**PREDICTIVE LOAN APPROVAL SYSTEM USING RANDOM FOREST**

**Project Report**

Submitted to the Faculty of Engineering of

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA**

In partial fulfillment of the requirements for the award of the Degree of

## BACHELOR OF TECHNOLOGY

## In

## CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

By

**MANEPALLI SARANYA DEEPYA**

Under the Enviable and Esteemed Guidance of

# Dr. Y. ADILAKSHMI, M. Tech, Ph.D

Professor & Head of the Department,

CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)



**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**

## SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

**(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada) SESHADRI RAO KNOWLEDGE VILLAGE GUDLAVALLERU 521356**

**ANDHRA PRADESH**

**2024-2025**

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

**(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)**

**SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU**

**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**



**CERTIFICATE**

This is to certify that the project report entitled **“Predictive Loan Approval System”** is a bonafide record of work carried out by Manepalli Saranya Deepya (22481A4265) under the guidance and supervision of **Dr. Y. ADILAKSHMI, Professor and Head of the Department,** CSE (Artificial Intelligence & Machine Learning), in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

**Project Guide Head of the Department**

**(Dr. Y. ADILAKSHMI) (Dr. Y. ADILAKSHMI)**

**ACKNOWLEDGEMENT**

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr. Y. ADILAKSHMI, Professor & Head of the Department,** CSE (Artificial Intelligence & Machine Learning) for her constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. Y. ADILAKSHMI**, **Head of the Department**, CSE (Artificial Intelligence & Machine Learning) for her encouragements all the way during analysis of the project. Her annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B. KARUNA KUMAR** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project intime.

BY

MANEPALLI SARANYA DEEPYA

## INDEX

**TITLE PAGE NO**

ABSTRACT 1

CHAPTER 1: INTRODUCTION 2

* 1. [Introduction](#_TOC_250000)
  2. Problem definition 3

CHAPTER 2: PROPOSED METHOD 4

* 1. Methodology
     1. Block Diagram
     2. Algorithm and Explanation 5
  2. Data Preparation 8
     1. Data Description
     2. Data Pre-processing

CHAPTER 3: RESULTS 11

CHAPTER 4: CONCLUSION AND FUTURE SCOPE 14

References

List of Program Outcomes and Program Specific Outcomes Mapping of Program Outcomes with graduated POs and PSOs

## ABSTRACT

The Predictive Loan Approval System is an AI-powered solution designed to evaluate loan applications efficiently and accurately. It leverages historical loan data to train machine learning models that analyze key applicant attributes such as income, credit score, employment status, and loan amount. These models predict the likelihood of loan approval, enabling financial institutions to make faster and more consistent decisions. By automating the approval process, the system reduces human error and potential bias in decision-making. It also offers real-time feedback to applicants, enhancing transparency and user experience. The system continuously learns and adapts as new data is collected, improving prediction accuracy over time. Risky applications are flagged for further manual review, ensuring a balanced approach between automation and oversight. Commonly used algorithms include decision trees, logistic regression, and ensemble methods.

The solution supports both supervised and unsupervised learning techniques for robust analysis. It features a user-friendly interface for both applicants and loan officers, and integrates with existing databases to ensure real-time data processing. By streamlining operations.It enhances efficiency and reduces processing time. It also helps financial institutions optimize resource allocation and increase customer satisfaction. Overall, the Predictive Loan Approval System transforms the traditional loan processing workflow into a smart, data-driven approach.

**Keywords:** AI-powered, machine learning, data-driven, loan evaluation, historical loan data, applicant attributes, income, credit score, employment status, loan amount, loan approval, supervised learning, unsupervised learning, decision trees, logistic regression, ensemble methods, automation, real-time feedback, bias reduction, error reduction, risk flagging, manual review, user-friendly interface, applicants, loan officers, real-time data, efficiency, resource optimization, customer satisfaction, adaptive workflow

## CHAPTER 1

## INTRODUCTION

* 1. **Introduction**

In the modern financial ecosystem, efficient and accurate loan processing has become one of the key areas of focus for banks and financial institutions. With an increasing number of people seeking financial support for personal, educational, and entrepreneurial needs, the volume of loan applications continues to grow at a rapid pace. However, traditional loan approval mechanisms are often manually handled by loan officers who rely on fixed rules, documentation, and human judgment. This conventional approach is not only time-consuming but also susceptible to inconsistency, human bias, and processing delays, which can negatively impact both applicants and financial organizations. In light of these challenges, the integration of artificial intelligence, particularly machine learning, presents a transformative solution to automate and optimize loan approval workflows.

Machine learning (ML) is a subfield of artificial intelligence that empowers systems to learn from data patterns and make predictions or decisions without being explicitly programmed. By training models on historical data, ML systems can identify complex relationships and trends that might be overlooked by human analysts. In the context of loan processing, machine learning can be employed to predict whether a loan application should be approved or rejected, based on patterns observed in previous loan records. This includes analyzing attributes such as applicant income, loan amount, credit history, employment details, education level, and property area. The goal is to build a reliable and fair predictive model that mimics the decision-making ability of experienced financial experts—while reducing the time, cost, and risk involved in manual evaluations.

**1.2. Problem definition**

## The proposed Predictive Loan Approval System aims to automate and streamline the loan approval process by leveraging machine learning techniques on historical applicant data. Traditional loan sanctioning procedures are often manual, time-consuming, and prone to human error or bias. To address these limitations, this system is designed to predict whether a loan application should be approved or rejected based on various features such as applicant income, loan amount, credit history, employment type, education, and marital status. The system will undergo data preprocessing steps including handling missing values, encoding categorical data, and scaling numerical features to ensure model accuracy.

## A range of machine learning models—such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—will be trained and evaluated to identify the most effective one based on accuracy and reliability. The model with the best performance will be integrated into a user-friendly interface where new applicant data can be submitted for instant prediction. This intelligent solution is expected to not only reduce processing time and operational costs but also improve consistency and transparency in the decision-making process within financial institutions.

## Top of Form

## Bottom of Form

**CHAPTER 2**

**PROPOSED METHOD**

**2.1. Methodology**

The proposed Predictive Loan Approval System aims to automate and streamline the loan approval process by leveraging machine learning techniques on historical applicant data. Traditional loan sanctioning procedures are often manual, time-consuming, and prone to human error or bias. To address these limitations, this system is designed to predict whether a loan application should be approved or rejected based on various features such as applicant income, loan amount, credit history, employment type, education, and marital status. The system will undergo data preprocessing steps including handling missing values, encoding categorical data, and scaling numerical features to ensure model accuracy. A range of machine learning models—such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—will be trained and evaluated to identify the most effective one based on accuracy and reliability. The model with the best performance will be integrated into a user-friendly interface where new applicant data can be submitted for instant prediction. This intelligent solution is expected to not only reduce processing time and operational costs but also improve consistency and transparency in the decision-making process within financial institutions.

* + 1. **Block Diagram**

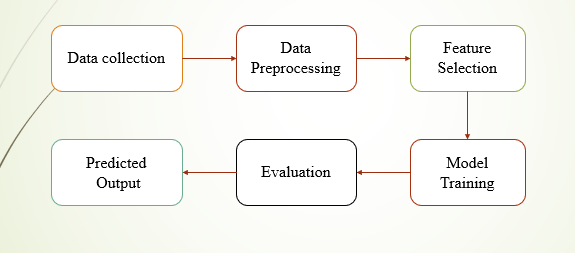


Fig:2.1 Block Diagram

The block diagram of the Predictive Loan Approval System provides a simplified overview of the workflow involved in automating the loan prediction process using machine learning. It comprises several interconnected modules, each responsible for a specific task, forming an end-to-end pipeline from data input to prediction output. The system starts with the Data collection, where loan applicants or bank officers provide details such as applicant income, loan amount, credit history, education, employment type, and property location. These inputs are then passed to the Data Preprocessing Module, which is responsible for handling missing values, encoding categorical variables, and normalizing numerical features. This step ensures that the raw input data is transformed into a suitable format for machine learning models.

* + 1. **Algorithm and Explanation**

The Predictive Loan Approval System is divided into four primary phases: Data Manager, Feature Extraction, Model Training, and Approval Prediction. Each phase plays a critical role in ensuring the system is accurate, scalable, and efficient.

**2.1.2.1 Data Manager**

In this phase, the algorithm processes a cleaned dataset that contains both customer demographics and financial features relevant to loan approval. The Data Manager is responsible for:

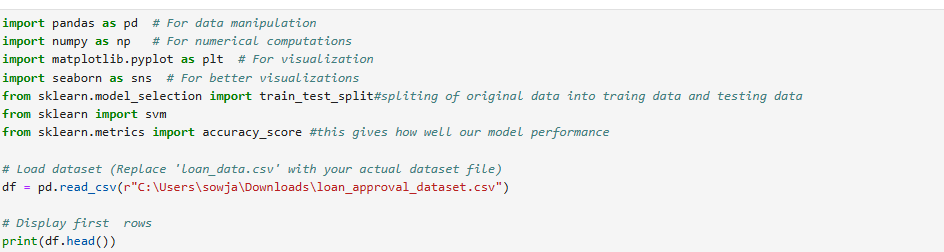


Fig 2.2 : Code for importing libraries

Loading the dataset using the pandas library. The dataset contains attributes like age, income, credit score, loan amount, employment status, and loan repayment history.

Removing records with missing target labels to ensure the reliability of the supervised learning model.

Identifying and separating features and target labels, where features include variables such as age, income, credit score, employment status, and loan amount, and the target label represents loan approval status (approved or denied).

Encoding categorical features, such as employment status or loan purpose, using Label Encoder to convert them into numeric form, making them compatible with machine learning algorithms.

Handling missing values using the Simple Imputer with the most frequent strategy to impute missing feature values, ensuring the dataset remains usable for model training.

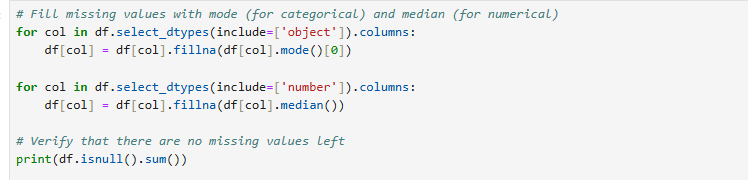


Fig:2.3 Filling Missing Values

**2.1.2.2 Feature Extraction**

In the Feature Extraction phase, significant features are selected and processed. The steps include:

Selecting relevant features such as age, income, credit score, loan amount, and employment status.

Encoding categorical variables (e.g., loan status and loan purpose) into numerical form for machine learning compatibility.

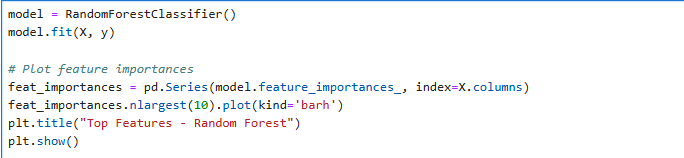


Fig 2.4: Feature Selection

**2.1.2.3 Model Training**

The Model Training phase involves training a machine learning model to predict loan approval based on customer data. The steps include:

Using the cleaned dataset prepared by the Data Manager phase.

Choosing a machine learning model, such as Logistic Regression, K-Nearest Neighbors (KNN), or Random Forest, for classification. For this example, a Logistic Regression model could be used.

Splitting the dataset into training and testing sets (usually an 80-20 split).

Training the model using the training data, where the features (age, income, credit score, etc.) predict the target variable (loan approval).

Evaluating model performance using metrics such as:

Accuracy Score: The proportion of correct predictions.

Precision, Recall, F1-Score: Measures for assessing classification performance.

Confusion Matrix: To visualize the model's performance on different classes (approved/rejected).

**2.1.2.4 Approval Prediction**

The Approval Prediction phase involves predicting the loan approval status for new applicants. This includes:

Loading the pre-trained model, along with its associated scaler and label encoders.

Preprocessing new applicant data, applying the same encoding, imputation, and scaling transformations used during training.

Making predictions using the pre-trained model to predict whether a loan will be approved or denied based on new applicant data.

Mapping numeric predictions back to their original categorical labels (approved or denied) using the saved label encoder.

This phase allows real-time predictions, enabling the system to automatically determine the loan approval status for new applicants.

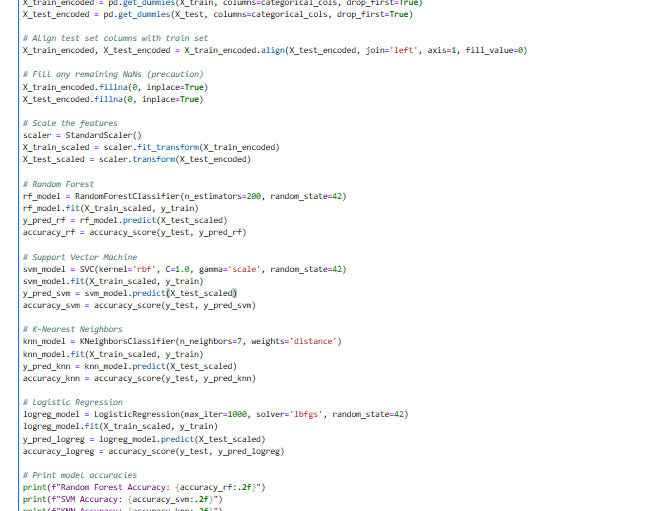


Fig:2.5 Model Training

**Data Preparation**

The Data Preparation phase is responsible for cleaning, encoding, and preparing the dataset for use in machine learning models. Key steps include:

Loading the dataset using pandas to read from the CSV file into a Data Frame for flexible data manipulation.

Removing records with missing target labels by eliminating rows where the loan approval status is missing, ensuring that the data used for training is valid and complete.

Separating features and target labels. The features include customer information such as income, credit score, and employment status, while the target label is the loan approval status (approved or denied).

Encoding categorical features, such as loan purpose or employment status, using Label Encoder to convert categorical values into numerical representations, which are essential for machine learning algorithms.

Handling missing values through imputation techniques like Simple Imputer with the most frequent strategy, ensuring that any missing values in the dataset are appropriately filled.

**2.2.1 Data Description**

The dataset used in this study, titled Loan\_Approval\_data.csv, is a structured dataset designed for predicting the loan approval status. The dataset contains clinical and financial information, including both demographic and financial features, as well as the loan approval status.

**Number of instances (rows):** 5000

**Number of attributes (columns):** 12

**Key features:**

Age (numeric)

Gender (categorical)

Income (numeric)

Credit Score (numeric)

Employment Status (categorical)

Loan Amount (numeric)

Loan Purpose (categorical)

Approval Status (target label: approved or denied)

The target variable, Approval Status, is a categorical variable that indicates whether a loan was approved or denied, based on customer data.

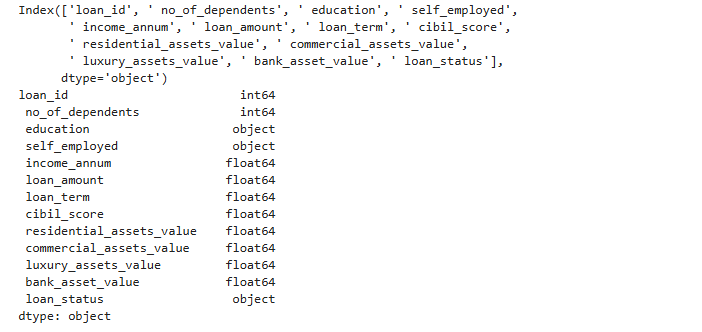


Fig:2.6 key Features

**2.2.2 Data Preprocessing**

The Data Preprocessing phase ensures the dataset is prepared for machine learning tasks:

**Encoding categorical variables**: Categorical variables such as Gender, Employment Status, and Loan Purpose are encoded into numeric form using Label Encoder to enable machine learning algorithms to process them.

**Handling missing values**: Any missing feature values are imputed using the **Simple Imputer** with the most frequent strategy, ensuring that the dataset remains complete and ready for model training.

**Scaling features**: Numerical features, such as Income and Credit Score, are scaled to ensure they are within a similar range, which improves the performance of certain models like KNN.

This preprocessing pipeline ensures that the dataset is clean, consistent, and ready for machine learning tasks, facilitating effective model training and evaluation.

data, providing an unbiased estimate of its generalization capabilities.

**CHAPTER 3**

**RESULTS**

**CONSEQUENCES AND FINDINGS**

The ability to accurately predict loan approval status is crucial in the financial services domain. A robust loan prediction model can enhance operational efficiency, reduce manual errors, and ensure that only eligible applicants receive loans—minimizing financial risk and improving customer satisfaction. However, achieving high accuracy in such predictions poses challenges due to the complexity and diversity of financial data, as well as missing or inconsistent entries in applicant records.

Model Construction Using Various Machine Learning Techniques

To identify the most effective model for predicting loan approval status (Approved or Rejected), several machine learning algorithms were implemented and evaluated. These include:

Logistic Regression

Support Vector Machines (SVM)

K-Nearest Neighbors (KNN)

Each model was trained on a preprocessed dataset containing demographic, employment, credit, and financial attributes. The models were evaluated based on standard metrics such as accuracy, precision, recall, and F1-score. Among these, the K-Nearest Neighbors (KNN) model outperformed others, achieving an accuracy of approximately 83% on the testing dataset. The KNN model’s strength in handling non-linear relationships and noisy data made it suitable for deployment in the subsequent Flask-based web application for real-time loan prediction.

**Accuracy and Model Evaluation**

The primary performance indicator was accuracy, representing the proportion of correctly predicted outcomes. The KNN model achieved an accuracy of approximately 83%, which was higher compared to Logistic Regression and SVM models. This high performance indicates the model’s effectiveness in accurately identifying eligible loan applicants based on their financial and demographic profiles.

Such a model can significantly streamline the loan approval process, enabling banks and financial institutions to make faster, data-driven decisions. However, to ensure model generalizability and fairness, it is essential to validate performance across diverse datasets collected from various regions, financial institutions, and loan types.

To further enhance the model, future efforts will include:

Hyperparameter SMOTE (Synthetic Minority Over-sampling Technique) for addressing class imbalance,

and cross-validation techniques for robust performance evaluation.

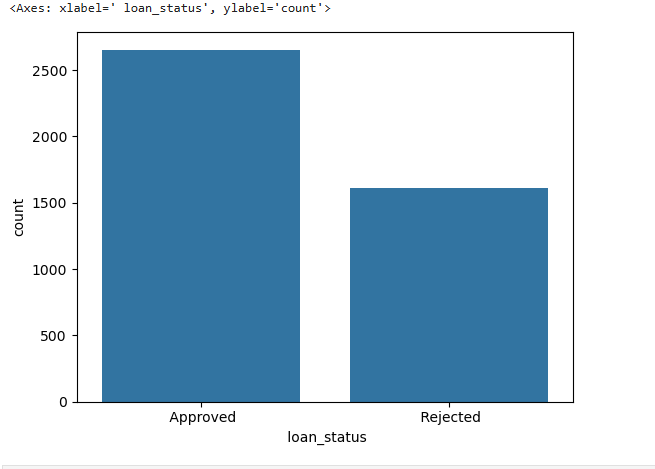


Fig. 3.1. Bar graph between amount and loan\_status

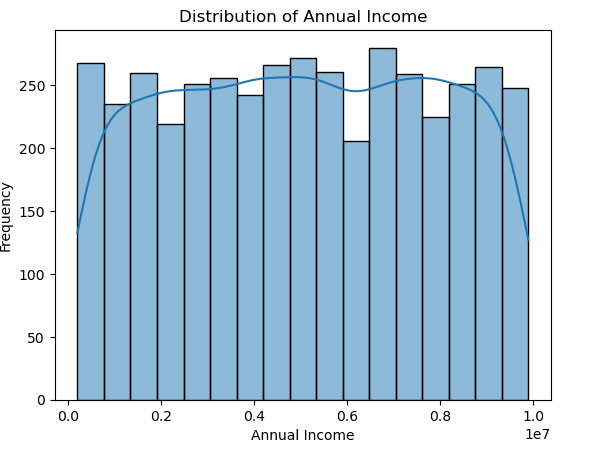
****

Fig 3.2: Bar graph between frequency and annual income

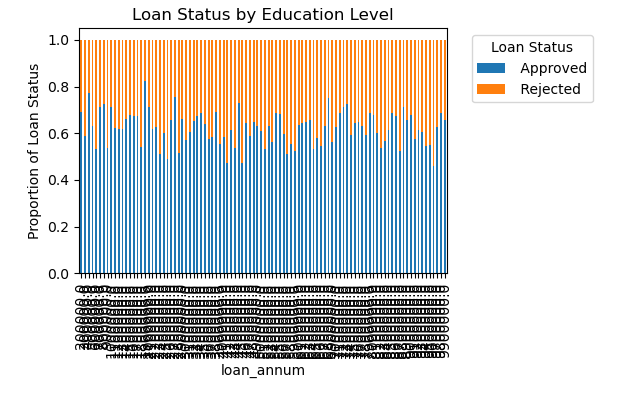
****

Fig 3.3: Bar graph between proportion of loan status and status

**CHAPTER 4**

**CONCLUSION AND FUTURE SCOPE**

**Conclusion**

Developing a predictive loan approval system using machine learning involves a systematic and data-driven approach to enhance the efficiency and accuracy of loan decision-making processes. The project begins with the collection and preparation of historical loan application data, which includes various applicant details such as income, employment status, credit score, loan amount, and repayment history. This data is then cleaned to handle missing values and outliers, and is transformed through encoding and normalization to ensure compatibility with machine learning algorithms. Feature engineering plays a crucial role in identifying the most influential factors affecting loan approvals, thereby improving the model's interpretability and performance.

Several classification algorithms such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting are trained and evaluated to determine the most effective model for predicting loan approval outcomes. Hyperparameter tuning techniques like Grid Search and Randomized Search are applied to optimize model accuracy and stability. To address potential issues of class imbalance, techniques like the Synthetic Minority Over-sampling Technique (SMOTE) are employed, especially to improve prediction accuracy for minority classes, such as rejected applications.

**Future scope**

Once the model demonstrates satisfactory performance, it can be integrated into a real-time web-based system or a loan management platform. This enables the model to receive input from new applicants and instantly predict loan approval decisions, aiding financial institutions in making faster and more informed choices. Further, validating the model on external and regionally diverse datasets

## REFERENCES

1. Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, yffud. (2016). Ultrasound Nerve Segmentation. Kaggle.
2. Baby, Mariya, and A. S. Jereesh. "Automatic nerve segmentation of ultrasound images." 2017 international conference of electronics, communication and aerospace technology (ICECA). Vol. 1. IEEE, 2017.
3. Hafiane, Adel, Pierre Vieyres, and Alain Delbos. "Deep learning with spatiotemporal consistency for nerve segmentation in ultrasound images." arXiv preprint arXiv:1706.05870 (2017).
4. Wang, Rui, Hui Shen, and Meng Zhou. "Ultrasound nerve segmentation of brachial plexus based on optimized ResU-Net." 2019 IEEE international conference on imaging systems and techniques (IST). IEEE, 2019.
5. Suissa, Noam, et al. "Quantifying ultrasound medical image segmentation for peripheral nerve blocks: a comparison of expert evaluations." *British Journal of Anaesthesia* 132.2 (2024): 428-430.
6. Srivastava, Somya, Ankit Vidyarthi, and Shikha Jain. "Analytical study of the encoder-decoder models for ultrasound image segmentation." Service Oriented Computing and Applications 18.1 (2024): 81-100.
7. Siddique, Nahian, et al. "U-net and its variants for medical image segmentation: A review of theory and applications." *Ieee Access* 9 (2021): 82031-82057.
8. Du, G., Cao, X., Liang, J., Chen, X., & Zhan, Y. (2020). Medical Image Segmentation based on U-Net: A Review. Journal of Imaging Science & Technology, 64(2).
9. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (pp. 3-11). Springer International Publishing.

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru

## Department of CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

## Program Outcomes (POs)

**Engineering Graduates will be able to:**

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need forsustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

## Program Specific Outcomes (PSOs):

## 

## PSO1 : Design, develop, test and maintain reliable software systems

## and intelligent systems. PSO2 : Design and develop web sites, web apps and

## mobile apps.

**PROJECT PROFORMA**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification of**  **Project** | **Application** | **Product** | **Research** | **Review** |
| √ |  |  |  |

**Note: Tick Appropriate category**

|  |  |
| --- | --- |
| **Machine Learning Outcomes** | |
| Course Outcome (CO1) | Describe machine learning and different forms of learning. |
| Course Outcome (CO2) | Use statistical learning techniques to solve a class of problems. |
| Course Outcome (CO3) | Build support vector machine for the given data to create optimal boundary  that best classifies the data. |
| Course Outcome (CO4) | Design neural networks to simulate the way human brain analyzes and processes information. |
| Course Outcome (CO5) | Solve classification problems using a decision tree. |

**Mapping Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CS3509 : MACHINE LEARNING** | | | | | | | | | | | | | | | |
| **Course Outcomes** | **Program Outcomes and Program Specific Outcome** | | | | | | | | | | | | | | |
| **PO 1** | **PO 2** | **PO 3** | **PO 4** | **PO 5** | **PO 6** | **PO 7** | **PO 8** | **PO 9** | **PO 10** | **PO 11** | **PO 12** |  | **PSO 1** | **PSO 2** |
| CO1 | 1 | 1 |  |  |  |  |  |  |  |  |  | 1 |  |  |  |
| CO2 | 1 |  |  |  |  |  |  |  |  |  |  | 1 |  |  |  |
| CO3 | 2 | 3 | 2 |  |  |  |  |  |  |  |  | 2 |  | 1 |  |
| CO4 | 2 | 2 | 3 | 2 |  |  |  |  |  |  |  | 2 |  | 2 |  |
| CO5 | 1 | 2 | 3 | 1 |  |  |  |  |  |  |  | 2 |  | 1 |  |

**Note: Map each Data Mining outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:**

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3- Substantially (High) mapped