Industrial Training

on

Predicting Diabetes Clinical Risk Factors with Various Machine Learning Techniques and Correcting Class Imbalance

SUBMITTED BY

Nehal Murdeshwar
210962021
Roll No:8

CSE(AI&ML) - B
nehalmurdeshwar@gmail.com
6364661543

Under the Guidance of:

Nagesh Singh

Executive Director

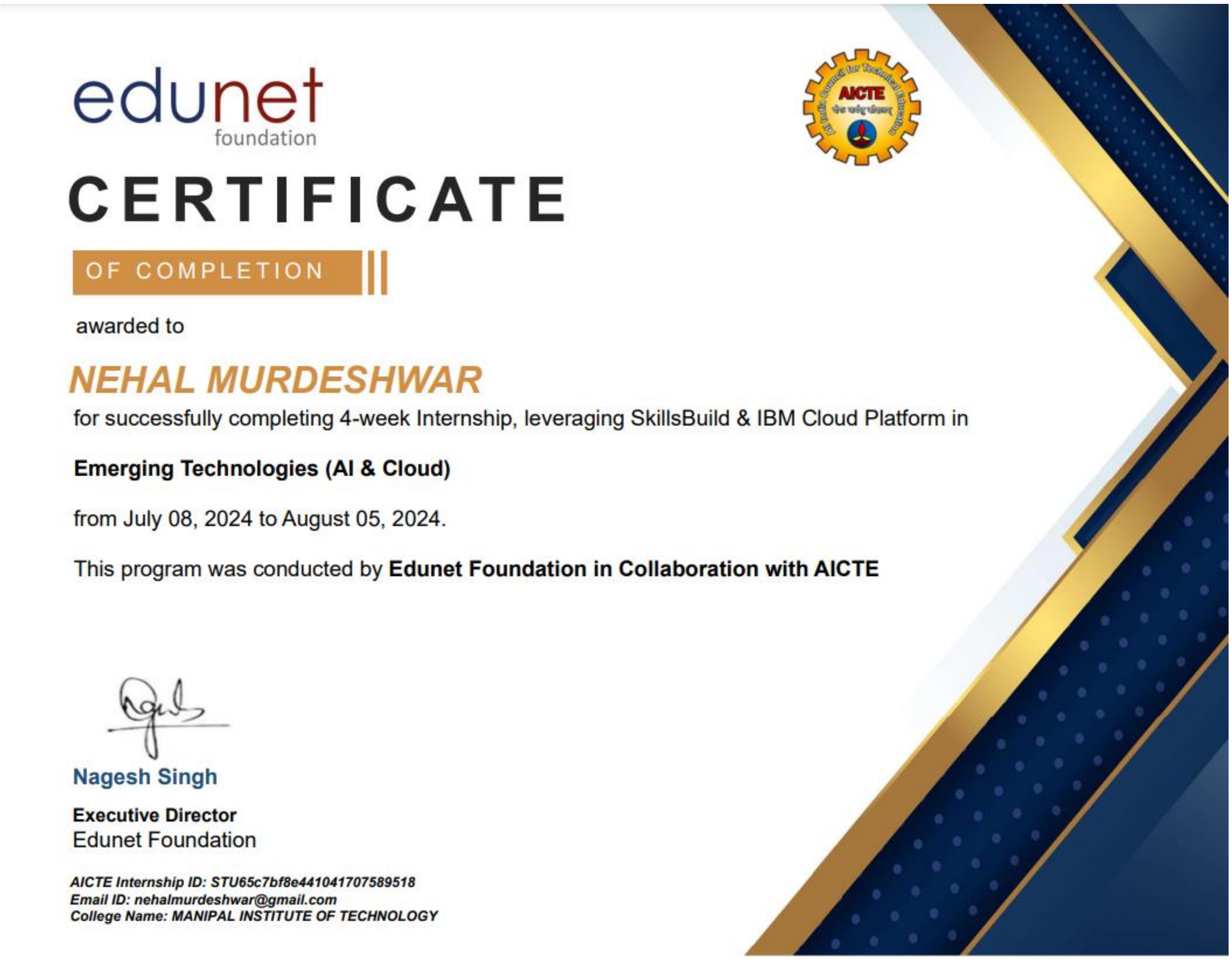
Edunet Foundation

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Department of Computer Science and Engineering

August 2024

Internship Certification



edunet
foundation

CERTIFICATE

OF COMPLETION

awarded to

NEHAL MURDESHWAR

for successfully completing 4-week Internship, leveraging SkillsBuild & IBM Cloud Platform in

Emerging Technologies (AI & Cloud)

from July 08, 2024 to August 05, 2024.

This program was conducted by **Edunet Foundation in Collaboration with AICTE**

Nagesh Singh

Executive Director
Edunet Foundation

AICTE Internship ID: STU65c7bf8e441041707589518
Email ID: nehalmurdeshwar@gmail.com
College Name: MANIPAL INSTITUTE OF TECHNOLOGY

1. Acknowledgements

I am incredibly appreciative of the profound opportunity I had at Edunet Foundation during my month-long summer internship. This experience has been nothing short of transformative, and I owe my gratitude to numerous individuals and the encouraging atmosphere that made this journey so enlightening.

To begin with, I want to sincerely thank my bosses and mentors. Their unwavering support, invaluable advice, and endless patience in imparting their knowledge have been nothing less than priceless. Their guidance has not only honed my technical proficiency but also broadened my understanding of the field in ways I never imagined. I deeply appreciate their constructive criticism, which has played a significant role in my growth both personally and professionally. Their extensive subject matter expertise, coupled with their exceptional teaching and communication skills, allowed me to grasp complex concepts quickly and apply them effectively to my Capstone project.

In addition to my mentors, I must also express my gratitude to the organization as a whole for their seamless support throughout the internship. The administrative team, HR department, and everyone involved in coordinating this experience deserve recognition for their tireless efforts in ensuring everything ran smoothly. Their contributions were vital in making this internship both enjoyable and productive.

Lastly, I would be remiss not to thank my friends and family for their unwavering support during this journey. Their encouragement and belief in me were constant sources of motivation, inspiring me to give my best effort every day.

As my internship comes to an end, I find myself reflecting on the wealth of knowledge I've gained, the incredible experiences I've had, and the strong network of amazing individuals I've been fortunate enough to connect with. My time at Edunet Foundation has been instrumental in shaping my career aspirations and has ignited an even greater passion for technology and innovation.

I am deeply grateful to Edunet Foundation for this wonderful opportunity. With renewed excitement and confidence, I eagerly look forward to the future and the challenges and opportunities that lie ahead.

2. Abstract

This Industrial Training focused on the project titled "Predicting Diabetes Clinical Risk Factors with Various Machine Learning Techniques and Correcting Class Imbalance". Diabetes, a highly prevalent and rapidly spreading disease, affects millions of individuals across all age groups annually, significantly reducing life expectancy. Due in large part to its enormous prevalence, diabetes can develop into more serious problems like stroke, renal failure, cardiovascular disease, and organ damage, all of which can be avoided with prompt diagnosis.

The project involved analyzing data from the combined NHANES dataset, spanning from 1999–2000 to 2015–2016. The objective was to apply various supervised machine learning techniques to identify and analyze the clinical risk factors associated with diabetes. A key aspect of this project was addressing class imbalance, a common issue in medical datasets, to ensure more accurate and reliable predictions.

By utilizing techniques such as ANOVA and logistic regression, the study aimed to uncover significant risk factors contributing to diabetes development. The integration of machine learning allowed for a comprehensive analysis, leading to a deeper understanding of how these risk factors correlate with the onset of diabetes. This project holds the potential to improve early detection and intervention strategies, ultimately reducing the burden of diabetes on individuals and healthcare systems.

Table Of Contents

3. Details of the Organization

Edunet Foundation is a prominent educational organization in India, dedicated to empowering learners and educators through a wide range of innovative programs. With a focus on skill development, digital literacy, and employability enhancement, the foundation partners with government bodies, educational institutions, and industry leaders to bridge the gap between academia and the evolving job market.

One of the key initiatives of Edunet Foundation is its internship programs, which provide students and young professionals with hands-on experience in various cutting-edge fields. These internships are designed to offer real-world exposure in areas such as artificial intelligence, machine learning, cloud computing, data analytics, and more. By working on live projects and receiving mentorship from industry experts, interns at Edunet Foundation gain invaluable skills that enhance their employability and career prospects.

In addition to technical training, the foundation's internships focus on holistic development, fostering critical thinking, problem-solving, and leadership abilities. Edunet Foundation is committed to ensuring that its interns are well-equipped to meet the demands of the modern workforce, making a smooth transition from academic environments to professional settings.

Beyond internships, Edunet Foundation offers a wide array of educational programs, including STEM education, vocational training, and entrepreneurship development. The foundation also engages in community development projects, providing access to education and resources for underprivileged communities. With a dedicated team of educators, trainers, and industry professionals, Edunet Foundation continues to drive positive change, impacting the lives of learners across India.

4. Problem Statement

The main problem statement addressed is as follows:

• **Design and develop a machine learning model for predicting diabetes risk factors using the NHANES dataset:**

- **Data Collection and Preprocessing:** Analyze various clinical and demographic features to prepare the dataset for predictive modeling.

- **Feature Selection:** Identify key features correlated with diabetes, using statistical techniques like ANOVA.

- **Modeling Techniques:** Apply multiple machine learning algorithms, such as logistic regression, support vector machine and random forest classification, to predict diabetes risk factors. Usage of a neural network is also possible.

• **Class Imbalance Resolution:** Address the challenge of class imbalance in the dataset by implementing appropriate strategies, ensuring that the model effectively identifies diabetes risk in underrepresented classes.

• **Model Evaluation:** Evaluate the performance of the developed models using key metrics, and determine the most effective model for predicting diabetes risk factors.

• **Deploy the Solution:** Develop a user-friendly solution that accurately predicts diabetes risk factors, with the potential for future deployment in clinical settings.

4.1 Technologies Used

- **Python:** Python is the core programming language utilized for this project due to its simplicity, versatility, and powerful libraries tailored for data science and machine learning. It enables seamless integration of various analytical and predictive modeling frameworks essential for this study.

- **Jupyter Notebook:** Jupyter Notebook provides an interactive coding environment that allows for the execution of Python code, visualization of data, and documentation all within a single interface. This tool was instrumental in developing, testing, and presenting the machine learning models, facilitating an iterative approach to improving model accuracy.

- **Pandas**: As a library for data manipulation and analysis, Pandas played a crucial role in preprocessing and analyzing the NHANES dataset. It was used extensively for cleaning data, handling missing values, and preparing the dataset for modeling, ensuring that the data was in an optimal format for analysis.

- **Scikit-learn**: This library provides simple and efficient tools for predictive data analysis and is integral to the machine learning aspect of the project. It was used for implementing various machine learning algorithms, including logistic regression and random forest classifiers, and for handling class imbalances through techniques like weighting classes.

- **Matplotlib and Seaborn**: Data visualisation and charting were done using both libraries. While Seaborn, which was built on top of Matplotlib, offered a high-level interface for creating visually appealing and educational statistical graphics, Matplotlib made it easier to create a variety of statistical charts.

- **Statsmodels**: This Python package facilitates statistical testing, statistical data exploration, and the estimation of numerous statistical models. Specifically, it was applied to ANOVA testing and logistic regression analysis to identify significant diabetes predictors.

- **NumPy**: This package is fundamental for scientific computing with Python. It supports a wide range of mathematical operations and is used for handling numerical data, which is crucial for the statistical analysis involved in this project.

- **IBM Cloud**: IBM Cloud was used to host and deploy the machine learning models, ensuring that they are scalable and accessible. The cloud environment provided the necessary computational resources to handle large datasets and intensive computations efficiently, facilitating a smooth deployment and operation of the predictive models.

5. Description

Developing successful predictive models for healthcare is an ongoing endeavor in the dynamic and frequently complex realm of medical research. To enhance the accuracy of clinical diagnostics and maximize early intervention, this project presents a novel approach to predicting diabetes risk factors using multiple machine learning techniques and resolving class imbalances. The approach is intended to optimize diagnostic precision while skillfully minimizing false negatives or positives, with a primary focus on early detection and prevention.

Medical research involves evaluating risk factors by analyzing statistical trends from health data, such as blood test results, BMI, age, and cholesterol levels. Unlike general medical assessments, which may rely on a broad range of diagnostic tests and medical history, this project focuses on identifying specific patterns and trends in clinical data to predict the risk of diabetes. Key tools and techniques include logistic regression, ANOVA, and advanced machine learning algorithms like random forest classification, support vector machines, decision trees, and artificial neural networks. Researchers believe that historical health data can offer valuable insights into future risk factors, as patterns in physiological and metabolic markers often repeat over time. This method is widely used by healthcare professionals to make informed diagnostic decisions, aiming to capitalize on early detection and optimize intervention strategies.

The aim of the project is to utilize machine learning to analyze clinical and demographic features to generate predictive insights. In the realm of diabetes risk assessment, this relates to examining and deciphering patterns from health records and lab results without relying solely on symptomatic data or direct patient examinations. When making diagnostic decisions, medical professionals pay close attention to statistical patterns, health trends, and important indicators like glucose and cholesterol levels. Logistic regression, random forest classifiers, SVM, DT, ANN, and ANOVA are examples of common tools used. With this method, healthcare providers can forecast potential health issues based on past patient data and epidemiological studies. It is based on the idea that all relevant clinical information is reflected in the collected health data.

The project makes use of various machine learning concepts:

- **Logistic Regression:** Used to identify the most significant predictors of diabetes, incorporating statistical significance and confidence intervals to validate findings.

- **Random Forest Classifier:** Applies ensemble learning techniques to improve prediction accuracy and address class imbalances effectively.
- **Support Vector Machine (SVM):** Utilized for its robust classification capabilities, particularly in high-dimensional spaces.
- **Decision Tree (DT):** Offers intuitive model interpretations, facilitating understanding of the decision-making process.
- **Artificial Neural Network (ANN):** Provides sophisticated modeling capabilities to capture complex nonlinear relationships in the data.
- **ANOVA:** Helps in determining which variables significantly impact the diabetes outcome, ensuring that the model focuses on relevant features.

By employing these techniques, the project seeks to establish a robust predictive model that can significantly aid in the early detection and management of diabetes, potentially reducing the incidence and severity of long-term complications.

5.1 Methodology

1. **Data Collection:**
    - Utilize the NHANES dataset, which spans from 1999-2000 to 2015-2016, encompassing demographic, examination, questionnaire, and laboratory data relevant to diabetes.
2. **Data Preprocessing:**
    - **Outlier Handling:** Identify and manage outliers using the Interquartile Range (IQR) method. Replace extreme values with median values to maintain data integrity.
    - **Class Imbalance Resolution:** Address class imbalance by assigning appropriate class weights during the training phase to ensure balanced model performance.
    - **Data Standardization:** Rescale the dataset to ensure the mean is 0 and the standard deviation is 1, which minimizes the impact of outliers on the model.
3. **Feature Selection:**
    - Apply statistical methods such as ANOVA to determine the significance of each feature. Select 25 independent variables that have the highest impact on predicting diabetes.

4. **Risk Factor Analysis:**

    - Use Logistic Regression and ANOVA to identify significant risk factors correlated with diabetes, such as age, blood-related diabetes, cholesterol, and BMI.

5. **Model Development**

    - Develop and train multiple machine-learning models, including:

        - **Logistic Regression (LR):** To analyse the relationship between the risk factors and the outcome variable.

        - **Support Vector Machine (SVM):** To classify the data based on a decision boundary that maximizes the margin between classes.

        - **Random Forest (RF):** To build an ensemble of decision trees for robust prediction.

        - **Decision Tree (DT):** To provide a simple yet effective classification model.

        - **Artificial Neural Network (ANN):** To model complex patterns and relationships in the data.

6. **Cross – Validation and Ensemble Voting Method:**

    - Perform cross-validation using the ShuffleSplit method to partition the data into training and testing sets. Evaluate model performance using metrics like accuracy and AUC (Area Under the Curve).

    - Combine the predictions from multiple models using a hard voting ensemble method to improve overall accuracy and robustness.

7. **Result Analysis:**

    - Compare the performance of different models in terms of accuracy and AUC scores. Identify the best-performing model, which in this case, is the Random Forest classifier with the highest accuracy and AUC score.

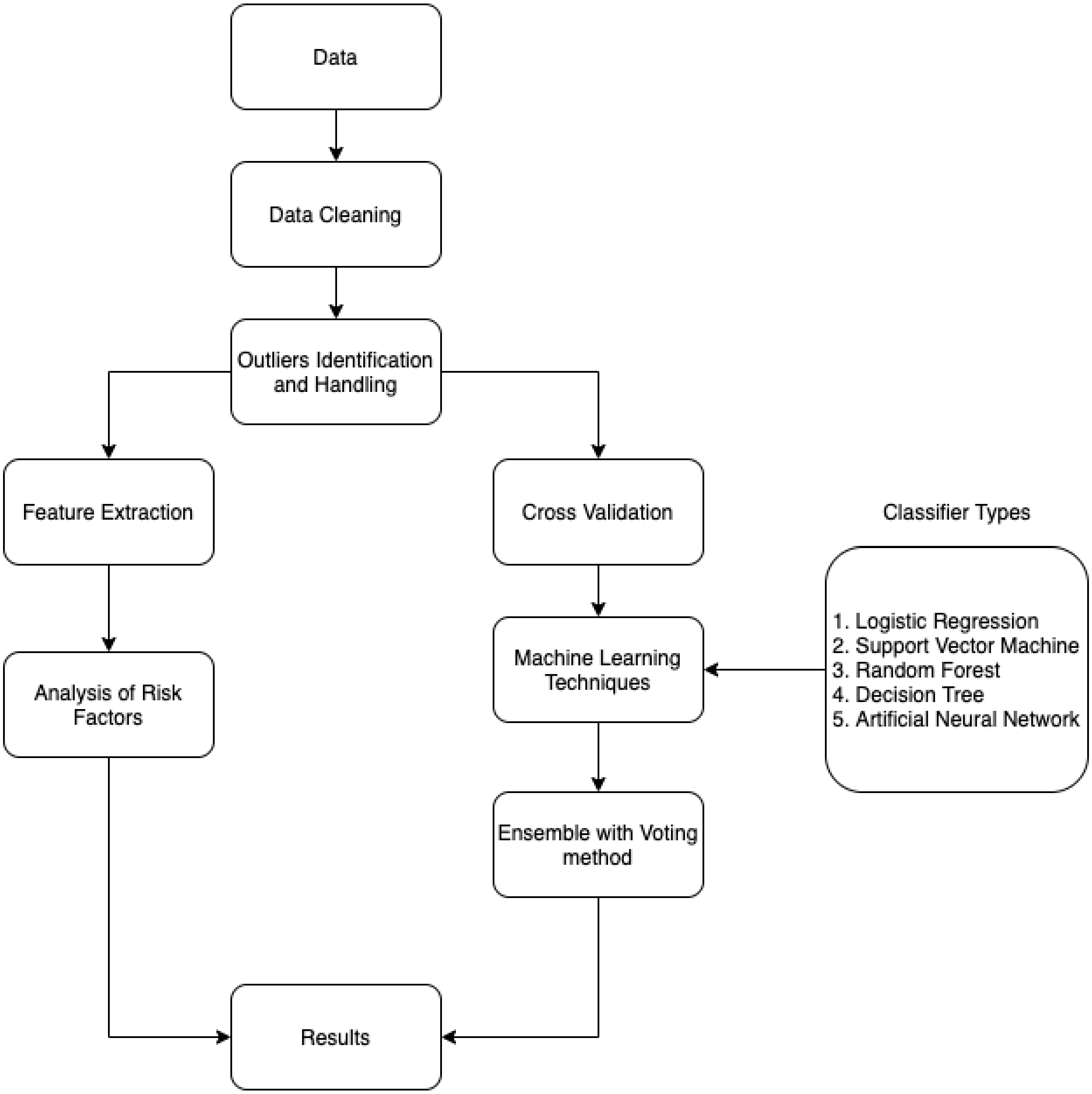


Fig 5.1.1 Methodology Flowchart

5.1.1 Data Collection and Exploratory Data Analysis (EDA)

The National Health and Nutrition Examination Survey provided the dataset for this study (NHANES). For this project, the assembled version NHANES data from 1999–2000 to 2015–2016 was used. NHANES is a health and nutrition assessment program that handles a variety of data types and is based on the US population. An aggregate of laboratory, survey, demographic, and examination data from 1999–2000 to 2015–2016 makes up the dataset used in this study. The dataset has also been used in a number other studies. The sample consists of responses from 37079 people who fall into 51 different categories of traits. Based on the impacts, a total of 25 independent factors were chosen in order to identify diabetic illnesses. In this study, diabetes is considered a dependent variable. There are 37079 responders in this dataset overall, made up of 32227 normal people and 4852 diabetes cases. The dataset includes an imbalance in classes based on the distribution of diabetes patients' classes; this issue is resolved by allocating class weights during training using various machine learning strategies. 51 distinct variables and 37079 respondents' information are included in the dataset.

Diabetes is considered as a dependent variable for this project. 25 independent variables were selected for this project based on their feature importance and effects on diabetes patients. Some factors (such as lymphocytes, mean cell volume, the yearly family income, eosinophils, basophils, the pulse rate in the 60s, the ratio of family income to poverty, monocytes, etc.) are not taken into consideration for additional analysis because they do not correlate with the outcome variable.

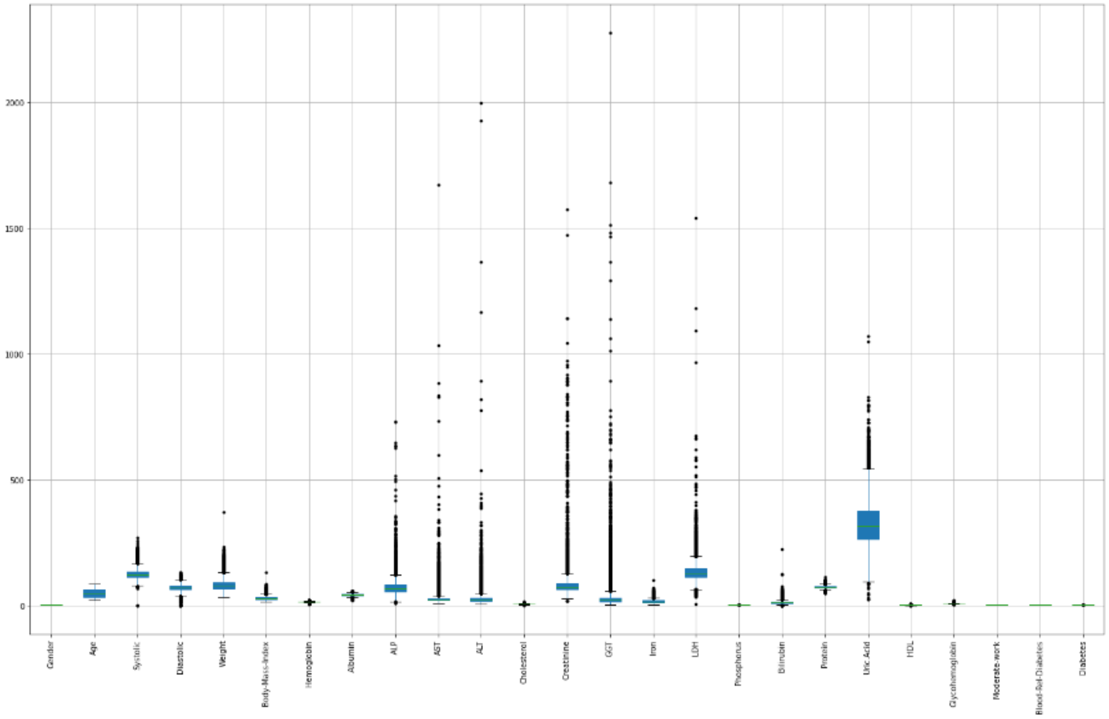


Fig 5.1.1.1 Box-Plot Visualization before Handling Outliers

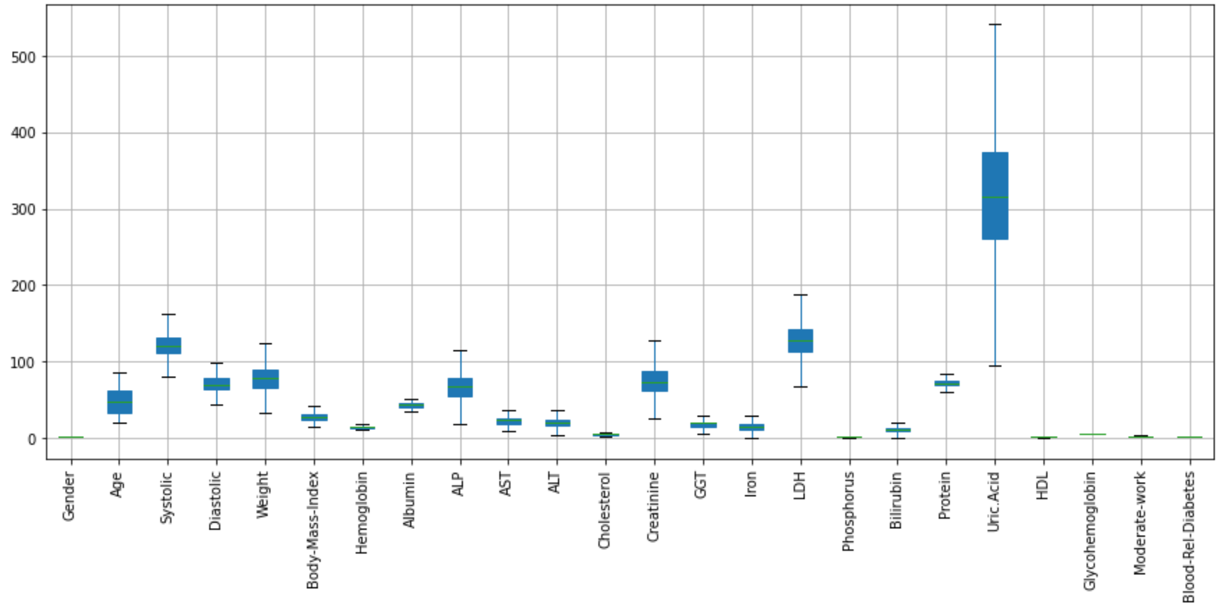


Fig 5.1.1.2 Box-Plot Visualization after Handling Outliers

5.1.2 Data Preprocessing

Outlier Handling

Outliers are data points that differ significantly from the rest of the data points in the dataset distribution. Anomalies or unusual values within a certain data distribution are called outliers. The effectiveness of statistical decision-making and machine learning algorithms is significantly impacted by outliers. The boxplot visualisations of the selected independent variables in the dataset that were produced for outlier analysis are shown in Fig. 5.1.2. Fig. 5.1.2 makes it clear that outliers affect the dataset. Ignoring them could have an impact on the efficiency of algorithms and the analysis of risk factors. The Interquartile Range (IQR) approach was used in this study to identify outliers. The values of the interquartile range from the first to the third quartile. The mathematical representation of IQR is:

$$IQR = Q3 - Q1$$

where Q3 and Q1 represent the third and first quartile respectively.

In order to identify the outliers and indicate the greatest and lowest value ranges, the lower and upper fences were computed. Outliers are values that deviate from both the lowest and maximum allowable range, as illustrated in Figure 5.1.2. The lower and upper fence's mathematical representation is as follows:

$$Upperfence = Q3 + (1.5 * IQR)$$

$$Lowerfence = Q1–(1.5 * IQR)$$

Class Imbalance Resolution

The dataset has 37079 respondents, of which 4852 have diabetes and 32227 are normal individuals. This suggests that there is a problem with class imbalance. Using the scikit-learn tool, class weights (Class 0: 0.52, Class 1: 0.47) are assigned to the classes in order to address the issue of class imbalance.

Data Standardisation

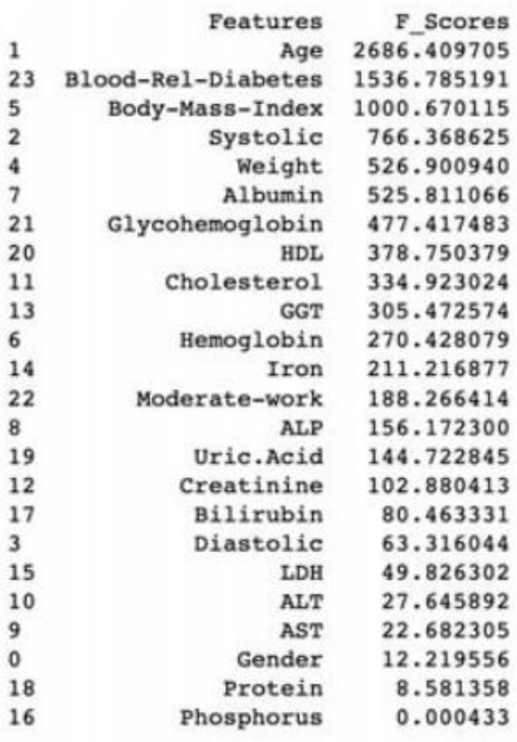
The process of standardisation involves rescaling the value distribution to maintain a mean of 0 and a standard deviation (std) of 1. Additionally, outliers have less of an impact . In this study, the scikit-learn tool was used to accomplish data standardisation following the management of the dataset's outliers.

5.1.3 Feature Selection

There are 51 independent variables in the dataset that was used in this study. However, a univariate feature selection approach and statistical logistic regression method were used for this study to account for 25 independent variables, representing their impacts on diabetes patients [1]. ANOVA F-scores for each variable verified the results of risk factors found by statistical logistic regression. Also, a comparison and validation of the resulting risk variables are made with the results of previous studies.

Analysis of Variance(ANOVA)

Analysis of Variance (ANOVA): There are twenty-five independent variables in the dataset used in this study. An algorithm for feature selection was used to identify the best features. Methods for feature selection also improve the accuracy of predictions. This study employed the analysis of variance (ANOVA), a feature selection method also used in other investigations. ANOVA is a straightforward and effective statistical technique for analysing the means of one or more groups. It also establishes the degree to which the groups differ from one another. The best features in this study were selected and arranged based on their ANOVA F-score, as indicated in Fig. 5.1.3.1 The F-score, which indicates the significance of each independent variable's particular characteristic, was calculated using the scikit-learn program.



```
              Features      F_Scores
1                  Age   2686.409705
23  Blood-Rel-Diabetes   1536.785191
5      Body-Mass-Index   1000.670115
2             Systolic    766.368625
4               Weight    526.900940
7              Albumin    525.811066
21     Glycohemoglobin    477.417483
20                 HDL    378.750379
11         Cholesterol    334.923024
13                 GGT    305.472574
6           Hemoglobin    270.428079
14                Iron    211.216877
22       Moderate-work    188.266414
8                  ALP    156.172300
19           Uric.Acid    144.722845
12          Creatinine    102.880413
17           Bilirubin     80.463331
3            Diastolic     63.316044
15                 LDH     49.826302
10                 ALT     27.645892
9                  AST     22.682305
0               Gender     12.219556
18             Protein      8.581358
16          Phosphorus      0.000433
```

Fig 5.1.3.1 Feature Importance based on ANOVA F-Score

5.1.4 Risk Factor Analysis

The logistic regression model is employed in this project to explore statistical data. I measured and determined the relative risk variables in our dataset using the statsmodels tool. Table 1 displays the p-value, confidence interval (C.I.), and odds ratio (O.R.) for each variable that was generated to examine the relative risks of the independent variables with respect to the outcome variable. These results indicate the risk and impacts of individual qualities on the dependent variable. The p-value is a useful tool for representing statistical significance. The range of the p-value is between 0 and 1. For statistical significance in this study, a p-value of less than 0.05 is considered acceptable. The most important risk factors, in addition to those linked to diabetes disease, include age, diastolic, cholesterol, and blood-related diabetics (Table 1). The relative risk connected to the outcome variable is also examined using the odds ratio (OR).

5.1.5 Model Development

Logistic Regression

Previous studies of diabetes patients' risk factors identification have employed logistic regression. Due to its binary structure, LR is one of the many popular supervised machine learning algorithms. The possible risk factors connected to diabetes patients are also examined in this study using statistical LR.

In LR, the dependent variable is a logit and the equation is:

$$logit(x) = log(\frac{x}{1-x}) = \beta_0 + \beta_1 y_1 + ... + \beta_y y_y$$

where, x represents the probability of diabetes patients when y=1 and 1-x represents the probability of non-diabetes patients when y = 0. The odd ratio and confidence intervals were found using the statistical logistic regression method, which also identifies possible risk variables related to diabetes. The features were selected and put in the order determined by the associated p-values. In this investigation, P-values less than 0.05 were considered to identify the particular risk factors. The risk factors are displayed in Table 1 along with their p-value, odd ratio, and confidence range.

Support Vector Machine

Support vector machines have also been used in earlier studies on the identification of diabetes risk factors (SVM). SVM is a popular supervised learning method that may be used to find hidden patterns as well as solve classification problems. Since SVM can locate a hyperplane, it is also known as a decision boundary function. The original train data is transformed using this method into a higher dimensional space while being monitored via nonlinear mapping. Afterwards, SVM distinguishes between patterns belonging to various classes. The mathematical equation of a hyperplane is:

$$W.Y + p = 0$$

where W represents the weighted vector and b represents the scalar data.

The Radial Basis Function (RBF) kernel of SVM is one of the classification techniques employed in this study. The RBF kernel's mathematical representation is:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$$

Decision Tree

The studies also employ classification based on decision trees. DT is a well-liked supervised learning technique that may be applied to both regression and classification models. It creates tree-structure-based regression or classification models. Using this method, the input features are built as nodes in a tree. It uses information gathered to choose the feature.

Random Forest

Prior research on diabetes diagnosis has also employed the Random Forest classification. It is possible to apply the random forest algorithm to regression and classification problems alike. It is a more complex variant of the bagging method, quick, and resistant to overfitting because it is based on the properties of decision trees.

Artificial Neural Network

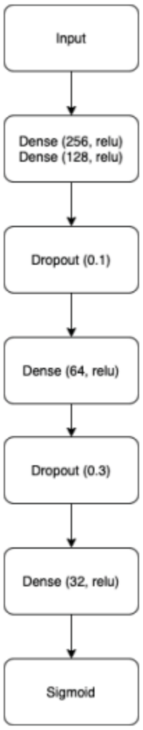


Fig 5.1.5.1 Proposed ANN Architecture

Several studies that predict diabetes also use artificial neural networks (ANN). In this study, dropout layers were included in addition to several dense layers to avoid overfitting. In order to address the issue of class imbalance, the dataset was trained using class weights. Figure 5.1.5.1 depicts the suggested architecture of the ANN.

5.1.6 Cross – Validation and Ensemble with Voting Method

A technique for data partitioning called cross-validation (CV) separates the dataset into two groups: train data and test/validation data. The ShuffleSplit function, which produces a user-defined number of independent train and validation splits, was used as the cross-validation technique in this study. By using this strategy, the samples are divided into train and test/validation sets by shuffling them around. Having control over the quantity of data samples on both sides of the train/validation set and the number of iterations is the reason this method was chosen. Using a cross-validation process, the dataset was split into 80% training set and 20% set/validation set. Using the scikit-learn tool, the data was evaluated using CV/split values of 3, 5, and 10 for the train and validation splits that had been set.

The ensemble method of assessing the model's performance is incredibly effective. In this study, the hard voting approach was used for ensembling. All of the earlier techniques were combined with the hard voting method in this study. The output equation for the hard voting method is:

$$X_{pred} = mode(P_1(y), P_2(y), P_3(y), ..., P_n(y))$$

where P1, P2, P3 is the expected value, represented by Xpred. P stands for the classifiers, y for the classifiers, and y for the classifier's input. The mode value of the labels that the classifiers anticipate is produced by the mode() function.

5.1.7 Result Analysis

<u>*Evaluation Metrics*</u>

Several statistical measures were utilised in this study as assessment metrics to quantify the risk factors and the impacts of individual variables on the independent variable, including the odd ratio (OR), P-value, and confidence interval (CI). In addition to this, I also employed accuracy and AUC score as evaluation metrics to assess how well our suggested model performed.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

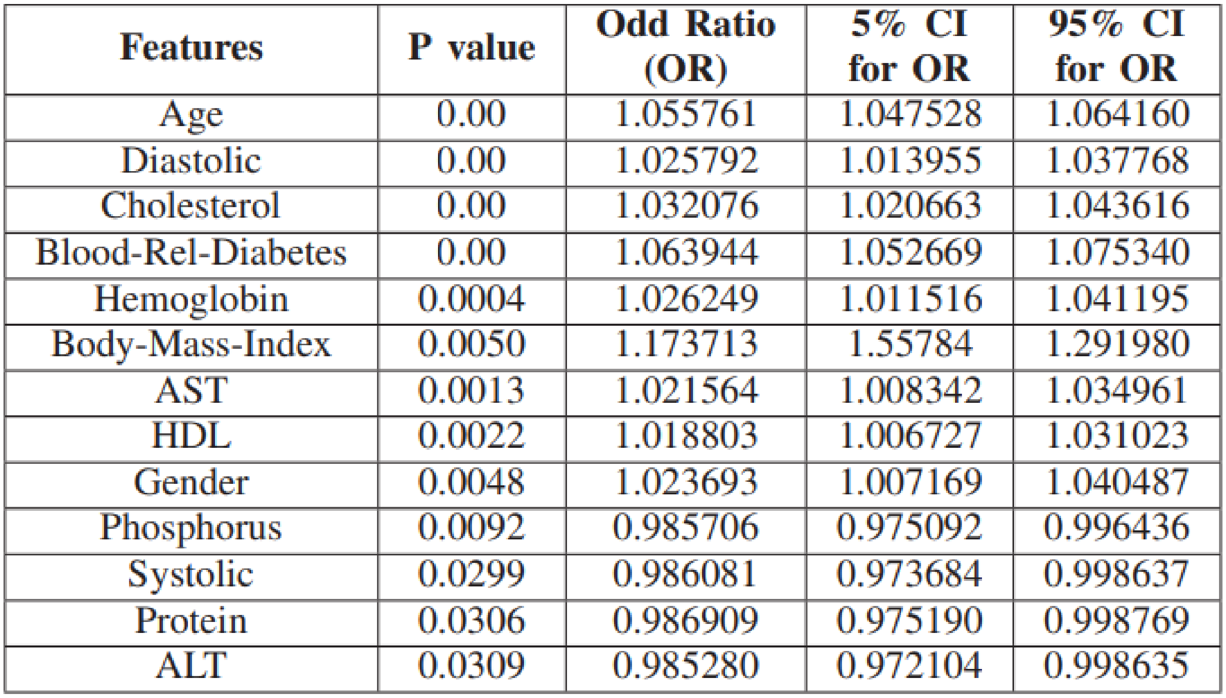
where the values for True Positives, True Negatives, False Positives, and False Negatives are represented, respectively, by the symbols TP, TN, FP, and FN. The AUC score, which plots sensitivity against 1 - specificity and indicates the overall efficacy of the model, is the product of the True Positive (TP) and False Positive (FP) rates.

Statistical Analysis of Risk Factors

The variables in this study were chosen for analysis based on their statistical significance and had a p-value less than 0.05. The most important risk factors linked to diabetes class include age, diastolic blood pressure, cholesterol, blood-related diabetes, BMI, etc. Figure 5.1.3.1 also displayed the feature importance following the application of Analysis of Variance (ANOVA) to handle the outliers. The chosen risk variables are validated based on the F-score. The most important risk variables for diabetes include age, blood-related diabetes, cholesterol, and BMI, according to the intersection of statistical logistic regression and ANOVA methodologies. By contrasting these risk factor findings with those from other investigations, they were verified and validated.

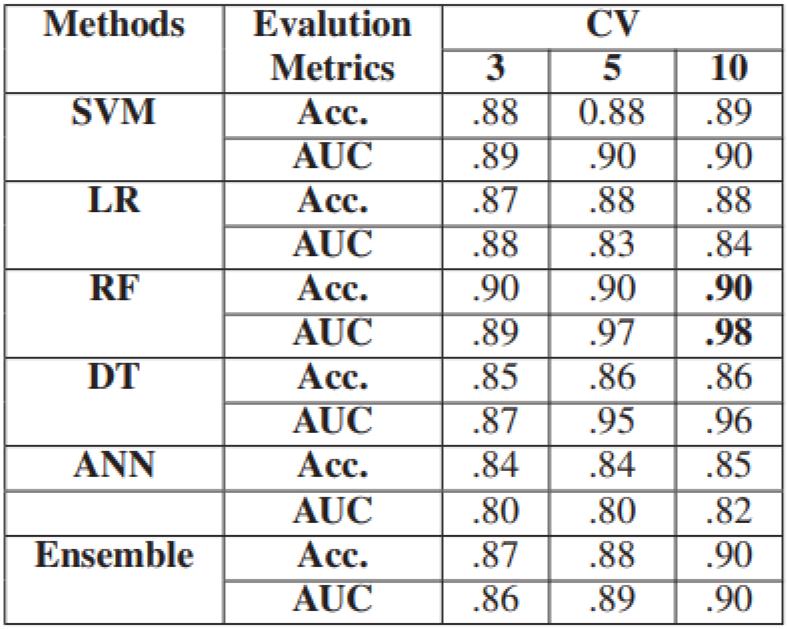
Final Results

The comparison of accuracy and AUC scores with various CV values is shown in Table 2. The dataset was tested using three, five, and 10 split/cv values. This study employed a number of classification techniques, including ensemble approach, support vector machine, logistic regression, random forest, decision tree, and artificial neural network. The assessment metrics listed in Table 2 have undergone considerable modifications. The random forest classification approach (RF) outperformed the other classification techniques in terms of accuracy (90%) and AUC score (0.98).



Features	P value	Odd Ratio (OR)	5% CI for OR	95% CI for OR
Age	0.00	1.055761	1.047528	1.064160
Diastolic	0.00	1.025792	1.013955	1.037768
Cholesterol	0.00	1.032076	1.020663	1.043616
Blood-Rel-Diabetes	0.00	1.063944	1.052669	1.075340
Hemoglobin	0.0004	1.026249	1.011516	1.041195
Body-Mass-Index	0.0050	1.173713	1.55784	1.291980
AST	0.0013	1.021564	1.008342	1.034961
HDL	0.0022	1.018803	1.006727	1.031023
Gender	0.0048	1.023693	1.007169	1.040487
Phosphorus	0.0092	0.985706	0.975092	0.996436
Systolic	0.0299	0.986081	0.973684	0.998637
Protein	0.0306	0.986909	0.975190	0.998769
ALT	0.0309	0.985280	0.972104	0.998635

Table 1: Individual Risk Factors with their P-value, odd ratio and confidence intervals using Logistic Regression



Methods	Evalution Metrics	CV		
		3	5	10
SVM	Acc.	.88	0.88	.89
	AUC	.89	.90	.90
LR	Acc.	.87	.88	.88
	AUC	.88	.83	.84
RF	Acc.	.90	.90	**.90**
	AUC	.89	.97	**.98**
DT	Acc.	.85	.86	.86
	AUC	.87	.95	.96
ANN	Acc.	.84	.84	.85
	AUC	.80	.80	.82
Ensemble	Acc.	.87	.88	.90
	AUC	.86	.89	.90

Table 2: Comparison of accuracy and AUC score with different CV values

6. Conclusion

The main goals of this study were to create an effective classification system for diabetes patients based on their different characteristics and to discover and evaluate potential clinical risk factors linked to the disease. This study successfully addressed major challenges in data preprocessing by carefully reviewing 37,079 patient records from the NHANES dataset. These challenges included handling outliers by replacing extreme values with medians and reducing class imbalance by assigning class weights during training.

The study determined that the most important risk variables for diabetes were age, blood-related diabetes, cholesterol, and BMI using Logistic Regression and ANOVA techniques. Several classification techniques, including as Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Artificial Neural Network, were applied in order to validate these results.

The use of these machine learning approaches produced encouraging outcomes, with the Random Forest algorithm surpassing other cutting-edge techniques with the highest accuracy (0.90) and AUC score (0.98). These results demonstrate how well the suggested method can categorise diabetic patients, and they also emphasise how crucial feature selection and model optimisation are to the field of predictive healthcare analytics.

Although this study's results are impressive, there is yet opportunity for development. To further improve the precision and resilience of the predictive models, future studies may examine the use of more sophisticated methods and the integration of new data sources. There is a great deal of promise to enhance diabetes early diagnosis and management through further improvement of these approaches, which will ultimately lead to improved patient outcomes.

6.1 Environmental and Societal Impact

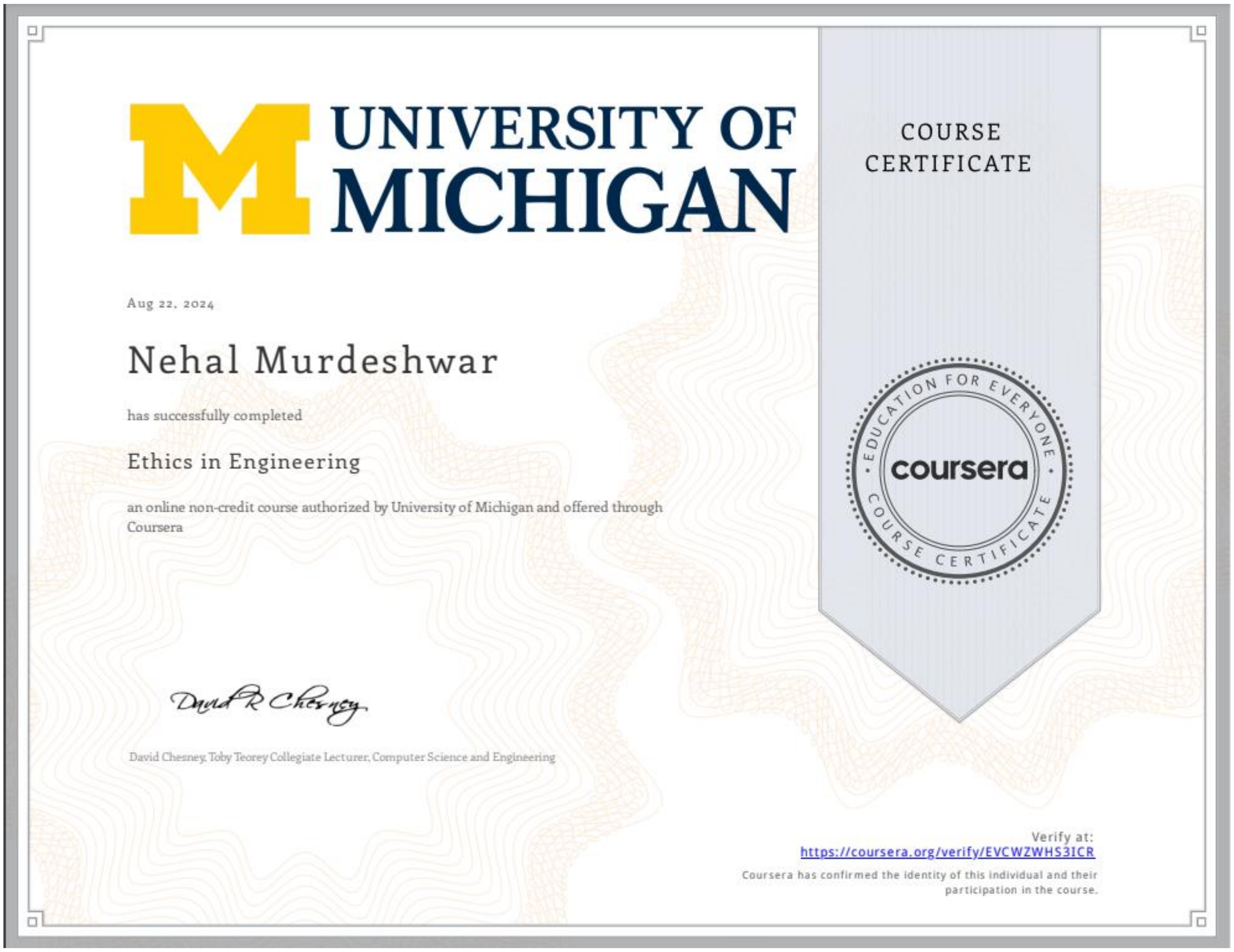
The development and implementation of machine learning models for predicting diabetes risk factors have a minimal direct environmental impact, as the process primarily involves digital resources and computational power. However, optimizing data processing algorithms and model training for energy efficiency can help minimize the carbon footprint associated with extensive computing tasks. Additionally, utilizing cloud services that adhere to green certifications and sustainable practices can further reduce the environmental impact.

From a societal perspective, this research has significant positive implications. By improving the early detection and management of diabetes, the project has the potential to contribute to better health outcomes and reduce the long-term burden of diabetes on healthcare systems. The use of machine learning models to identify high-risk individuals can lead to timely interventions, preventing the progression of diabetes and its associated complications. This, in turn, can enhance the quality of life for patients and reduce healthcare costs related to diabetes management.

Moreover, the project underscores the importance of technological innovation in healthcare. By integrating advanced data analytics into medical diagnostics, the research promotes the adoption of cutting-edge technologies in clinical settings. This contributes to the broader societal goal of advancing healthcare through technology, making it more efficient, accurate, and accessible. Additionally, the project fosters a deeper understanding of the role of data science in public health, encouraging further research and education in this critical field.

Overall, this project not only advances the field of healthcare analytics but also highlights the importance of sustainability and technological education in addressing global health challenges.

7. Ethics in Engineering Certification



UNIVERSITY OF MICHIGAN

COURSE CERTIFICATE

Aug 22, 2024

Nehal Murdeshwar

has successfully completed

Ethics in Engineering

an online non-credit course authorized by University of Michigan and offered through Coursera

David R Chesney

David Chesney, Toby Teorey Collegiate Lecturer, Computer Science and Engineering

EDUCATION FOR EVERYONE coursera COURSE CERTIFICATE

Verify at:
https://coursera.org/verify/EVCWZWHS3ICR

Coursera has confirmed the identity of this individual and their participation in the course.

8. NBA/IET Mapping

NBA - PROGRAM OUTCOMES (PO) & PROGRAM SPECIFIC OUTCOMES (PSO)

Engineering Graduates will be able to:

PO1:Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: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.

PO3: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.

PO4: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.

PO5: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.

PO6: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.

PO7:Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8:Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10:Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: 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.

PO12: 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.

PSO1: Analyse and solve real world problems by applying a combination of hardware and software. PSO2: Formulate & build optimised solutions for systems level software & computationally intensive applications.

PSO3: Design & model applications for various domains using standard software engineering practices. PSO4: Design & develop solutions for distributed processing & communication.

NBA CO PO Mapping

CSE 4298-ITR	**CO**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**	**PSO1**	**PSO2**	**PSO3**	**PSO4**
CSE 4298.1	Understand the functioning of the industry	1	0	0	0	0	0	0	2	2	0	0	0	1	0	0	0
CSE 4298.2	Understand the requirements of real world applications	2	1	3	1	1	1	1	0	3	2	0	1	2	1	1	0
CSE 4298.3	Demonstrate skills to use modern engineering tools, software and equipment to analyze problems	0	0	2	2	3	1	1	0	3	2	0	1	2	1	1	0
CSE 4298.4	Demonstrate an ability to envisage and work on laboratory and multidisciplinary tasks	0	0	2	1	3	1	1	0	3	2	0	1	2	1	1	0
CSE 4298 (Avg. correlation level)		**0.75**	**0.25**	**1.75**	**1**	**1.75**	**0.75**	**0.75**	**0.5**	**2.75**	**1.5**	**0**	**0.75**	**1.75**	**0.75**	**0.75**	**0**

NBA Program articulation Matrix

Course Code	Course Title	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
CSE 4298	Industrial Training	0.75	0.25	1.75	1	1.75	0.75	0.75	0.5	2.75	1.5	0	0.75	1.75	0.75	0.75	0

IET-AHEP Learning outcome statements	
Code	***Learning outcome (LO)***
C1.	Apply knowledge of mathematics, statistics, natural science and engineering principles to the solution of complex problems. Some of the knowledge will be at the forefront of the particular subject of study
C2.	Analyse complex problems to reach substantiated conclusions using first principles of mathematics, statistics, natural science and engineering principles
C3.	Select and apply appropriate computational and analytical techniques to model complex problems, recognising the limitations of the techniques employed
C4.	Select and evaluate technical literature and other sources of information to address complex problems
C5.	Design solutions for complex problems that meet a combination of societal, user, business and customer needs as appropriate. This will involve consideration of applicable health & safety, diversity, inclusion, cultural, societal, environmental and commercial matters, codes of practice and industry standards
C6.	Apply an integrated or systems approach to the solution of complex problems
C7.	Evaluate the environmental and societal impact of solutions to complex problems and minimise adverse impacts

C8.	Identify and analyse ethical concerns and make reasoned ethical choices informed by professional codes of conduct
C9.	Use a risk management process to identify, evaluate and mitigate risks (the effects of uncertainty) associated with a particular project or activity
C10.	Adopt a holistic and proportionate approach to the mitigation of security risks
C11.	Adopt an inclusive approach to engineering practice and recognise the responsibilities, benefits and importance of supporting equality, diversity and inclusion
C12.	Use practical laboratory and workshop skills to investigate complex problems
C13.	Select and apply appropriate materials, equipment, engineering technologies and processes, recognising their limitations
C14.	Discuss the role of quality management systems and continuous improvement in the context of complex problems
C15.	Apply knowledge of engineering management principles, commercial context, project and change management, and relevant legal matters including intellectual property rights
C16.	Function effectively as an individual, and as a member or leader of a team
C17.	Communicate effectively on complex engineering matters with technical and non-technical audiences
C18.	Plan and record self-learning and development as the foundation for lifelong learning/CPD

VII SEMESTER - CSE 4298 – Industrial Training - CLO - AHEP LO Mapping

Course Learning Outcome	Statement	C1	C2	C3	C4	C6	C7	C8	C11	C15	C18
CLO 4298.1	Apply mathematics, science and engineering skills to identify, formulate, synthesize and solve the problems from various areas of computer science engineering.	✓		✓							
CLO 4298.2	Have knowledge of new trends in engineering/technology by developing programmable coding in various computer programming languages.				✓						
CLO 4298.3	Use the industry standard tools to analyze, design, develop and test software engineering based applications.			✓							
CLO 4298.4	Apply theoretical knowledge to real-world engineering problems and manage complex engineering projects.		✓			✓					
CLO 4298.5	Understand the adverse societal impacts of the solutions to complex problems during project development.						✓				
CLO 4298.6	Identify and analyze ethical concerns to be followed during the practice school/research project internships and make reasoned moral choices guided by the internal and external supervisors during the tenure.							✓			
CLO 4298.7	Recognize the responsibilities deemed by the external and internal supervisors and understand the importance of supporting equality, diversity, and inclusion between peers.								✓		
CLO 4298.8	Apply knowledge of engineering management principles and understand why project and change management may be required during practice school and project work.									✓	
CLO 4298.9	Indicate the future direction of the project development and appreciate how it can be realized with collaborative, lifelong learning, and self-learning.										✓

Declaration: Through this Industrial Training, I have accomplished the above stated program articulation and IET learning outcomes.

9. References

[1] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, p. 7, 2020.

[2] N. S. El Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," 2018. [Online]. Available: https://www.researchgate.net/publication/326119049_Diabetes_Prediction_Using_Artificial_Neural_Network. [Accessed: 22-Aug-2024].

[3] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," *Expert Systems with Applications*, p. 113408, 2020.

[4] D. M. Hawkins, *Identification of Outliers*. Springer, 1980.

[5] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Medical Research Methodology*, vol. 14, no. 1, p. 135, 2014.

[6] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

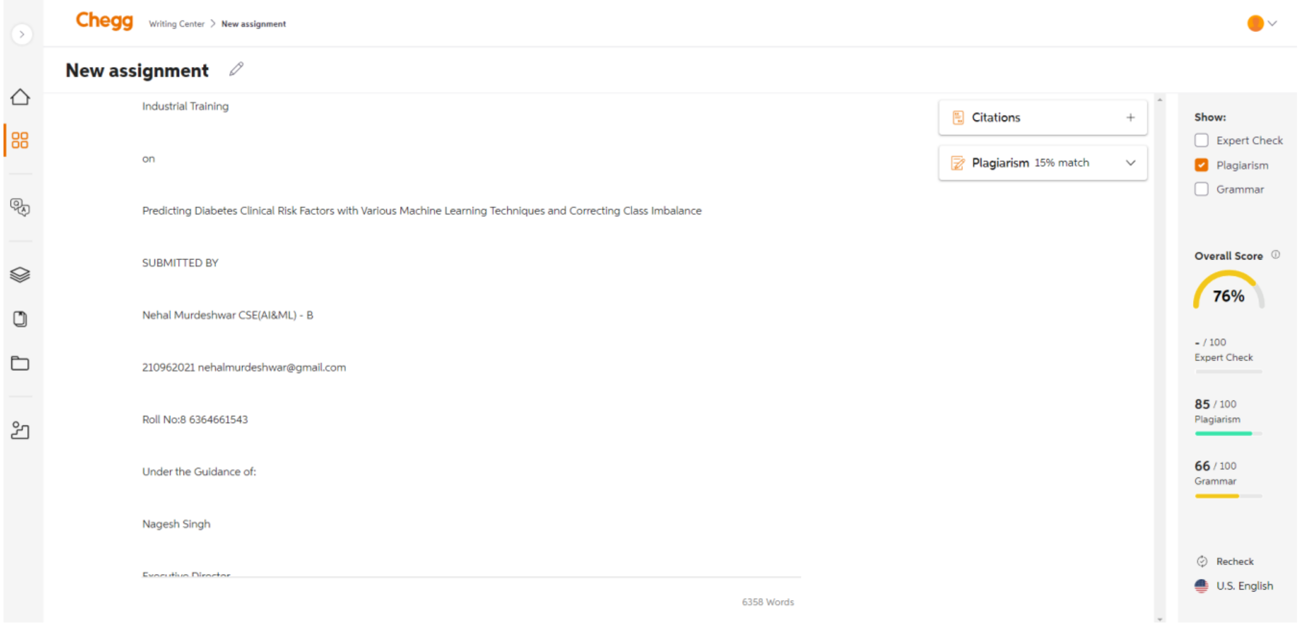
[7] A. Smiley, D. King, J. Harezlak, P. Dinh, and A. Bidulescu, "The association between sleep duration and lipid profiles: the NHANES 2013–2014," *Journal of Diabetes & Metabolic Disorders*, vol. 18, no. 2, pp. 315–322, 2019.

[8] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, 2011.

[9] T. G. Dietterich, "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[10] Centers for Disease Control and Prevention (CDC), "National Health and Nutrition Examination Survey (NHANES)," [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm. [Accessed: 22-Aug-2024].

10. Plagiarism Report



Chegg Writing Center > New assignment

New assignment

Industrial Training

on

Predicting Diabetes Clinical Risk Factors with Various Machine Learning Techniques and Correcting Class Imbalance

SUBMITTED BY

Nehal Murdeshwar CSE(AI&ML) - B

210962021 nehalmurdeshwar@gmail.com

Roll No:8 6364661543

Under the Guidance of:

Nagesh Singh

Executive Director

6358 Words

Citations +

Plagiarism 15% match

Show:

□ Expert Check

☑ Plagiarism

□ Grammar

Overall Score

76%

- / 100
Expert Check

85 / 100
Plagiarism

66 / 100
Grammar

Recheck

U.S. English