Journal*, 52(1): 51–66.
- Baghdadi, A. and Kishk, M.** (2015). Saudi Arabian Aviation Construction Projects: Identification of Risks and Their Consequences, *Procedia Engineering*, 123: 32–40.
- Baratta, A.** (2006). The triple constraint, a triple illusion, In *PMI® Global Congress*, Vol. 202.
- Bilal, M. and Oyedele, L. O.** (2020). Guidelines for applied machine learning in construction industry—A case of profit margins estimation, *Advanced Engineering Informatics*, 43: 101013.
- Bin Seddeeq, A., Assaf, S., Abdallah, A. and Hassanain, M. A.** (2019). Time and Cost Overrun in the Saudi Arabian Oil and Gas Construction Industry, *Buildings*, 9(2): 41.
- Breiman, L.** (2001). Random Forests, *Machine Learning*, 45(1): 5–32.
- Bufaied, A. S.** (1987). *Risks in the construction industry: their causes and their effects at the product level*. (PhD Thesis). The University of Manchester, Faculty of Technology, Manchester.
- Bu-Qammaz, A. S., Dikmen, I. and Birgonul, M. T.** (2009). Risk assessment of international construction projects using the analytic network process, *Canadian Journal of Civil Engineering*, 36(7): 1170–1181.
- Calik, I., Gürgün, A. and Özçoban, M.** (2023). Methods and Tools Used for Management of Geotechnical Risks in Construction Projects. *Proceedings of International Structural Engineering and Construction*, 10(1), CPM-07-1: CPM-07-6.
- Carpio-de Los Pinos, A. J., González-García, M. D. L. N., Pentelhão, L. C. and Baptista, J. S.** (2021). Zero-Risk Interpretation in the Level of Preventive Action Method Implementation for Health and Safety in Construction Sites, *International Journal of Environmental Research and Public Health*, 18(7): 3534.
- Ceric, A.** (2014). Minimizing Communication Risk in Construction: A Delphi Study of the Key Role of Project Managers, *Journal Of Civil Engineering and Management*, 20(6): 829–838.
- Cevikbas, M., Okudan, O. and Işık, Z.** (2024). Identification and assessment of disruption claim management risks in construction projects: a life cycle-based approach, *Engineering, Construction and Architectural Management*, 31(1): 1–27.
- Chadee, A., Martin, H., Gallage, S. and Rathnayake, U.** (2023). Reducing Cost Overrun in Public Housing Projects: A Simplified Reference Class Forecast for Small Island Developing States, *Buildings*, 13(4): 998.

- Chan, D. W., Cristofaro, M., Nassereddine, H., Yiu, N. S. and Sarvari, H.** (2021). Perceptions of safety climate in construction projects between workers and managers/supervisors in the developing country of Iran, *Sustainability*, 13(18): 10398.
- Chan, E. H. W. and Au, M. C. Y.** (2007). Building contractors' behavioural pattern in pricing weather risks, *International Journal of Project Management*, 25(6): 615–626.
- Chan, S. L. and Park, M.** (2005). Project cost estimation using principal component regression, *Construction Management and Economics*, 23(3): 295–304.
- Chandra, H. P.** (2015). Structural Equation Model for Investigating Risk Factors Affecting Project Success in Surabaya, *Procedia Engineering*, 125: 53–59.
- Chang, S. E. and Shinozuka, M.** (1996). Life-Cycle Cost Analysis with Natural Hazard Risk, *Journal of Infrastructure Systems*, 2(3): 118–126.
- Cheng, M.Y. and Hoang, N.D.** (2014). Interval Estimation of Construction Cost at Completion using Least Squares Support Vector Machine, *Journal of Civil Engineering And Management*, 20(2): 223–236.
- Choi, J., Gu, B., Chin, S. and Lee, J.S.** (2020). Machine learning predictive model based on national data for fatal accidents of construction workers, *Automation in Construction*, 110: 102974.
- Chou, J.S.** (2011). Cost simulation in an item-based project involving construction engineering and management, *International Journal of Project Management*, 29(6): 706–717.
- Darko, A., Chan, A. P. C., Yang, Y. and Tetteh, M. O.** (2020). Building information modeling (BIM)-based modular integrated construction risk management – Critical survey and future needs, *Computers in Industry*, 123: 103327.
- Das, P., Kashem, A., Hasan, I. and Islam, M.** (2024). A comparative study of machine learning models for construction costs prediction with natural gradient boosting algorithm and SHAP analysis, *Asian Journal of Civil Engineering*, 25(4): 3301–3316.
- De La Cruz, M. P., Del Caño, A. and De La Cruz, E.** (2006). Downside Risks in Construction Projects Developed by the Civil Service: The Case of Spain, *Journal of Construction Engineering and Management*, 132(8): 844–852.
- Derakhshanlavijeh, R. and Teixeira, J. M. C.** (2017). Cost overrun in construction projects in developing countries, Gas-Oil industry of Iran as a case study, *Journal of Civil Engineering and Management*, 23(1): 125–136.
- Desai, V. S. and Joshi, S.** (2010). Application of Decision Tree Technique to Analyze Construction Project Data, In S. K. Prasad, H. M. Vin, S. Sahni, M. P. Jaiswal, & B. Thipakorn (Eds.), *Information Systems, Technology and Management*, Vol. 54, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 304–313. Retrieved from http://link.springer.com/10.1007/978-3-642-12035-0_30

- Diab, M. F., Varma, A. and Panthi, K.** (2017). Modeling the Construction Risk Ratings to Estimate the Contingency in Highway Projects, *Journal of Construction Engineering and Management*, 143(8): 04017041.
- Dikmen, I., Birgonul, M. T. and Han, S.** (2007). Using fuzzy risk assessment to rate cost overrun risk in international construction projects, *International journal of project management*, 25(5): 494–505.
- Ding, L. Y., Zhong, B. T., Wu, S. and Luo, H. B.** (2016). Construction risk knowledge management in BIM using ontology and semantic web technology, *Safety Science*, 87: 202–213.
- Do, T. M. D. and Lam, T. Q. K.** (2019). Analysis of Risk Problems in Construction by R software, *International Journal of Engineering and Advanced Technology*, 8(5): 1872–1875.
- Doloi, H.** (2012). Understanding Impacts of Time and Cost Related Construction Risks on Operational Performance of PPP Projects, *International Journal of Strategic Property Management*, 16(3): 316–337.
- Doloi, H.** (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects, *Journal of construction engineering and management*, 139(3): 267–279.
- Douglas, M. and Wildavsky, A.** (1990). Theories of risk perception: who fears what and why, *Deadalus*, 119(4): 41–60.
- Drozd, W.** (2017). Logistic regression in the identification of hazards in construction, In *IOP Conference Series: Materials Science and Engineering*, Vol. 245, IOP Publishing, p. 062012. Retrieved from <https://iopscience.iop.org/article/10.1088/1757-899X/245/6/062012/meta>
- El-Adaway, I. H. and Kandil, A. A.** (2010). Construction Risks: Single versus Portfolio Insurance, *Journal of Management in Engineering*, 26(1): 2–8.
- Elhegazy, H., Chakraborty, D., Elzarka, H., Ebid, A. M., Mahdi, I. M., Aboul Haggag, S. Y. and Abdel Rashid, I.** (2022). Artificial Intelligence for Developing Accurate Preliminary Cost Estimates for Composite Flooring Systems of Multi-Storey Buildings, *Journal of Asian Architecture and Building Engineering* 21(1): 120–132.
- El-Sayegh, Sameh M., Manjikian, S., Ibrahim, A., Abouelyousr, A. and Jabbour, R.** (2021). Risk identification and assessment in sustainable construction projects in the UAE, *International Journal of Construction Management*, 21(4): 327–336.
- El-Sayegh, S. M.** (2008). Risk assessment and allocation in the UAE construction industry, *International Journal of Project Management*, 26(4): 431–438.
- Eskander, R. F. A.** (2018). Risk assessment influencing factors for Arabian construction projects using analytic hierarchy process, *Alexandria Engineering Journal*, 57(4): 4207–4218.
- Etemadinia, H. and Tavakolan, M.** (2021). Using a hybrid system dynamics and interpretive structural modeling for risk analysis of design phase of the

construction projects, *International Journal of Construction Management*, 21(1): 93–112.

- Fan, C., Binchao, D. and Yin, Y.** (2023). Hierarchical structure and transfer mechanism to assess the scheduling-related risk in construction of prefabricated buildings: an integrated ISM–MICMAC approach, *Engineering, Construction and Architectural Management*, 30(7): 2991–3013.
- Fazeli, A., Dashti, M. S., Jalaei, F. and Khanzadi, M.** (2021). An integrated BIM-based approach for cost estimation in construction projects, *Engineering, Construction and Architectural Management*, 28(9): 2828–2854.
- Fernandez-Dengo, M., Naderpajouh, N. and Hastak, M.** (2013). Risk Assessment for the Housing Market in Mexico, *Journal of Management in Engineering*, 29(2): 122–132.
- Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Holm, M. S., Lovallo, D. and Lunn, D.** (2018). Five things you should know about cost overrun, *Transportation research part A: policy and practice*, 118: 174–190.
- Fong, S. W.** (1987). Risk management, *The Cost Engineer*, 25: 12–16.
- Forgues, D., Iordanova, I., Valdivesio, F. and Staub-French, S.** (2012). Rethinking the Cost Estimating Process through 5D BIM: A Case Study, In *Construction Research Congress 2012*, Presented at the Construction Research Congress 2012, West Lafayette, Indiana, United States: American Society of Civil Engineers, pp. 778–786. Retrieved from <http://ascelibrary.org/doi/10.1061/9780784412329.079>
- Frimpong, Y., Oluwoye, J. and Crawford, L.** (2003). Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study, *International Journal of project management*, 21(5): 321–326.
- Gondia, A., Siam, A., El-Dakhakhni, W. and Nassar, A. H.** (2020). Machine learning algorithms for construction projects delay risk prediction, *Journal of Construction Engineering and Management*, 146(1): 04019085.
- Guo, D., Song, Z., Xu, T., Zhang, Y. and Ding, L.** (2022). Coupling Analysis of Tunnel Construction Risk in Complex Geology and Construction Factors, *Journal of Construction Engineering and Management*, 148(9): 04022097.
- Gurcanli, G. E., Bilir, S. and Sevim, M.** (2015). Activity based risk assessment and safety cost estimation for residential building construction projects, *Safety science*, 80: 1–12.
- Gurgun, A. P., Arditi, D. and Vilar, P. C.** (2016). Impacts of construction risks on costs in leed-certified projects, *Journal of Green Building*, 11(4): 163–181.

- Hallowell, M.** (2010). Safety risk perception in construction companies in the Pacific Northwest of the USA, *Construction Management and Economics*, 28(4): 403–413.
- Hassim, S., Jaafar, M. S. and Hasan Saz, S. A. A.** (2009). The Contractor Perception Towers Industrialised Building System Risk in Construction Projects in Malaysia, *American Journal of Applied Sciences*, 6(5): 937–942.
- Hatefi, S. M. and Tamošaitienė, J.** (2019). An integrated fuzzy dematel-fuzzy anp model for evaluating construction projects by considering interrelationships among risk factors, *Journal of civil engineering and management*, 25(2): 114–131.
- Hegazy, T. and Ayed, A.** (1998). Neural network model for parametric cost estimation of highway projects, *Journal of construction engineering and management*, 124(3): 210–218.
- Hendrickson, C. and Au, T.** (1989). *Project management for construction: Fundamental concepts for owners, engineers, architects, and builders*, Chris Hendrickson.
- Hillson, D.** (2003). Using a risk breakdown structure in project management, *Journal of Facilities management*, 2(1): 85–97.
- Hlaing, N. N., Singh, D., Tiong, R. L. K. and Ehrlich, M.** (2008). Perceptions of Singapore construction contractors on construction risk identification, *Journal of Financial Management of Property and Construction*, 13(2): 85–95.
- Hull, J. K.** (1990). Application of risk analysis techniques in proposal assessment, *International Journal of Project Management*, 8(3): 152–157.
- Hussain, O. A. I., Moehler, R. C., Walsh, S. D. C. and Ahiaga-Dagbui, D. D.** (2024). Minimizing Cost Overrun in Rail Projects through 5D-BIM: A Conceptual Governance Framework, *Buildings*, 14(2): 478.
- Ismail, N. A. A., Yousof, M. N. M. and Adnan, H.** (2021). BIM adoption in managing construction risks amongst Malaysian quantity surveyors: current practice and challenges, *International Journal of Sustainable Construction Engineering and Technology*, 12(3): 166–175.
- ISO 9001:2015—Quality Management Systems—Requirements.** (2015). Geneva, Switzerland: ISO.
- ISO 31000:2018(en), Risk management — Guidelines.** (2018). Geneva, Switzerland: ISO.
- James, G., Witten, D., Hastie, T. and Tibshirani, R.** (2013). *An introduction to statistical learning*, Vol. 112, Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J.** (2023). Statistical learning, In *An Introduction to Statistical Learning: With Applications in Python*, Springer, pp. 15–67.
- Jarkas, A. M. and Haupt, T. C.** (2015). Major construction risk factors considered by general contractors in Qatar, *Journal of Engineering, Design and Technology* 13(1): 165–194.

- Jayasudha, K. and Vidivelli, B.** (2016). Analysis of major risks in construction projects, *ARPN journal of engineering and applied sciences*, 11(11): 6943–6950.
- Jha, K. N. and Devaya, M. N.** (2008). Modelling the risks faced by Indian construction companies assessing international projects, *Construction Management and Economics*, 26(4): 337–348.
- Ji, C., Su, X., Qin, Z. and Nawaz, A.** (2022). Probability analysis of construction risk based on noisy-or gate bayesian networks, *Reliability Engineering & System Safety*, 217: 107974.
- Ji, S.H., Park, M. and Lee, H.S.** (2011). Cost estimation model for building projects using case-based reasoning, *Canadian Journal of Civil Engineering*, 38(5): 570–581.
- Ji, S.H., Park, M. and Lee, H.S.** (2012). Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea, *Journal of Construction Engineering and Management*, 138(1): 43–52.
- Johnson, R. M. and Babu, R. I. I.** (2020). Time and cost overruns in the UAE construction industry: a critical analysis, *International Journal of Construction Management*, 20(5): 402–411.
- Juszczyk, M.** (2020). On the Search of Models for Early Cost Estimates of Bridges: An SVM-Based Approach, *Buildings*, 10(1): 2.
- Kaming, P. F., Olomolaiye, P. O., Holt, G. D. and Harris, F. C.** (1997). Factors influencing construction time and cost overruns on high-rise projects in Indonesia, *Construction Management and Economics*, 15(1): 83–94.
- Kang, L. S., Kim, S.-K., Moon, H. S. and Kim, H. S.** (2013). Development of a 4D object-based system for visualizing the risk information of construction projects, *Automation in Construction*, 31: 186–203.
- Kantianis, D. D.** (2022). Design morphology complexity and conceptual building project cost forecasting, *Journal of Financial Management of Property and Construction*, 27(3): 387–414.
- Khan, A. H.** (2013). The Contractors Perception of Risk Management in Pakistan, *Pakistan Acedmy of Sciences*, 50: 189–200.
- Khanzadi, M., Eshtehardian, E. and Mokhlespour Esfahani, M.** (2017). Cash flow forecasting with risk consideration using bayesian belief networks (BBNS), *Journal of Civil Engineering and Management*, 23(8): 1045–1059.
- Khodeir, L. M. and Mohamed, A. H. M.** (2015). Identifying the latest risk probabilities affecting construction projects in Egypt according to political and economic variables. From January 2011 to January 2013, *HBRC Journal* 11(1): 129–135.
- Kim, G. H., Seo, D. S. and Kang, K. I.** (2005). Hybrid Models of Neural Networks and Genetic Algorithms for Predicting Preliminary Cost Estimates, *Journal of Computing in Civil Engineering*, 19(2): 208–211.

- Kim, G.H., An, S.-H. and Kang, K.I.** (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning, *Building and Environment*, 39(10): 1235–1242.
- Kim, K. J. and Kim, K.** (2010). Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms, *Journal of Computing in Civil Engineering*, 24(6): 499–505.
- Koc, K., Ekmekcioğlu, Ö. and Gurgun, A. P.** (2021). Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers, *Automation in Construction*, 131: 103896.
- Koc, K., Ekmekcioğlu, Ö. and Gurgun, A. P.** (2022). Accident prediction in construction using hybrid wavelet-machine learning, *Automation in Construction*, 133: 103987.
- Kotu, V. and Deshpande, B.** (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*, Morgan Kaufmann.
- Koulinas, G. K., Xanthopoulos, A. S., Sidas, K. A. and Koulouriotis, D. E.** (2021). Risks Ranking in a Desalination Plant Construction Project with a Hybrid AHP, Risk Matrix, and Simulation-Based Approach, *Water Resources Management*, 35(10): 3221–3233.
- Kuhn, M. and Johnson, K.** (2013). *Applied Predictive Modeling*, New York, NY: Springer New York. Retrieved from <http://link.springer.com/10.1007/978-1-4614-6849-3>
- Kuo, Y.C. and Lu, S.T.** (2013). Using fuzzy multiple criteria decision making approach to enhance risk assessment for metropolitan construction projects, *International Journal of Project Management*, 31(4): 602–614.
- Lapidus, A. A., Kuzmina, T. K. and Bolshakova, P. V.** (2024). Developing administrative and process-related solutions for the benefit of residential construction projects, *Construction materials and products*, 7(6): 2–2.
- Lapidus, A., Topchiy, D., Kuzmina, T. and Chapidze, O.** (2022). Influence of the Construction Risks on the Cost and Duration of a Project, *Buildings*, 12(4): 484.
- Laryea, S. and Hughes, W.** (2008). How contractors price risk in bids: theory and practice, *Construction Management and Economics*, 26(9): 911–924.
- Lee, S.K., Kim, K.R. and Yu, J.H.** (2014). BIM and ontology-based approach for building cost estimation, *Automation in construction*, 41: 96–105.
- Leśniak, A. and Zima, K.** (2018). Cost Calculation of Construction Projects Including Sustainability Factors Using the Case Based Reasoning (CBR) Method, *Sustainability*, 10(5): 1608.
- Li, D.C., Fang, Y.H. and Fang, Y. F.** (2010). The data complexity index to construct an efficient cross-validation method, *Decision Support Systems*, 50(1): 93–102.

- Lin, S.S., Zhang, N., Zhou, A. and Shen, S.L.** (2022). Risk evaluation of excavation based on fuzzy decision-making model, *Automation in Construction*, 136: 104143.
- Linnenluecke, M. K., Marrone, M. and Singh, A. K.** (2020). Conducting systematic literature reviews and bibliometric analyses, *Australian Journal of Management*, 45(2): 175–194.
- Liu, L. and Napier, Z.** (2010). The accuracy of risk-based cost estimation for water infrastructure projects: preliminary evidence from Australian projects, *Construction Management and Economics*, 28(1): 89–100.
- Long, N. D., Ogunlana, S., Quang, T. and Lam, K. C.** (2004). Large construction projects in developing countries: a case study from Vietnam, *International Journal of project management*, 22(7): 553–561.
- Lowe, D. J., Emsley, M. W. and Harding, A.** (2006). Predicting Construction Cost Using Multiple Regression Techniques, *Journal of Construction Engineering and Management*, 132(7): 750–758.
- Luu, V. T., Kim, S.-Y., Tuan, N. V. and Ogunlana, S. O.** (2009). Quantifying schedule risk in construction projects using Bayesian belief networks, *International Journal of Project Management*, 27(1): 39–50.
- Ma, Z., Wei, Z. and Zhang, X.** (2013). Semi-automatic and specification-compliant cost estimation for tendering of building projects based on IFC data of design model, *Automation in Construction*, 30: 126–135.
- Mahmoodzadeh, A., Nejati, H. R. and Mohammadi, M.** (2022). Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects, *Automation in Construction*, 139: 104305.
- Mahmoudi, A., Abbasi, M., Deng, X., Ikram, M. and Yeganeh, S.** (2020). A novel model for risk management of outsourced construction projects using decision-making methods: a case study, *Grey Systems: Theory and Application*, 10(2): 97–123.
- Makovšek, D.** (2014). Systematic construction risk, cost estimation mechanism and unit price movements, *Transport Policy*, 35: 135–145.
- Mañelele, I. and Muya, M.** (2008). Risk identification on community-based construction projects in Zambia, *Journal of Engineering, Design and Technology*, 6(2): 145–161.
- Mansfield, N. R., Ugwu, O. O. and Doran, T.** (1994). Causes of delay and cost overruns in Nigerian construction projects, *International journal of project Management*, 12(4): 254–260.
- Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T. and Voordijk, H.** (2022). An artificial neural network approach for cost estimation of engineering services, *International Journal of Construction Management*, 22(7): 1274–1287.
- Mhady, A. A., Budayan, C. and Gurgun, A. P.** (2024). Estimate-at-completion (EAC) prediction using Archimedes optimization with adaptive fuzzy and neural networks, *Automation in Construction*, 166: 105653.

- Mistikoglu, G., Gerek, I. H., Erdis, E., Usmen, P. M., Cakan, H. and Kazan, E. E.** (2015). Decision tree analysis of construction fall accidents involving roofers, *Expert Systems with Applications*, 42(4): 2256–2263.
- Moreno, P. D. C.** (2020). Machine learning: Choosing the right estimator, *Medium*. Retrieved from <https://caiomsouza.medium.com/machine-learning-choosing-the-right-estimator-scikit-learn-algorithm-cheat-sheet-c51500772ac2>
- Motawa, I. A., Anumba, C. J. and El-Hamalawi, A.** (2006). A fuzzy system for evaluating the risk of change in construction projects, *Advances in Engineering Software*, 37(9): 583–591.
- Namian, M., Albert, A., Zuluaga, C. M. and Behm, M.** (2016). Role of Safety Training: Impact on Hazard Recognition and Safety Risk Perception, *Journal of Construction Engineering and Management*, 142(12): 04016073.
- Nasirzadeh, F., Afshar, A., Khanzadi, M. and Howick, S.** (2008). Integrating system dynamics and fuzzy logic modelling for construction risk management, *Construction Management and Economics*, 26(11): 1197–1212.
- Nasirzadeh, F., Khanzadi, M. and Rezaie, M.** (2014). Dynamic modeling of the quantitative risk allocation in construction projects, *International Journal of Project Management*, 32(3): 442–451.
- Naveen, N., Harshit, N., Muthu, D., Venkatasubramaniam, C. and Sharmila, P.** (2018). Post Project Assessment of Risk in Construction Projects, *International Journal of Engineering & Technology*, 7(3.12): 185.
- Nyqvist, R., Peltokorpi, A. and Seppänen, O.** (2024). Uncertainty network modeling method for construction risk management, *Construction Management and Economics*, 42(4): 346–365.
- Odeck, J.** (2004). Cost overruns in road construction—what are their sizes and determinants?, *Transport Policy*, 11(1): 43–53.
- Ofori-Boadu, A. N.** (2015). Exploring regression models for forecasting early cost estimates for high-rise buildings, *The Journal of Technology, Management, and Applied Engineering*, 31(5).
- Ökmen, Ö. and Öztaş, A.** (2010). Construction cost analysis under uncertainty with correlated cost risk analysis model, *Construction Management and Economics*, 28(2): 203–212.
- Öztaş, A. and Ökmen, Ö.** (2004). Risk analysis in fixed-price design–build construction projects, *Building and environment*, 39(2): 229–237.
- Park, D. and Yun, S.** (2023). Construction Cost Prediction Using Deep Learning with BIM Properties in the Schematic Design Phase, *Applied Sciences*, 13(12): 7207.
- Park, U., Kang, Y., Lee, H. and Yun, S.** (2022). A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs, *Applied Sciences*, 12(19): 9729.

- Patterson, F. D. and Neailey, K.** (2002). A risk register database system to aid the management of project risk, *International journal of project management*, 20(5): 365–374.
- Petroutsatou, K., Georgopoulos, E., Lambropoulos, S. and Pantouvakis, J. P.** (2012). Early Cost Estimating of Road Tunnel Construction Using Neural Networks, *Journal of Construction Engineering and Management*, 138(6): 679–687.
- PMBOK Guide | Project Management Institute.** (2017). Retrieved January 18, 2025, from <https://www.pmi.org/standards/pmbok>.
- Pooladvand, S. and Hasanzadeh, S.** (2023). Impacts of Stress on Workers' Risk-Taking Behaviors: Cognitive Tunneling and Impaired Selective Attention, *Journal of Construction Engineering and Management*, 149(8): 04023060.
- Radi Al-Aga, S. F. and Burhan, A. M.** (2022). Appropriate Risk Response Planning of Build Operate Transfer Contracts for the Transportation Projects in Iraq, *Civil and Environmental Engineering*, 18(2): 430–437.
- Rafiei, M. H. and Adeli, H.** (2018). Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes, *Journal of Construction Engineering and Management*, 144(12): 04018106.
- Rani, H. A., Syammaun, T., Azzahra, F., Ayob, A., Amlus, M. H. and Aqsha, M. S.** (2024). Risk Management Planning by Risk Register in Building Construction Project, In *IOP Conference Series: Earth and Environmental Science*, Vol. 1303, IOP Publishing, p. 012034. Retrieved from <https://iopscience.iop.org/article/10.1088/1755-1315/1303/1/012034/meta>.
- Raschka, S.** (2020). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Retrieved from <http://arxiv.org/abs/1811.12808>
- Rehman, M. A. and Ishak, M. S. B.** (2022). Role of government acts, laws and policies between effective communication and construction risk management, *Journal of Applied Structural Equation Modeling*, 6(2): 1–22.
- Rezaie, K., Amalnik, M. S., Gereie, A., Ostadi, B. and Shakhseeniaee, M.** (2007). Using extended Monte Carlo simulation method for the improvement of risk management: Consideration of relationships between uncertainties, *Applied Mathematics and Computation*, 190(2): 1492–1501.
- Rostami, A. and Oduoza, C. F.** (2017). Key risks in construction projects in Italy: contractors' perspective, *Engineering, Construction and Architectural Management*, 24(3): 451–462.
- Rowlinson, S., YunyanJia, A., Li, B. and ChuanjingJu, C.** (2014). Management of climatic heat stress risk in construction: A review of practices, methodologies, and future research, *Accident Analysis & Prevention*, 66: 187–198.

- Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A. and Shamshirband, S.** (2020). A new K-nearest neighbors classifier for big data based on efficient data pruning, *Mathematics*, 8(2): 286.
- Sastoque, L. M., Arboleda, C. A. and Ponz, J. L.** (2016). A Proposal for Risk Allocation in Social Infrastructure Projects Applying PPP in Colombia, *Procedia Engineering*, 145: 1354–1361.
- Sathishkumar, V., Raghunath, P. N. and Suguna, K.** (2015). Critical Factors Influencing to Management Risk in Construction Projects, *The International Journal of Engineering and Science*, 4(1): 37–46.
- Scott, E. and Bommu, R.** (2024). Efficient Construction Management: AI-Driven Strategies to Combat Cost Overruns, *International Journal of Advanced Engineering Technologies and Innovations*, 1(3): 222–240.
- Senthil, J., Muthukannan, M., Urbański, M., Stępień, M. and Kądziałowski, G.** (2021). MSCA based Deep Recurrent Neural Network for Statistics Risk Management in Construction Projects., *Acta Montanistica Slovaca*, 26(3).
- Shaktawat, A. and Vadhera, S.** (2021). Risk management of hydropower projects for sustainable development: a review, *Environment, Development and Sustainability*, 23(1): 45–76.
- Shan, M., Le, Y., Yiu, K. T. W., Chan, A. P. C., Hu, Y. and Zhou, Y.** (2018). Assessing collusion risks in managing construction projects using artificial neural network, *Technological and Economic Development of Economy*, 24(5): 2003–2025.
- Shane, J. S., Molenaar, K. R., Anderson, S. and Schexnayder, C.** (2009). Construction Project Cost Escalation Factors, *Journal of Management in Engineering*, 25(4): 221–229.
- Shen, L.-Y., Platten, A. and Deng, X. P.** (2006). Role of public private partnerships to manage risks in public sector projects in Hong Kong, *International Journal of Project Management*, 24(7): 587–594.
- Shevchenko, G., Ustinovichius, L. and Andruškevičius, A.** (2008). Multi-attribute analysis of investments risk alternatives in construction / statybos alternatyvų investicinės rizikos daugiatislė analizė taikant saw ir clara metodus, *Technological and Economic Development of Economy*, 14(3): 428–443.
- Shevchenko, G., Ustinovichius, L. and Walasek, D.** (2019). The Evaluation of the Contractor's Risk in Implementing the Investment Projects in Construction by Using the Verbal Analysis Methods, *Sustainability*, 11(9): 2660.
- Shoar, S., Chileshe, N. and Edwards, J. D.** (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression, *Journal of Building Engineering*, 50: 104102.
- Siraj, N. B. and Fayek, A. R.** (2019). Risk Identification and Common Risks in Construction: Literature Review and Content Analysis, *Journal of Construction Engineering and Management*, 145(9): 03119004.

- Skorupka, D.** (2008). Identification and Initial Risk Assessment of Construction Projects in Poland, *Journal of Management in Engineering*, 24(3): 120–127.
- Smith, N. J., Merna, T. and Jobling, P.** (2014). *Managing risk in construction projects*, John Wiley & Sons.
- Sohrabi, H. and Noorzai, E.** (2024). Risk-supported case-based reasoning approach for cost overrun estimation of water-related projects using machine learning, *Engineering, Construction and Architectural Management*, 31(2): 544–570.
- Sousa, V., Almeida, N. M. and Dias, L. A.** (2014). Risk-based management of occupational safety and health in the construction industry – Part 1: Background knowledge, *Safety Science*, 66: 75–86.
- Taroun, A.** (2014). Towards a better modelling and assessment of construction risk: Insights from a literature review, *International Journal of Project Management*, 32(1): 101–115.
- Thangavel, D. N. and Manikandan, M.** (2015). Ascertaining Key Risks in the Construction Projects From Stakeholder and Life Cycle Perspectives, *Management Research*, 3(1).
- Thomas, A. V., Kalidindi, S. N. and Ananthanarayanan, K.** (2003). Risk perception analysis of BOT road project participants in India, *Construction Management and Economics*, 21(4): 393–407.
- Thomas, N. and Thomas, A.** (2016). Regression Modelling for Prediction of Construction Cost and Duration, *Applied Mechanics and Materials*, 857: 195–199.
- Tian, Z., Chen, Q. and Zhang, T.** (2022). A method for assessing the crossed risk of construction safety, *Safety Science*, 146: 105531.
- Tixier, A. J.-P., Hallowell, M. R., Albert, A., Van Boven, L. and Kleiner, B. M.** (2014). Psychological Antecedents of Risk-Taking Behavior in Construction, *Journal of Construction Engineering and Management*, 140(11): 04014052.
- Tran, D. Q. and Molenaar, K. R.** (2014). Impact of Risk on Design-Build Selection for Highway Design and Construction Projects, *Journal of Management in Engineering*, 30(2): 153–162.
- Trangkanont, S. and Charoenngam, C.** (2014). Private partner's risk response in PPP low-cost housing projects, *Property Management*, 32(1): 67–94.
- Trillo-Cabello, A. F., Carrillo-Castrillo, J. A. and Rubio-Romero, J. C.** (2021). Perception of risk in construction. Exploring the factors that influence experts in occupational health and safety, *Safety Science*, 133: 104990.
- Tummala, V. R. and Burchett, J. F.** (1999). Applying a risk management process (RMP) to manage cost risk for an EHV transmission line project, *International journal of project management*, 17(4): 223–235.
- Türkel, E. B. and Ergen, E.** (2016). Relationship between height and footprint area with cost for tunnel form high-rise buildings. *Pamukkale University Journal of Engineering Sciences*, 22(6): 418–426

- Valipour, A., Yahaya, N., Md Noor, N., Kildienė, S., Sarvari, H. and Mardani, A.** (2015). A fuzzy analytic network process method for risk prioritization in freeway ppp projects: an iranian case study, *Journal of Civil Engineering and Management*, 21(7): 933–947.
- Van Eck, N. J. and Waltman, L.** (2014). Visualizing Bibliometric Networks, In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact*, Cham: Springer International Publishing, pp. 285–320. Retrieved from https://link.springer.com/10.1007/978-3-319-10377-8_13
- Vidyapriya, P., Mohanasundari, M., Sundharesalingam, P. and Sailendharani, A.** (2019). Impact and assessment of cost overrun due to material cost in construction projects, *International Journal of Civil Engineering and Technology*, 10(2): 1099–1115.
- Viswanathan, S. K. and Jha, K. N.** (2020). Critical risk factors in international construction projects: An Indian perspective, *Engineering, Construction and Architectural Management*, 27(5): 1169–1190.
- Wang, C. M., Xu, B. B., Zhang, S. J. and Chen, Y. Q.** (2016). Influence of personality and risk propensity on risk perception of Chinese construction project managers, *International Journal of Project Management*, 34(7): 1294–1304.
- Wang, D., Dai, F. and Ning, X.** (2015). Risk Assessment of Work-Related Musculoskeletal Disorders in Construction: State-of-the-Art Review, *Journal of Construction Engineering and Management*, 141(6): 04015008.
- Wang, J. Y. and Yuan, H. P.** (2011). Major cost-overrun risks in construction projects in China, *International Journal of Project Organisation and Management*, 3(3/4): 227.
- Wang, R., Asghari, V., Cheung, C. M., Hsu, S.C. and Lee, C.J.** (2022). Assessing effects of economic factors on construction cost estimation using deep neural networks, *Automation in Construction*, 134: 104080.
- Wauters, M. and Vanhoucke, M.** (2014). Support vector machine regression for project control forecasting, *Automation in Construction*, 47: 92–106.
- Weaver, P.** (2008). The meaning of risk in an uncertain world, *PMIVR Global Congress 2008*, Malta, Project Management Institute.
- Wong, C. H.** (2004). Contractor Performance Prediction Model for the United Kingdom Construction Contractor: Study of Logistic Regression Approach, *Journal of Construction Engineering and Management*, 130(5): 691–698.
- Xia, N., Zhong, R., Wu, C., Wang, X. and Wang, S.** (2017). Assessment of Stakeholder-Related Risks in Construction Projects: Integrated Analyses of Risk Attributes and Stakeholder Influences, *Journal of Construction Engineering and Management*, 143(8): 04017030.
- Xia, N., Zou, P. X. W., Griffin, M. A., Wang, X. and Zhong, R.** (2018). Towards integrating construction risk management and stakeholder management: A systematic literature review and future research

agendas, *International Journal of Project Management*, 36(5): 701–715.

Xie, W., Deng, B., Yin, Y., Lv, X. and Deng, Z. (2022). Critical Factors Influencing Cost Overrun in Construction Projects: A Fuzzy Synthetic Evaluation, *Buildings*, 12(11): 2028.

Xie, Y., Ebad Sichani, M., Padgett, J. E. and DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review, *Earthquake Spectra*, 36(4): 1769–1801.

Yan, H., Gao, C., Elzarka, H., Mostafa, K. and Tang, W. (2019). Risk assessment for construction of urban rail transit projects, *Safety Science*, 118: 583–594.

Yang, S.W., Moon, S.W., Jang, H., Choo, S. and Kim, S.A. (2022). Parametric method and building information modeling-based cost estimation model for construction cost prediction in architectural planning, *Applied sciences*, 12(19): 9553.

Yao, B., Wang, L., Gao, H. and Ren, L. (2024). Bridge Construction Risk Assessment Based on Variable Weight Theory and Cloud Model, *Buildings*, 14(3): 576.

Yildiz, A. E., Dikmen, I., Birgonul, M. T., Ercoskun, K. and Alten, S. (2014). A knowledge-based risk mapping tool for cost estimation of international construction projects, *Automation in Construction*, 43: 144–155.

Yoon, Y., Tamer, Z. and Hastak, M. (2015). Protocol to Enhance Profitability by Managing Risks in Construction Projects, *Journal of Management in Engineering*, 31(5): 04014090.

Yousri, E., Sayed, A. E. B., Farag, M. A. M. and Abdelalim, A. M. (2023). Risk Identification of Building Construction Projects in Egypt, *Buildings*, 13(4): 1084.

Zhang, L. and Li, H. (2022). Construction risk assessment of deep foundation pit projects based on the projection pursuit method and improved set pair analysis, *Applied Sciences*, 12(4): 1922.

Zhang, Y., Javanmardi, A., Liu, Y., Yang, S., Yu, X., Hsiang, S. M., Jiang, Z. and Liu, M. (2020). How Does Experience with Delay Shape Managers' Making-Do Decision: Random Forest Approach, *Journal of Management in Engineering*, 36(4): 04020030.

Zhang, Z., Li, W. and Yang, J. (2021). Analysis of stochastic process to model safety risk in construction industry, *Journal of civil engineering and management*, 27(2): 87–99.

Zhao, D., McCoy, A. P., Kleiner, B. M., Mills, T. H. and Lingard, H. (2016). Stakeholder perceptions of risk in construction, *Safety Science*, 82: 111–119.

Zhou, C., Ding, L. Y., Skibniewski, M. J., Luo, H. and Zhang, H. T. (2018). Data based complex network modeling and analysis of shield tunneling performance in metro construction, *Advanced Engineering Informatics*, 38: 168–186.

- Zhou, H., Tang, S., Huang, W. and Zhao, X.** (2023). Generating risk response measures for subway construction by fusion of knowledge and deep learning, *Automation in Construction*, 152: 104951.
- Zhou, H., Zhao, Y., Shen, Q., Yang, L. and Cai, H.** (2020). Risk assessment and management via multi-source information fusion for undersea tunnel construction, *Automation in Construction*, 111: 103050.
- Zhu, X. and Davidson, I.** (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*, Information Science Reference, Hershey, New York.
- Zou, P. X. W., Zhang, G. and Wang, J.** (2007). Understanding the key risks in construction projects in China, *International Journal of Project Management*, 25(6): 601–614.
- Zou, Patrick. X. W. and Zhang, G.** (2009). Managing Risks in Construction Projects: Life Cycle and Stakeholder Perspectives, *International Journal of Construction Management*, 9(1): 61–77.
- Zou, Y., Kiviniemi, A., Jones, S. W. and Walsh, J.** (2019). Risk Information Management for Bridges by Integrating Risk Breakdown Structure into 3D/4D BIM, *KSCE Journal of Civil Engineering*, 23(2): 467–480.



APPENDICES

APPENDIX A: Risk breakdown structure and references table.

APPENDIX B: Flow chart of the research methodology.

APPENDIX C: Python codes machine learning regression algorithms.

APPENDIX D: Python codes machine learning classification algorithms.



[illegible]

Figure A.1 : Risk breakdown structure and references

APPENDIX B

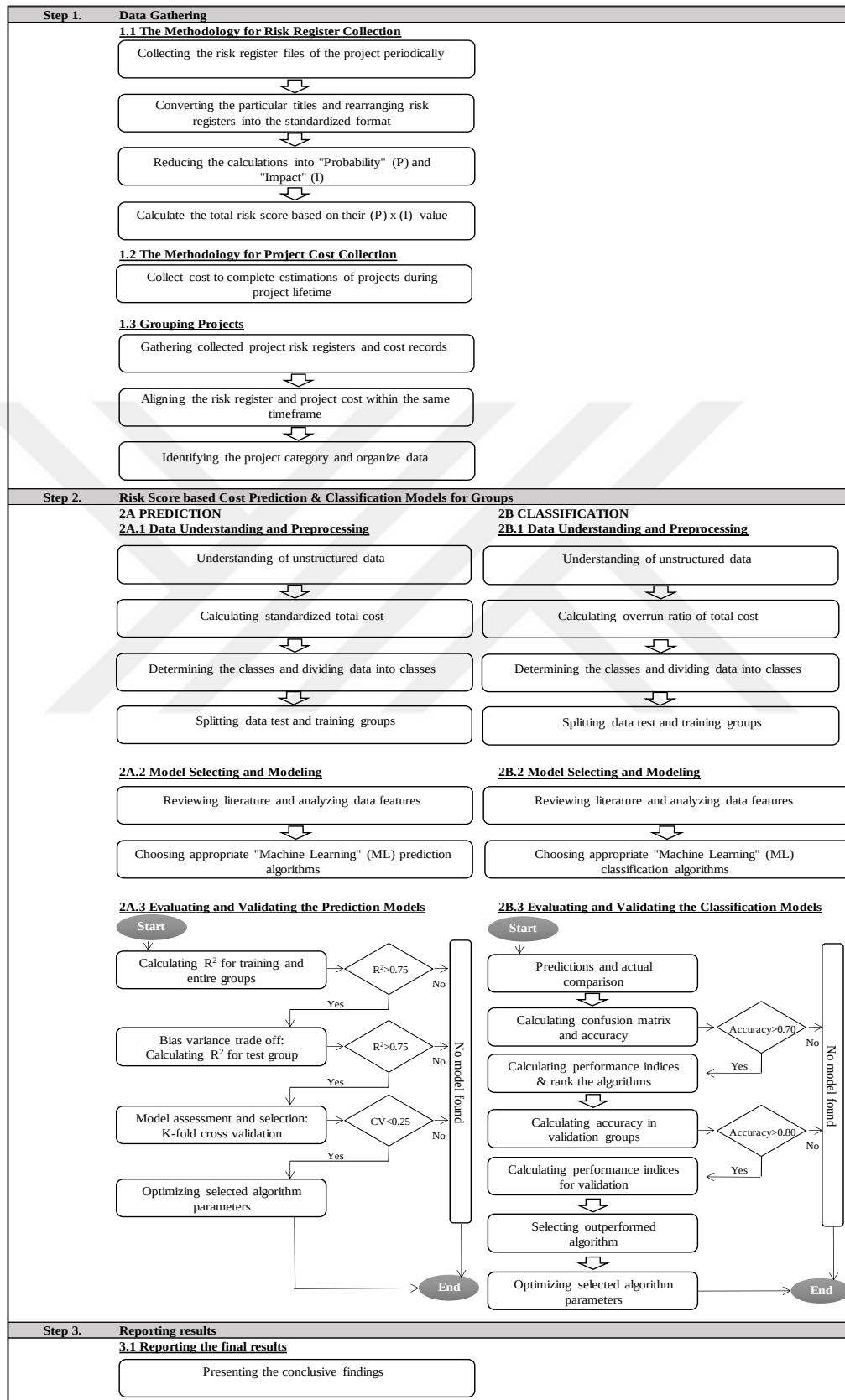


Figure A.2 : Flow chart of the research methodology.

APPENDIX C

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Wed Oct 4 09:49:49 2023
```

```
@author: AynurHurriyet.Sesen
```

```
"""
```

```
# 1. Loading Libraries
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
# 2. Data Preprocessing
```

```
# 2.1. Data Upload
```

```
veriler = pd.read_csv('risks_cost_PT.01_r0.csv')
```

```
# Slicing Data Frame
```

```
x = veriler.iloc[:,1]
```

```
y = veriler.iloc[:,1:]
```

```
#Convert data frame to NumPY array
```

```
X = x.values
```

```
Y = y.values
```

```
#Splittinng train and test groups
```

```
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.33, random_state = 0)
```

```
# LINEAR REGRESSION
```

```
from sklearn.linear_model import LinearRegression
```

```
lin_reg = LinearRegression()
```

```
lin_reg.fit(x_train, y_train)
```

```
Prediction_LR = lin_reg.predict(x_test)
```

```
# Visualization
```

```
#x_train = x_train.sort_index()
```

```
#plt.title('TC z_values depending on Risk Score')
```

```
#plt.xlabel('Risk Score')
```

```

plt.ylabel('TC z_values')

plt.scatter(x_train,y_train,color='red')
plt.plot(x_test, lin_reg.predict(x_test))
plt.show()

# R2 (coefficient of determination) regression score function
from sklearn.metrics import r2_score

print('Linear Regression R2 Score_train group')
print(r2_score(y_train,lin_reg.predict(x_train)))

lin_reg.fit(X, Y)
print('Linear Regression R2 Score')
print(r2_score(Y,lin_reg.predict(X)))

# POLYNOMIAL REGRESSION

# 2nd Degree Polynomial Regression

from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree = 2)
x_poly = poly_reg.fit_transform(x_train)
lin_reg2 = LinearRegression()
lin_reg2.fit(x_poly,y_train)

Prediction_PR2 = lin_reg2.predict(poly_reg.fit_transform(x_test))

plt.scatter(x_train,y_train,color='red')
plt.plot(x_test,lin_reg2.predict(poly_reg.fit_transform(x_test)),color='blue')
plt.show()

# R2 (coefficient of determination) regression score function
print('2nd Degree Polynomial Regression R2 Score_train group')
print(r2_score(y_train,lin_reg2.predict(poly_reg.fit_transform(x_train))))

x_poly = poly_reg.fit_transform(X)
lin_reg2 = LinearRegression()
lin_reg2.fit(x_poly,y)
print('2nd Degree Polynomial Regression R2 Score')
print(r2_score(Y,lin_reg2.predict(poly_reg.fit_transform(X))))

# 4th Degree Polynomial Regression

poly_reg4 = PolynomialFeatures(degree = 4)
x_poly4 = poly_reg4.fit_transform(x_train)
lin_reg4 = LinearRegression()
lin_reg4.fit(x_poly4,y_train)

Prediction_PR4 = lin_reg4.predict(poly_reg4.fit_transform(x_test))

```

```

plt.scatter(x_train,y_train,color='red')
plt.plot(x_test,lin_reg4.predict(poly_reg4.fit_transform(x_test)),color='blue')
plt.show()

# R2 (coefficient of determination) regression score function
print('4th Degree Polynomial Regression R2 Score_train group')
print(r2_score(y_train,lin_reg4.predict(poly_reg4.fit_transform(x_train))))

x_poly4 = poly_reg4.fit_transform(X)
lin_reg4 = LinearRegression()
lin_reg4.fit(x_poly4,y)

print('4th Degree Polynomial Regression R2 Score')
print(r2_score(Y,lin_reg4.predict(poly_reg4.fit_transform(X))))

# SUPPORT VECTOR REGRESSION (SVR)
# Scaling (Oznitelik Olcekleme _Verilerin Olceklenmesi_)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()

x_train_scaled = sc.fit_transform(x_train)
x_test_scaled = sc.fit_transform(x_test)

#Y_train = sc.fit_transform(y_train)
#Y_test = sc.fit_transform(y_test)

from sklearn.svm import SVR
svr_reg = SVR(kernel = 'rbf')
svr_reg.fit(x_train_scaled,y_train)

Prediction_SVR = svr_reg.predict(x_test_scaled)

plt.scatter(x_train_scaled, y_train,color='red')
plt.plot(x_train_scaled, svr_reg.predict(x_train_scaled),color='blue')
plt.show()

# R2 (coefficient of determination) regression score function
print('Support Vector Regression R2 Score_train group')
print(r2_score(y_train,svr_reg.predict(x_train_scaled)))

print('Support Vector Regression R2 Score')
print(r2_score(Y,svr_reg.predict(sc.fit_transform(X))))

# DECISION TREE
from sklearn.tree import DecisionTreeRegressor
r_dt = DecisionTreeRegressor(random_state=0)
r_dt.fit(x_train,y_train)

Prediction_DT = r_dt.predict(x_test)

```

```

plt.scatter(X,Y, color = 'red')
plt.plot(X,r_dt.predict(X), color = 'blue')
plt.show()

# R2 (coefficient of determination) regression score function
print('Decision Tree R2 Score_train group')
print(r2_score(y_train,r_dt.predict(x_train)))

print('Decision Tree R2 Score')
print(r2_score(Y,r_dt.predict(X)))

# RANDOM FOREST

from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor(n_estimators = 10, random_state = 0)
rf_reg.fit(x_train,y_train.ravel())

Prediction_RF = rf_reg.predict(x_test)

plt.scatter(X,Y, color='red')
plt.plot(X,rf_reg.predict(X),color='blue')

# R2 (coefficient of determination) regression score function

print('Random Forest R2 Score_train group')
print(r2_score(y_train,rf_reg.predict(x_train)))

print('Random Forest R2 Score')
print(r2_score(Y,rf_reg.predict(X)))

# PREDICTIONS
print(lin_reg.predict([[10]])) #linear regression
print(lin_reg2.predict(poly_reg.fit_transform([[10]]))) #2nd degree polynomial
regression
print(lin_reg4.predict(poly_reg4.fit_transform([[10]]))) #4th degree polynomial
regression
print(svr_reg.predict([[10]])) #SVR
print(r_dt.predict([[10]])) #Decision Tree
print(rf_reg.predict([[10]])) #Random Forest

```

APPENDIX D

```
# -*- coding: utf-8 -*-  
"""
```

Created on Fri Oct 13 12:19:44 2023

```
@author: AynurHurriyet.Sesen  
"""
```

```
# 1. Loading Libraries
```

```
import pandas as pd  
import numpy as np
```

```
#import numpy as np  
#import matplotlib.pyplot as plt
```

```
# 2. Data Preprocessing  
# 2.1. Data Upload
```

```
veriler = pd.read_csv('classification_PT.01_r0.csv')
```

```
# Slicing Data Frame & Convert data frame to NumPY array (.values)
```

```
x = veriler.iloc[:,0:2].values  
y = veriler.iloc[:,2].values
```

```
#Splitting train and test groups  
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.33, random_state =  
0)
```

```
# Data scaling  
from sklearn.preprocessing import StandardScaler
```

```
sc=StandardScaler()
```

```
X_train = sc.fit_transform(x_train)  
X_test = sc.transform(x_test)
```

```
#LOGISTIC REGRESSION CLASIFIER (LRC)
```

```
from sklearn.linear_model import LogisticRegression  
logr = LogisticRegression(random_state=0)  
logr.fit(X_train,y_train)
```

```
y_pred_logr = logr.predict(X_test)
```

```
# K-NEAREST NEIGHBORS (KNN)
```

```

from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=1, metric='minkowski')
knn.fit(X_train,y_train)

y_pred_knn = knn.predict(X_test)

# SUPPORT VECTOR CLASSIFIER (SVC)
from sklearn.svm import SVC
svc = SVC(kernel='rbf')
svc.fit(X_train,y_train)

y_pred_svc = svc.predict(X_test)

# NAIVE BAYES (NB)
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train,y_train)

y_pred_gnb = gnb.predict(X_test)

# DECISION TREE CLASSIFIER (DTC)
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier(criterion='entropy')
dtc.fit(X_train,y_train)

y_pred_dtc = dtc.predict(X_test)

# RANDOM FOREST CLASSIFIER (RFC)
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=10, criterion='entropy')
rfc.fit(X_train,y_train)

y_pred_rfc = rfc.predict(X_test)

# CONFUSION MATRIX

from sklearn.metrics import confusion_matrix

print('Confusion matrix for LRC')
cm_logr=confusion_matrix(y_test, y_pred_logr)
print(cm_logr)
print('Confusion matrix for KNN')
cm_knn=confusion_matrix(y_test, y_pred_knn)
print(cm_knn)
print('Confusion matrix for SVC')
cm_svc=confusion_matrix(y_test, y_pred_svc)
print(cm_svc)
print('Confusion matrix for GNB')
cm_gnb= confusion_matrix(y_test, y_pred_gnb)
print(cm_gnb)

```

```

print('Confusion matrix for DTC')
cm_dtc=confusion_matrix(y_test, y_pred_dtc)
print(cm_dtc)
cm_rfc=confusion_matrix(y_test, y_pred_rfc)
print(cm_rfc)

# CLASSIFICATION REPORT
from sklearn.metrics import classification_report
target_names = ['TC%≤0%', '0%<TC%≤5%', '5%≤TC%<10%', '10%≤TC%']
print('Classification Report for RFC')
print(classification_report(y_test, y_pred_rfc, target_names=target_names))

# PREDICTIONS
print('Prediction for LRC')
print(logr.predict([[0.5,0.3]]))

#PARAMETER OPTIMIZATION (GRID SEARCH)
# for SVM
from sklearn.model_selection import GridSearchCV
p = [{'C':[1,2,3,4,5], 'kernel':['linear']},
      {'C':[1,2,3,4,5], 'kernel':['rbf'], 'gamma':[1,0.5,0.1,0.01,0.001]}]

gs=GridSearchCV(estimator=svc, #SVM algoritması
                 param_grid=p,
                 scoring='accuracy',
                 cv=10,
                 n_jobs=-1)

grid_search = gs.fit(X_train,y_train)
eniyisonuc = grid_search.best_score_
eniyiparametreler = grid_search.best_params_

print(eniyisonuc)
print(eniyiparametreler)

# for DTC
p_dtc = [{'criterion':['gini', 'entropy', 'log_loss'], 'splitter':['best','random']}]

gs=GridSearchCV(estimator=dtc,
                 param_grid=p_dtc,
                 scoring='accuracy',
                 cv=10,
                 n_jobs=-1)

grid_search = gs.fit(X_train,y_train)
bestscore_dtc = grid_search.best_score_
bestparams_dtc = grid_search.best_params_

print(bestscore_dtc)
print(bestparams_dtc)

```

CURRICULUM VITAE

Name Surname : Aynur Hürriyet TÜRKYILMAZ

EDUCATION :

- **B.Sc.** : 2010, Middle East Technical University, Faculty of Engineering, Department of Civil Engineering
- **M.Sc.** : 2012, Northeastern University, Faculty of Engineering, Department of Civil Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2010-2012 Research Assistant at Boston Northeastern University, MA.
- 2012-2014 Technical Office Engineer, TAV Construction, Izmir Adnan Menderes New Domestic Terminal Building and Multistorey Carpark Project, Izmir.
- 2014-2015 Technical Office Engineer, TAV-Sera Joint Venture, Emaar Square MP-02-00 N1 & N2 Seismic Zones Raft & Substructure Construction Works Extension, Istanbul.
- 2015-2017 Cost Control Specialist, TAV-Sera Libadiye Joint Venture, Emaar Square MP-03-00 Mall & Carpark and North Tower, Istanbul.
- 2017-2024 Cost Control Chief, TAV Construction Headquarters, Istanbul.
- 2024-Present Cost Control Manager, TAV Construction Headquarters, Istanbul.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Turkyilmaz, A. H., and Polat, G.** 2024: Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study, *Buildings*, 14(11), 3541.
- **Turkyilmaz, A. H., Polat, G., and Gurgun, A. P.** (2024). Risk Perception In The Construction Industry: A Bibliometric Review. The 2024 Latin American Structural Engineering and Construction Conference, March 4-8, 2024, Quito, Ecuador.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Turkyilmaz, A. H., and Polat, G.** (2025). Risk-based completion cost prediction approach in construction projects utilizing machine learning, *Journal of Information Technology in Construction (ITcon)*, 30(16), 375-396.

- **Turkyilmaz, A. H., Polat, G., and Gurkan, A.** (2024). Application of Construction 4.0 Technologies: Empirical Findings from the Turkish Construction Industry, *Journal of Information Technology in Construction (ITcon)*, 29(9), 179-197.
- **Turkyilmaz, A. H., Polat, G., and Gurkan, A.** (2024). Why Are Construction 4.0 Technologies Not Widely Used in The Construction Industry? In *Creative Construction Conference 2024 Proceedings*, June 29-July 2, 2024, Prague, Czechia.
- **Turkyilmaz, A. H., Polat, G., and Armangil Gokalp, A.F.** (2024). A Bibliometric Review of Disputes in The Construction Projects. In *Creative Construction Conference 2024 Proceedings*, June 29-July 2, 2024, Prague, Czechia.
- **Turkyilmaz, A. H., Polat, G., and Gurkan, A.** (2023). Implementation of Construction 4.0 Technologies in the Turkish Construction Industry. In *Creative Construction Conference 2023 Proceedings*, June 20-23, 2023, Budapest, Hungary.
- **Taş, E. F., Bulğan, J., Turkeyilmaz, A. H., Bozdemir, S., Akarcalı, G. Z., Akin, F. D., Şimşek, E., Kızılkaya, Ş.B., and Ergin, A.** (2020). An Assesment of Lean and Agile Principles in Building Life Cycle Phases. *6th International Project and Construction Management Conference (e-IPCMC2020)* Istanbul Technical University, 12-14 November 2020, Istanbul, Türkiye.