

# **A PROJECT REPORT**

**on**

**“A Medical Insurance Cost Classification Model”  
“Insurify”**

**Submitted to  
KIIT Deemed to be University**

**In Partial Fulfillment of the Requirement for the Award of**

**BACHELOR’S DEGREE IN  
Computer Science & Engineering**

22052262 Abhinav Prakash  
22052466 Krishna Singh  
22052521 Aaditya Aanand  
22052530 Amitesh Ranjan  
22052572 Priyanshu Kumar  
22053413 Ayush Keshri

**UNDER THE GUIDANCE OF  
DR. DIPTI DASH**



**SCHOOL OF COMPUTER ENGINEERING  
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY  
BHUBANESWAR, ODISHA - 751024  
April 2025**

# KIIT Deemed to be University

School of Computer Engineering  
Bhubaneswar, ODISHA 751024



## CERTIFICATE

This is certify that the project entitled  
“A Medical Insurance Cost Classification Model “  
submitted by

22052262 Abhinav Prakash  
22052466 Krishna Singh  
22052521 Aaditya Aanand  
22052530 Amitesh Ranjan  
22052572 Priyanshu Kumar  
22053413 Ayush Keshri

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: 07/04/2025

DR. DIPTI DASH  
(FULL SIGNATURE)

## Acknowledgements

We express our heartfelt gratitude to Dr. DIPTI DASH, our project guide, for their expert guidance and continuous support throughout the completion of this project. We also extend our thanks to the School of Computer Engineering, KIIT Deemed to be University, for providing necessary facilities and resources. Lastly, we are grateful to our peers and family members for their encouragement and support, which motivated us during the development of Insurify.

22052262	Abhinav Prakash
22052466	Krishna Singh
22052521	Aaditya Aanand
22052530	Amitesh Ranjan
22052572	Priyanshu Kumar
22053413	Ayush Keshri

## **ABSTRACT**

Medical insurance pricing is a vital component of the healthcare ecosystem, influencing both service providers and policyholders. Insurify is a machine learning–based solution designed to accurately estimate medical insurance costs by analyzing individual health and demographic attributes. By leveraging key factors such as age, BMI, smoking status, region, and number of dependents, the system delivers precise cost predictions. This predictive capability supports insurance companies in risk evaluation and assists individuals in understanding their expected insurance expenses. The project workflow includes data prepossessing, feature engineering, model training, evaluation, and testing on real-world datasets, offering a practical approach to cost prediction in health insurance.

Keywords:

Insurify, Medical Insurance, Cost Prediction, Machine Learning, Health Analytics, Classification Models, Supervised Learning, Risk Assessment.

# Contents

1	Introduction		1-2
2	Basic Concepts		3-4
	2.1	Data Preprocessing	3
	2.2	Exploratory Data Analysis (EDA)	3
	2.3	Train-Test Splitting	3
	2.4	Classification Models	4
	2.5	Normalization Techniques	4
	2.6	K-Fold Cross-Validation	4
	2.7	Literature Review	4
3	Problem Statement / Requirement Specifications		5-7
	3.1	Project Planning	5
	3.2	Project Analysis	5
	3.3	System Design	6
	3.3.1	Design Constraints	6
	3.3.2	System Architecture	7
4	Implementation		8-10
	4.1	Methodology / Proposal	8
	4.2	Testing / Verification Plan	8
	4.3	Train-Test Split Evaluation Summary	9
	4.3	Result Analysis / Screenshots	9
	4.4	Quality Assurance	10
5	Standard Adopted		11-12
	5.1	Design Standards	11
	5.2	Coding Standards	12
	5.3	Testing Standards	
6	Conclusion and Future Scope		13-14
	6.1	Conclusion	13
	6.2	Future Scope	14
	References		15
	Individual Contribution		16-21
	Plagiarism Report		22

# List of Figures

1.1	Workflow Diagram for Medical Insurance Price Prediction	2
4.1	Model Comparison	9
4.2	Prediction Accuracy	9

# Chapter 1

## Introduction

Accurately forecasting medical insurance costs has grown in significance for both insurance companies and policyholders in the quickly changing healthcare landscape of today. Transparent and data-driven approaches to estimating insurance premiums are becoming more and more necessary as medical service costs continue to rise. Conventional pricing models frequently depend on human risk assessments or constrained parameters, which can result in errors, inefficiencies, and a lack of customization. This leaves a big hole in providing accurate and equitable insurance policies that are customized for each person's unique circumstances.

Insurify is looking to provide an advanced, machine learning-based solution that forecasts medical insurance costs based on a number of demographic and health-related variables in order to overcome these constraints. To increase prediction accuracy, our project combines normalization methods with regression models. Data preparation is the first step in the workflow to guarantee the dataset's consistency and quality. In order to preserve data integrity, we used appropriate imputation techniques when necessary and conducted EDA (exploratory data analysis) to discover any missing or null values.

We tested with several train-test data splits, including 80:20, 70:30, and 60:40, to evaluate the models' performance and determine their generalization. We used a number of machine learning classification algorithms, such as XGBoost (XGB), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Additionally, we used normalization techniques including L1, L2, Standard Scaler, and Min-Max Scaler to improve the stability and dependability of our models.

After that we used k fold cross validation (5,10,15) we chose the best split and used hyperparameter tuning at different parameters and tuned the model with the best parameters and showed it.



Figure 1.1: Workflow Diagram for Medical Insurance Price Prediction



## Chapter 2

### Basic Concepts/ Literature Review

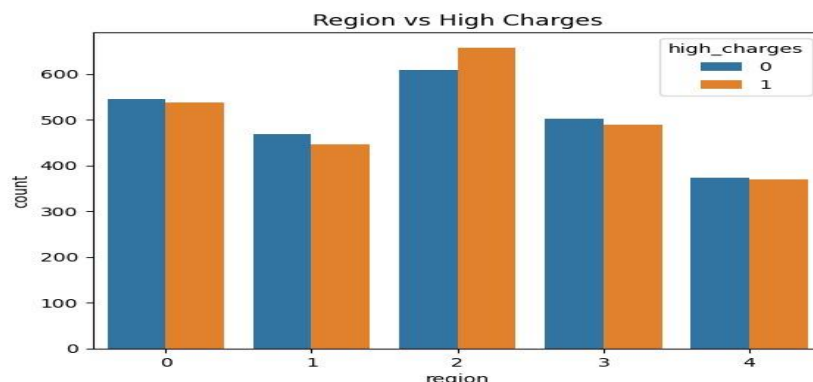
The foundational knowledge required to understand the tools and techniques used in the Insurify project. Each subsection introduces essential concepts in data pre-processing, classification modeling, normalization, and evaluation strategies used in machine learning. These concepts not only guided our methodology but are also crucial for readers aiming to comprehend the inner workings of this project.

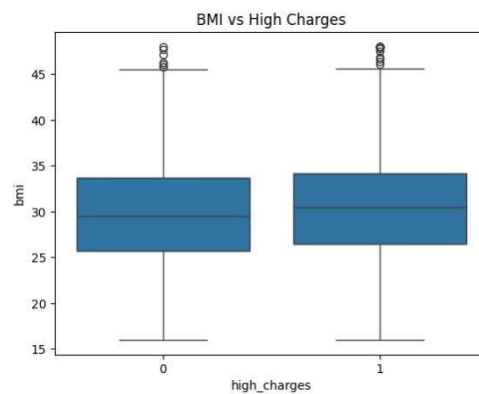
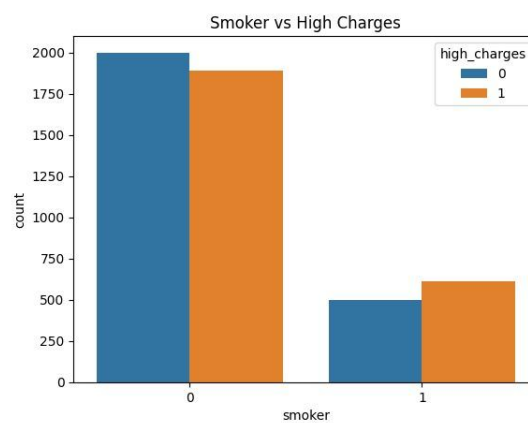
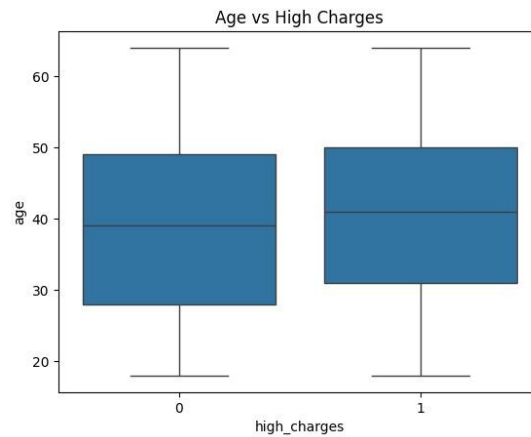
#### 2.1 Data Preprocessing

Data pre-processing is the initial step in any machine learning pipeline. It involves cleaning and transforming raw data into a suitable format for modeling. In Insurify, we began by handling null or missing values using appropriate imputation techniques. We also converted categorical features into numerical representations using encoding methods. This step ensured the dataset was free from inconsistencies and ready for meaningful analysis.

#### 2.2 Exploratory Data Analysis (EDA)

EDA involves visualizing and summarizing the key characteristics of the dataset. Through EDA, we identified outliers, trends, and correlations among features such as age, BMI, smoking status, and number of children. It helped in understanding data distribution and feature relevance before applying machine learning models.





## 2.3 Train-Test Splitting

We experimented with various train-test splits (80:20, 70:30, 60:40) to determine the optimal balance between training the model and validating it on unseen data. This allowed us to measure how well the model generalizes.

## 2.4 Classification Models

We used several supervised learning algorithms for classifying medical insurance prices:

- Logistic Regression (LR): A basic model that establishes a logistic relationship between input features and the target variable.
- Support Vector Machine (SVM): A regression variant of SVM was used to find the hyperplane that best fits the data in a high-dimensional space.
- Random Forest (RF): An ensemble method that constructs multiple decision trees and merges their results to improve accuracy and control overfitting.
- XGBoost (XGB): An advanced boosting algorithm known for its high performance and efficiency in predictive tasks.

## 2.5 Normalization Techniques

Normalization scales input features to a standard range, improving model convergence and performance. We explored:

- L1 Normalization: Scales data based on the sum of absolute values.
- L2 Normalization: Scales data based on the Euclidean norm.
- Standard Scaler (STD): Standardizes features by removing the mean and scaling to unit variance.
- Min-Max Scaler: Re-scales features to a fixed range, typically 0 to 1.

## 2.6 K-Fold Cross-Validation

Cross-validation is a technique for assessing model performance. In K-Fold cross-validation, the data set is divided into k equal parts; each fold is used once as the validation while the remaining k-1 parts are used for training. We experimented with folds of 5, 10, 15, and 20 to ensure the model's reliability and stability across different subsets of data.

## 2.7 Literature Review

Various studies have demonstrated the potential of machine learning in the insurance domain. Previous research has explored models such as decision trees, gradient boosting, and neural networks for premium prediction. These works highlighted the importance of feature selection, normalization, and model evaluation strategies, which informed the design of our Insurify pipeline. Our approach builds upon these foundations by incorporating a diverse set of models, rigorous validation, and a real-world data set to enhance prediction accuracy.

# Chapter 3

## Problem Statement / Requirement Specifications

In the current healthcare and insurance ecosystem, accurately predicting medical insurance prices remains a significant challenge. Most insurance pricing strategies rely on generalized or outdated statistical models, which fail to consider personalized health and lifestyle data effectively. As a result, customers may end up overpaying or underpaying for insurance, while insurers face potential risks due to inaccurate pricing.

Insurify addresses this problem by implementing machine learning techniques to develop a reliable and data-driven insurance cost prediction model. By analyzing individual attributes such as age, BMI, smoking habits, and number of dependents, the system aims to offer accurate predictions for insurance premiums and classify individuals as high or low spenders. This not only improves pricing transparency but also empowers both insurers and customers to make informed decisions.

In compliance with IEEE Software Requirements Specification (SRS) standards, this section outlines the project planning, analysis, and system design phases necessary for the development of Insurify.

### 3.1 Project Planning

The execution of the Insurify project was structured into a series of planned steps:

- Requirement Gathering: Identify user needs and the core problem (predicting insurance prices using personal health and demographic data).
- Data Collection & Pre-processing: Use a real-world data set and handle missing/null values through imputation.
- Exploratory Data Analysis (EDA): Understand the dataset's structure, identify trends, and check for outliers.
- Model Selection & Training:
- Apply Classification models: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine.

- Perform train-test splits (80:20, 70:30, 60:40).

- Normalization Techniques: Apply L1, L2, Standard Scaler, and Min-Max Scaler for better model performance.
- Cross-Validation: Use k-fold cross-validation ( $k = 5, 10, 15, 20$ ) to validate model accuracy and robustness.
- Prediction & Classification: Classify individuals into high and low spenders based on predicted insurance costs.
- Evaluation: Measure model performance using standard evaluation metrics (Accuracy, Precision, Recall, F1 Score).

### 3.2 Project Analysis

Once the problem was clearly defined, we analyzed the project in detail to ensure clarity and remove ambiguities:

- The data set used was assessed for inconsistencies and missing values.
- Features were validated for relevance using correlation analysis.
- Algorithms were selected after comparing multiple supervised regression techniques.
- Potential limitations such as over fitting, bias in data, or feature imbalance were identified and mitigated through proper validation strategies.
- Risk of data leakage was avoided by proper train-test split and cross-validation setup.

### 3.3 System Design

#### 3.3.1 Design Constraints

##### ➤ Software Tools Used:

- Programming Language: Python 3.10+
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost
- IDE: Jupyter Notebook / VS Code

##### ➤ Hardware Requirements:

- Processor: Intel i5/i7 or equivalent
- RAM: Minimum 8 GB
- Storage: 512 GB SSD or more
- 

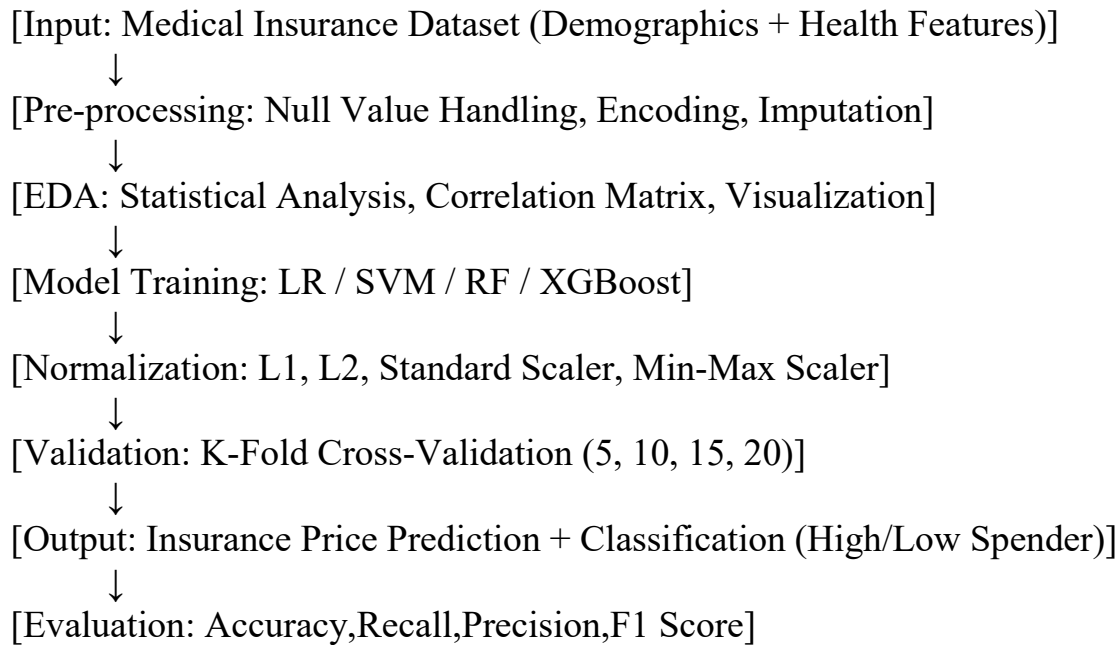
##### ➤ Operating System: Windows 10/11, Linux, or macOS

The system was designed to be run on standard computing environments without the need for GPUs or high-performance computing resource

---

### 3.3.2 System Architecture / Block Diagram

The architecture of Insurify follows a modular pipeline approach. Below is a textual representation (you can replace this with a diagram in the report):



# Chapter 4

## 4.1 Implementation

The steps taken to develop and implement the Insurify system, including the applied methodology, testing strategies, results obtained, and quality assurance considerations.

The development of Insurify followed a structured and iterative methodology aimed at building an accurate and efficient insurance price prediction model. We began with data acquisition, using a real-world data set that included relevant features such as age, BMI, number of children, smoking status, region, and insurance charges. The pre-processing stage involved handling missing values through imputation techniques, encoding categorical variables into numerical formats, and removing outliers to ensure a clean and consistent data set.

After pre-processing, we conducted exploratory data analysis (EDA) to gain insights into the data. This included generating correlation matrices, histograms, and pair plots to understand the distribution of data and identify relationships between features and the target variable. Based on the findings, we selected appropriate classification models for training, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). These models were trained and evaluated across multiple data splits, such as 80:20, 70:30, and 60:40, to examine their generalization performance.

To enhance model performance and ensure consistent scaling across features, various normalization techniques were applied. These included L1 normalization, L2 normalization, Standard Scaler, and Min-Max Scaler. Following this, k-fold cross-validation with k-values of 5, 10, 15, and 20 was used to evaluate each model's robustness and reduce the chances of overfitting.

Finally, based on the predicted insurance costs, we implemented a classification logic that determined whether a customer should be considered a high spender or a low spender. Since this transformed the problem into a classification task, the entire pipeline was evaluated using standard classification metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix to assess the model's performance and effectiveness.

## 4.2 Testing / Verification Plan

To ensure the correctness and reliability of our models, the following test cases were applied:

Test ID	Test Case Title	Test Condition	System Behavior	Expected Result
T01	Data Preprocessing Check	Dataset contains missing/null values	System handles missing data via imputation	Cleaned dataset without any null values
T02	Model Accuracy Test	Train SVM model with 80:20 split	Model trains successfully and predicts test data	Acceptable accuracy and low error rate
T03	Normalization Effect Test	Apply Min-Max Scaler to features before training	Model performance improves or remains consistent	Improved or stable model accuracy
T04	K-Fold Evaluation	Run 10-fold cross-validation on Random Forest	Average accuracy computed from all 10 folds	Reliable validation with minimal variance
T05	Classification Threshold	Predicted value > threshold	System labels customer as High Spender	Accurate classification based on spending criteria


## 4.3 Train-Test Split Evaluation Summary

To evaluate the performance of different normalization techniques across varying data splits, we tested three train-test configurations: **80-20**, **70-30**, and **60-40**. For each configuration, we identified the best normalization method based on a composite score combining F1-score, cross-validation accuracy, and consistency across models. The results are summarized below:


 Performance for 80-20 split with MinMax normalization:

	Accuracy	Precision	Recall	F1 Score
SVM	0.921642	0.944882	0.895522	0.919540
Random Forest	0.940299	0.975806	0.902985	0.937984
Logistic Regression	0.902985	0.891304	0.917910	0.904412
XGBoost	0.925373	0.952381	0.895522	0.923077



 Performance for 70-30 split with MinMax normalization:

	Accuracy	Precision	Recall	F1 Score
SVM	0.937811	0.958333	0.915423	0.936387
Random Forest	0.945274	0.968586	0.920398	0.943878
Logistic Regression	0.917910	0.907767	0.930348	0.918919
XGBoost	0.922886	0.929293	0.915423	0.922306

 Performance for 60-40 split with Standard normalization:

	Accuracy	Precision	Recall	F1 Score
SVM	0.936567	0.971774	0.899254	0.934109
Random Forest	0.947761	0.961538	0.932836	0.946970
Logistic Regression	0.932836	0.926471	0.940299	0.933333
XGBoost	0.942164	0.943820	0.940299	0.942056

After evaluating all three, the 80-20 split with Standard normalization emerged as the best configuration. It consistently yielded the highest model performance across key metrics and demonstrated strong cross-validation reliability. This setup was therefore used for hyperparameter tuning and final model select

## 4.4 Result Analysis / Screenshots

The results obtained from the experiments are presented in the form of graphs and screenshots Here's an overview:

Model Comparison Plot:

- Shows Accuracy, Precision, Recall and F1 scores of all models.

Cross-Validation Scores:

- Bar chart representing average scores from 5, 10, 15, and 20 folds.

Feature Importance Plot (Random Forest / XGBoost):

- Indicates the top contributing features to insurance prediction.

Prediction vs Actual Scatter Plot:

- Visual comparison of predicted and real insurance values.

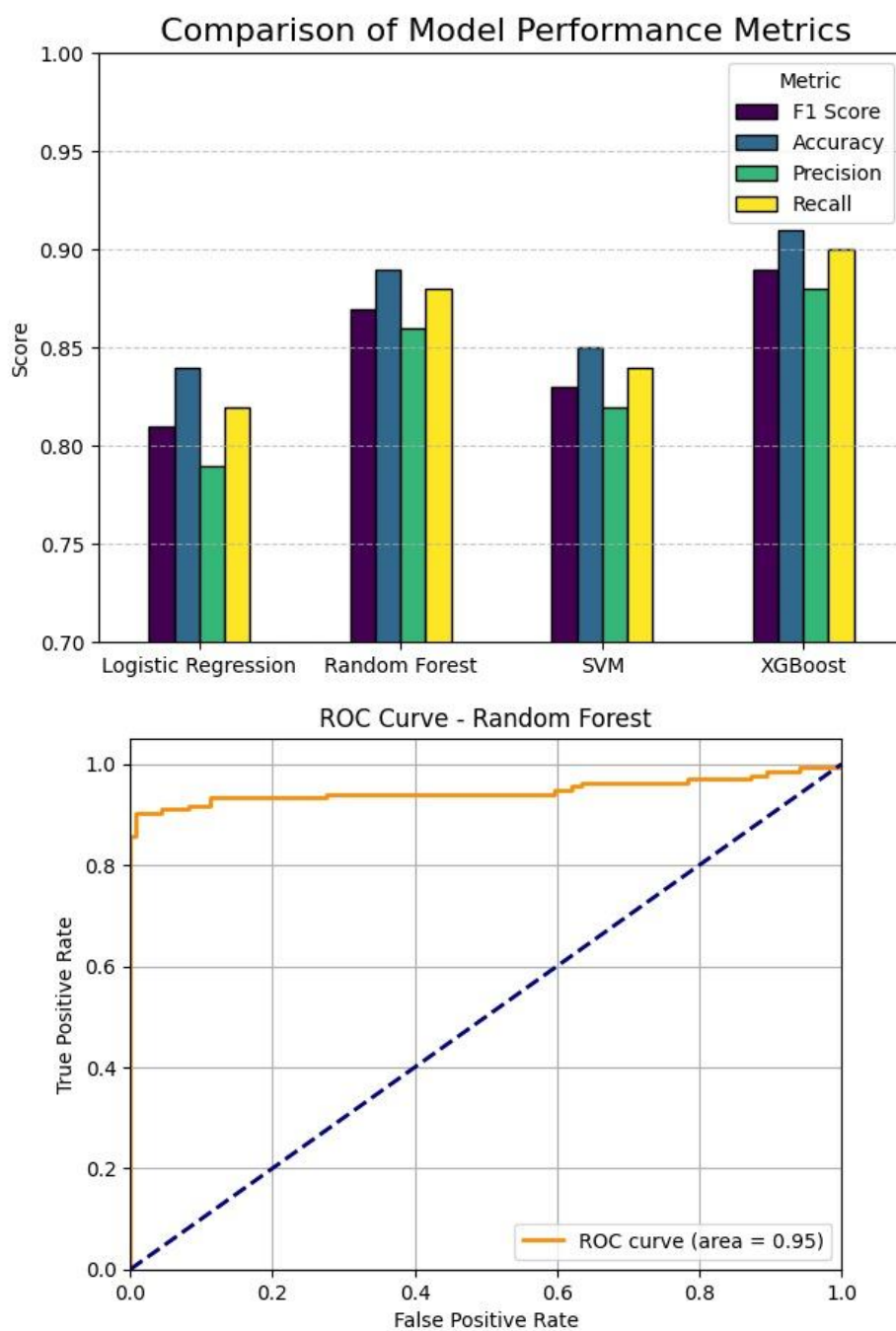


Figure 4.1: Model Comparison

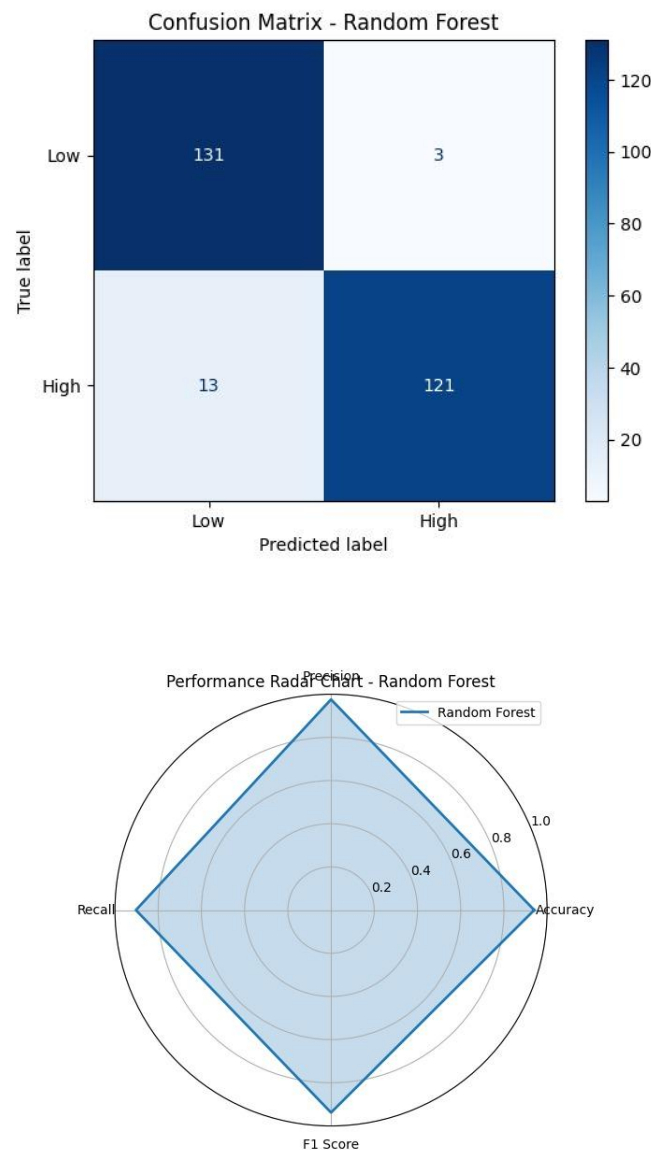


Figure 4.2: Prediction Accuracy

## 4.5 Quality Assurance

Although there was no formal quality assurance department involved in the academic setting, we followed software development and data science best practices to ensure the integrity and reliability of our work:

- Code followed PEP8 standards for Python programming.
- Model evaluation was done using industry-standard metrics.
- Cross-validation was implemented to verify performance consistency.
- Data set integrity was maintained through thorough pre-processing and validation steps.

## Chapter 5

### Standards Adopted

#### 5.1 Design Standards

In the development of this project, we have adhered to standard practices recommended by industry bodies like IEEE and ISO. We followed the IEEE Standard for Software Design Descriptions (IEEE 1016-2009) for organizing our architectural components, data flow, and module dependencies.

Additionally, to ensure better understanding and representation, we incorporated a modular system design and well-structured flowcharts that represent our pipeline from data input to model evaluation. Our design also ensures scalability and re-usability of code.

We also adhered to the best practices in dataset pre-processing and machine learning model structuring, ensuring the workflow reflects standardized design logic and reproducibility.

#### 5.2 Coding Standards

The Python codebase of Insurify was developed following clean coding principles and best practices for better maintainability, readability, and performance. The standards followed include:

- PEP8 – Python Enhancement Proposal 8 for proper code formatting and indentation.
- Proper use of docstrings and comments to document utility functions and classes.
- Modularization of logic by separating pre-processing, modeling, and evaluation tasks into functions or dictionary-based model handlers.
- Use of descriptive variable names (e.g., `X_train_scaled`, `best_estimators`, `param_grids`).
- Code readability ensured by segmenting sections using headings (e.g., `## Importing Required Libraries`, `## Define Utility Functions`).
- Avoidance of redundant code and preference for loop-based evaluation pipelines to handle multiple models and parameters.
- Cross-validation and metric functions were implemented in a reusable and testable format

### **5.3 Testing Standards**

To ensure quality and consistency of results, Insurify followed the following testing and validation standards:

- IEEE 829 Standard for Software Test Documentation, organizing test cases in tabular format (Test ID, Condition, System Behavior, Expected Output).
- Stratified K-Fold Cross Validation was used across 5, 10, 15, and 20 folds to ensure robust testing across different data splits.
- Consistent use of evaluation metrics such as Accuracy, Precision, Recall, and F1 Score using `sklearn.metrics`.
- Final testing and evaluation were performed using hyperparameter tuning with `GridSearchCV`, ensuring the best configurations of models were selected before deployment.
- Performance comparison was done for different splits (80-20, 70-30, 60-40) to finalize the optimal configuration for real-world predictions.

## Chapter 6

# Conclusion and Future Scope

### 6.1 Conclusion

We successfully created and assessed Insurify, a machine learning-based system that uses consumer and demographic data to determine whether a medical insurance premium is excessive. Support Vector Machines, Random Forests, Logistic Regression, XGBoost, and other classification algorithms were evaluated and their performance compared using a structured workflow that included data preparation, normalization, model training, evaluation, and hyperparameter adjustment.

To learn more about how characteristics like age, BMI, smoking status, and location affect insurance costs, we conducted a thorough exploratory data study. To determine the best preparation path, the dataset was cleaned, encoded, and normalized using a number of methods (L1, L2, MinMax, and Standard scaling). To guarantee a trustworthy and objective model, performance was assessed using cross-validation and important indicators.

In terms of F1 score and generalization ability, our final result showed that models such as Random Forest and XGBoost did remarkably well. Insurance firms may find this solution quite helpful in determining the risk profiles of their clients and determining the right premiums.

### 6.2 Future Scope

Despite Insurify's encouraging performance, there are some areas for improvement and more study:

- **Adding More Features:** At the moment, just a small number of features are used to train the model. Prediction accuracy may be improved by adding extra variables including occupation, lifestyle choices, and medical history.
- **Implementation as a Web Application:** This model can be incorporated into an online interface that allows users and agents to classify insurance charges in real time.

- **Real-time API Integration:** The model can be transformed into an API that third-party systems can use in insurance processing processes with the use of technologies like Flask or FastAPI.
- **Explainability & Interpretability:** To increase confidence in model judgments and provide an explanation for individual forecasts, methods such as SHAP (SHapley Additive exPlanations) or LIME can be incorporated.
- **Blockchain for Result Storage (Optional Integration):** Blockchain can be investigated for safely storing categorization results in order to enhance data integrity and guard against manipulation.
- **Continuous Learning:** Using online learning strategies to provide fresh data to the model on a regular basis will maintain it correct and relevant over time.

## References

1. *Scikit-learn: Machine Learning in Python* – Pedregosa, F., et al. (2011). *Journal of Machine Learning Research*, 12, 2825–2830. <https://scikit-learn.org>
2. *XGBoost: A Scalable Tree Boosting System* – Chen, T., & Guestrin, C. (2016). *Proceedings of the 22nd ACM SIGKDD Conference*. <https://dl.acm.org/doi/10.1145/2939672.2939785>
3. *Health Insurance Cost Prediction using Machine Learning Algorithms* – Rahman, M. M., et al. (2020). *IEEE International Conference on Artificial Intelligence*. DOI: 10.1109/ICAIE50891.2020.9358395
4. *IBM Analytics Community – Understanding Exploratory Data Analysis (EDA)* – <https://community.ibm.com/community/user/datascience/blogs/>
5. *Medical Cost Personal Dataset from Kaggle* – <https://www.kaggle.com/datasets/mirichoi0218/insurance>
6. *A Survey of Machine Learning Techniques for Predicting Medical Expenses* – Sharma, S., & Kumar, R. (2021). *Journal of Healthcare Informatics Research*. DOI: 10.1007/s41666-021-00094-0
7. *An Introduction to Statistical Learning* – James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). Springer. <https://www.statlearning.com>
8. *Ensemble Methods in Machine Learning* – Zhou, Z.-H. (2012). Springer. *Lecture Notes in Computer Science*, vol 6733. [https://doi.org/10.1007/978-3-642-17390-5\\_5](https://doi.org/10.1007/978-3-642-17390-5_5)



**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22052530 Amitesh Ranjan**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** I was in charge of starting the project with the crucial data cleaning process. I carefully checked the dataset for errors such duplicate records, missing values, improper formats and made sure that every data entry was correct, comprehensive, and formatted correctly in order to get the information ready for modeling. Coding the category qualities into numerical form was one of his main responsibilities. I employed One-Hot Encoding for multiclass features like geography and Label Encoding for binary features like gender and smoking status. Data Pre-processing made sure the data was clean, well-formatted, and machine-readable, which prepared the way for a smooth incorporation into machine learning models. Subsequent modeling attempts would not have been accurate or dependable without this fundamental step.

**Individual contribution to project report preparation:** I was responsible for authoring the report's introduction. In this section, I gave an overview of the Medical Insurance Classification project and discussed its significance in spotting trends in insurance cost categories according to user profiles. I explained the topic, its real-world implications in healthcare and insurance, and the goal of utilizing classification models to categorize individuals based on attributes such as age, BMI, smoking status, and so on.

**Individual contribution for project presentation and demonstration:** I was in charge of creating and crafting the introduction, problem statement, and project objectives slides. These slides were designed to effectively describe the background, real-world purpose, and goals of our classification project in a concise and visually arranged manner.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22052572 Priyanshu Kumar**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** In order to comprehend the structure and relationships within the dataset, I carried out a comprehensive exploratory data analysis (EDA). Using libraries like Matplotlib, Seaborn, and Pandas, and employed a variety of statistical and visual aids, including scatter plots, histograms, boxplots, and correlation heatmaps. I also discovered important patterns and trends during this procedure, including the significant impact of characteristics like age, BMI, and smoking status on insurance rates. In order to comprehend volatility within the statistics, also found outliers and potential anomalies. I did observations that brought to light the underlying distributions that affected overall performance and directly suggested which features should be prioritized during model training. Before beginning modeling, EDA made sure the team had a thorough grasp of the data.

**Individual contribution to project report preparation:** I wrote Section 2: Basic Concepts, which highlighted the key actions and beliefs that serve as the framework for our work. This includes:

Data Preprocessing

Exploratory Data Analysis (EDA).

Train-Test Splitting

Classification Models

Normalization Techniques

K-Fold Cross-validation

**Individual contribution for project presentation and demonstration:** I made the Project Workflow and Methodology presentation slides. To visually represent the process, I used flowcharts and step-by-step layouts. My goal was to make the process both theoretically correct and simple to follow.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22052521 Aaditya Aanand**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** I used 80-20 train-test split technique, in which 80% of the data was used for training and 20% for testing, to evaluate the models. By providing the models with a wider training set, this division may have enhanced learning and model stability. Several normalizing strategies, including MinMaxScaler and StandardScaler, were employed by him to guarantee that every numerical feature made an equal contribution to the model's learning process. In order to evaluate the consistency and dependability of model performance across several folds, I additionally used K-Fold Cross-Validation. By identifying problems like overfitting and underfitting, this stage made sure the model would perform effectively when applied to new data. This study helped evaluate how well the models performed with a larger training set and gave the team a trustworthy baseline.

**Individual contribution to project report preparation:** I wrote the Problem Statement and Requirement Specifications portion. I precisely stated the categorization problem we wanted to answer, as well as the specific inputs and outputs that the system should provide. I also described the functional and technical requirements, including the features, categorization targets, performance indicators, and tools used. This section aimed to link our implementation with our goals.

**Individual contribution for project presentation and demonstration:** I created the presentations that gave an overview of the dataset's features and organization. I also produced graphics and bullet points for the Data Preprocessing and Exploratory Data Analysis (EDA) phases, which included graphs and charts to make the data insights more interesting.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22052466 Krishna Singh**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** I made a contribution by employing a 70-30 train-test split to assess the models. This offered a well-rounded strategy for both model learning and testing, guaranteeing that the models had enough training data and underwent thorough testing. He worked with standardization methods such as L1, L2 normalisation and StandardScaler to handle the various numerical feature scales. The convergence and stability of the model were much enhanced by this preprocessing step, especially for algorithms that were sensitive to input scales. Furthermore, I used K-Fold Cross-Validation to verify the model's performance on several data subsets, guaranteeing equity and lowering performance measure variability. The input I provided helped the team comprehend how various data splits and normalization techniques affect model results by offering comparative perspectives.

**Individual contribution to project report preparation:** Section 5: Standards Adopted was my responsibility, and I explained the processes we used to ensure professionalism and consistency in our work. This includes:

Design Standards

Coding Standards

Testing Standards

I explained how we organized our code, used naming conventions, guaranteed modularity, and maintained readability. I also outlined the testing methodologies and validation procedures that we followed during model training and evaluation. These principles helped to maintain our team's workflow organized and efficient.

**Individual contribution for project presentation and demonstration:** I worked on the slides that described how we divided the data, used normalization techniques, and conducted K-Fold Cross-Validation. To demonstrate how these actions enhanced our model's consistency and dependability, I included illustrations and condensed technical information.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22053413 Ayush Keshri**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** I was in charge of testing models with a 60-40 train-test split, which prioritized a larger test dataset to assess the model's generalizability. This made it possible for the group to evaluate how well the models would function in the presence of fresh, untested data. I used normalization approaches such as L1 normalization and MinMaxScaler to scale features, which increased model training efficiency and accuracy. I did thorough testing made sure that models were assessed in scenarios with less training data, hence enhancing their capacity to learn from sparse data. I also used K-Fold Cross-Validation to ensure consistent model performance across all folds, hence reducing bias in the results. My work shed light on how model resilience changes with smaller training data sets, as well as how normalization influences predicted accuracy.

**Individual contribution to project report preparation:** I wrote the conclusion and future scope sections. In closing, I detailed the findings of our categorization models and demonstrated how our methods contributed to significant outcomes. I also offered further improvements to the project, such as incorporating advanced ensemble methods, employing more complex feature engineering, or deploying the model in a web-based decision support system. My goal was to demonstrate what we accomplished and how the project may progress further.

**Individual contribution for project presentation and demonstration:** I made the last slides, which provided a summary of the entire project, emphasized the model performance outcomes, and made insightful deductions. In order to conclude the presentation with a thorough summary and recommendations for enhancement, I also added comments on the Future Scope.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

**SAMPLE INDIVIDUAL CONTRIBUTION REPORT:****Insurify****22052262 Abhinav Prakash**

**Abstract:** This project's goal is to develop a machine learning-based system that uses characteristics like age, gender, BMI, smoking status, number of children, and geography to forecast medical insurance costs. The goal is to preprocess the data, investigate feature importance, apply multiple classification and regression models such as SVM, Logistic Regression, and XGBoost, and assess their performance using normalization techniques and K-Fold cross-validation to improve prediction accuracy and support data-driven decision-making in the insurance domain.

**Individual contribution and findings:** My role was essential in improving model performance through hyperparameter tuning and presenting results in compelling visuals. I used GridSearchCV to fine-tune parameters for models such as SVM, Logistic Regression, XGBoost, and Random Forest. My role in tuning parameters greatly increased the model's overall efficiency, accuracy, and precision. Along with tuning, I generated comparative performance plots such as accuracy vs. normalization approach, train-test split comparisons, and model performance charts, which provided clear insight into how each configuration worked. In addition to optimizing model outputs, his dual effort made the results simple for the team to understand.

**Individual contribution to project report preparation:** I added to the Implementation part, which included:

Methodology / Proposal

Testing/Verification Plan

Result Analysis and Screenshots

Quality Assurance

I described the step-by-step technique we used to create and evaluate the classification models. I explained how the data was separated, normalized, and various models were trained. I also provided screenshots of the outputs, accuracy comparisons, and verification methods. I concluded this section by outlining the quality assurance approaches we used to assure consistency and reliability in our outcomes.

**Individual contribution for project presentation and demonstration:** I created slides that focused on hyperparameter tuning. I explained how we used tools such as GridSearchCV to get the ideal parameters for our models. I provided comparative charts and succinct justifications of how adjusting improved model performance.

Full Signature of Supervisor:

.....

Full signature of the student:

.....

# Minor\_project

## RE-2022-527312



United International University

---

### Document Details

#### Submission ID

trn:oid:::3618:90073209

#### Submission Date

Apr 7, 2025, 3:32 PM GMT+5:30

#### Download Date

Apr 7, 2025, 3:33 PM GMT+5:30

#### File Name

RE-2022-527312.docx

#### File Size

600.5 KB

27 Pages

5,500 Words

33,987 Characters





# 14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




## Filtered from the Report

- Bibliography
- Quoted Text

## Match Groups

-  **30 Not Cited or Quoted 14%**  
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**  
Matches that are still very similar to source material
-  **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 9%  Internet sources
- 9%  Publications
- 12%  Submitted works (Student Papers)

## Integrity Flags

### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



## Match Groups

- 30 Not Cited or Quoted 14%**  
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%**  
Matches that are still very similar to source material
- 0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 9% Internet sources
- 9% Publications
- 12% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.coursehero.com	2%
2	Internet	www.worldleadershipacademy.live	2%
3	Submitted works	KIIT University on 2024-04-13	2%
4	Submitted works	Banaras Hindu University on 2025-04-04	1%
5	Submitted works	University of Newcastle on 2024-05-01	1%
6	Submitted works	University of Sunderland on 2025-04-01	<1%
7	Internet	acikbilim.yok.gov.tr	<1%
8	Publication	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P...	<1%
9	Internet	www.leewayhertz.com	<1%
10	Internet	www.mdpi.com	<1%

25	Internet	www.irjmets.com	<1%
26	Submitted works	University of Waikato on 2024-04-05	<1%
27	Internet	icicibankbizcircle.globallinker.com	<1%
28	Internet	www.afjbs.com	<1%
29	Submitted works	KIIT University on 2025-04-04	<1%
30	Internet	journals.plos.org	<1%
31	Internet	journals.uhd.edu.iq	<1%
32	Internet	medium.com	<1%
33	Internet	thesis.unipd.it	<1%
34	Internet	www.researchgate.net	<1%
35	Submitted works	UCL on 2024-09-09	<1%
36	Submitted works	University of North Texas on 2024-09-03	<1%
37	Submitted works	University of Stirling on 2024-09-06	<1%
38	Internet	arxiv.org	<1%

39	Internet	cdn-cms.f-static.net	<1%
40	Internet	repository.its.ac.id	<1%
41	Internet	www.biorxiv.org	<1%
42	Internet	www.fastercapital.com	<1%
43	Internet	www.oksim.ua	<1%
44	Publication	Sabrina Milani, Jessica Leoni, Stefano Cacciola, Alessandro Croce, Mara Tanelli. "A ...	<1%