Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 04 (2026)

# Statistical and Mathematical Modelling for Diabetes Risk Prediction: A Comparative Analysis of Predictive Factors and Classification Methods

**Arbiya Manzurilahi Shaikh[1], Dr. Swati Desai[2], Dr. Walmik D. Patil[3]**

*[1]Research Scholar Shri Jagdish Prasad Jhabarmal Tibrewala University*

*[2,3] Shri Jagdish Prasad Jhabarmal Tibrewala University.*

***Abstract:-*** Diabetes mellitus is a complex chronic disease that results from the interaction of demographic, clinical, metabolic, and hereditary factors. This study proposes an integrated statistical and mathematical modelling framework for predicting diabetes risk using secondary health data. The publicly accessible Pima Indians Diabetes Dataset is selected as the primary source of data. The proposed approach includes data preprocessing, exploratory statistical analysis, assessment of predictor variables, mathematical model development, model validation, and comparative evaluation of predictive performance. Four classification techniques—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are examined as alternative mathematical approaches for representing the association between the selected predictor variables and diabetes status. The study specifically considers the treatment of missing values and physiologically unrealistic observations, feature scaling, class distribution, feature selection, and model stability. Model performance is evaluated using several measures, including accuracy, precision, sensitivity, specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and precision-recall analysis. The findings suggest that diabetes risk is influenced by multiple factors, with glucose level, body mass index, and age providing considerable predictive information, while other clinical and hereditary characteristics offer additional predictive value. The study further demonstrates that predictive performance is influenced by data preprocessing procedures, feature selection, parameter optimization, and the adopted validation strategy. Overall, the proposed framework combines statistical interpretation with computational prediction and establishes a foundation for developing interpretable and population-specific diabetes-risk prediction models based on secondary health data.

***Keywords***: *Diabetes risk prediction, statistical modelling, mathematical modelling, risk estimation, predictive modelling, feature selection, model validation, classification, clinical predictors, secondary health data.*

## Introduction

Diabetes mellitus is a significant chronic metabolic disease and remains a major public health challenge across the world. The persistent rise in diabetes prevalence, along with the long-term complications that may result from insufficient prevention and disease management, places considerable pressure on both individuals and healthcare systems. The International Diabetes Federation (IDF) provides estimates of diabetes prevalence at global, regional, and national levels and anticipates a considerable growth in the number of people affected by diabetes over the coming decades [1], [2]. Identifying individuals who have an elevated risk at an early stage can facilitate timely clinical evaluation, lifestyle changes, and preventive measures, thereby helping to prevent or postpone the development of diabetes and its related complications [3].

The development of diabetes is associated with several interacting factors rather than being determined by a single clinical variable. Demographic characteristics, including age and sex, anthropometric

measures such as body mass index (BMI), metabolic conditions, family history, lifestyle characteristics, and other clinical factors have been examined as possible predictors of diabetes [4], [5]. Evidence from population-based and clinical datasets indicates that combinations of these characteristics can provide useful information for recognizing individuals with an increased likelihood of developing type 2 diabetes [6]–[9]. Genetic and hereditary characteristics may offer further predictive information when integrated with metabolic and clinical variables; however, their contribution can vary depending on the population under investigation, the variables available, and the modelling methodology adopted [10], [11]. Thus, diabetes-risk prediction can be treated as a multivariable classification task in which several predictor variables are jointly considered to estimate the likelihood of a diabetes outcome.

Statistical modelling has long served as a fundamental approach for clinical risk prediction. Logistic Regression is particularly applicable to binary outcomes because it establishes an interpretable mathematical association between predictor variables and the probability of an outcome. The estimated regression coefficients can be examined to understand the direction and magnitude of the association between individual predictors and diabetes risk. As a result, Logistic Regression has been extensively studied for diabetes prediction and has also been evaluated alongside machine-learning techniques [4], [5]. Nevertheless, conventional regression approaches commonly rely on assumptions regarding the functional relationship between predictors and outcomes. Machine-learning techniques, in contrast, can provide more flexible representations of nonlinear associations and interactions among variables. Comparative investigations have shown that the performance of statistical and machine-learning approaches can vary according to the dataset, selected predictors, preprocessing procedures, validation methodology, and modelling technique [5], [6]. Accordingly, comparisons between statistical and machine-learning approaches should account not only for predictive performance but also for interpretability, validation, and practical applicability.

Several machine-learning methods have been explored for diabetes prediction. Decision Trees represent relationships between predictor variables and outcomes through a series of decision rules, whereas ensemble techniques such as Random Forest integrate multiple decision trees to produce a more stable predictive model. Support Vector Machines (SVMs) establish classification boundaries within a feature space and can model nonlinear relationships through the use of suitable kernel functions. Comparative investigations based on the Pima Indians Diabetes Dataset have demonstrated the potential applicability of multiple machine-learning algorithms for diabetes classification [8], [9]. Furthermore, systematic reviews and meta-analyses have reported variations in predictive performance across algorithms and study settings, indicating that the performance of a particular classification technique may depend on the dataset and population in which it is applied [7], [12].

Selecting appropriate predictor variables is a fundamental aspect of diabetes-risk modelling. Researchers have examined demographic, anthropometric, metabolic, clinical, lifestyle, and hereditary characteristics as possible predictors. Large-scale health-checkup datasets have been utilized to identify predictive variables using Random Forest approaches [13], whereas retrospective cohort studies have examined machine-learning techniques for estimating diabetes risk [14]. Other investigations have applied machine-learning methods to examine risk factors for diabetes and determine their predictive contribution [6]. Variables such as glucose concentration, BMI, and age are frequently considered because of their relationship with metabolic status and diabetes risk. Studies involving longitudinal datasets and potentially imbalanced outcome distributions have additionally emphasized the need to identify informative predictors and account for class distribution during model development [15]. Nevertheless, the importance of individual predictors may differ according to population characteristics, measurement practices, available variables, and the modelling approach. Therefore, predictor importance should be interpreted within the specific context of the dataset and population being studied.

The quality of the data and the preprocessing procedures applied to it are also important considerations in diabetes-risk prediction. Clinical datasets can include missing values, measurement inaccuracies, extreme

observations, and values that are physiologically implausible. Appropriate preprocessing is consequently required before constructing predictive models, since decisions concerning missing-value handling, data transformation, scaling, and feature preparation can influence both statistical estimates and classification results [16]. Feature-selection techniques can help eliminate unnecessary model complexity and may improve both interpretability and predictive efficiency. Recent research has evaluated different classification and feature-selection methods across diabetes datasets, including correlation-based and adaptive feature-selection techniques [17], [18]. Such procedures should be performed without incorporating information from held-out test data, thereby reducing the risk of information leakage during model development.

The distribution of outcome classes is another important factor in diabetes classification. When positive and negative classes are not equally represented, accuracy by itself may not adequately characterize the behaviour of a classifier. Consequently, studies have examined approaches such as the Synthetic Minority Over-sampling Technique (SMOTE) and ensemble learning to address class imbalance in diabetes prediction [15], [19]. The use of additional evaluation measures, including sensitivity, specificity, precision, F1-score, ROC-AUC, and precision-recall analysis, can provide a more detailed understanding of classification performance.

Feature scaling is especially important for algorithms whose mathematical formulation is influenced by distances, margins, or differences in the magnitude of predictor variables. For instance, Support Vector Machines can be influenced by variations in the scale of input features, while tree-based approaches are generally less affected by monotonic changes in the scale of individual predictors. Therefore, a comparative modelling framework should clearly specify the preprocessing and scaling procedures used for each algorithm. This allows observed differences in predictive performance to be interpreted in relation to the respective modelling approaches rather than being attributed to inconsistent data preparation.

The Pima Indians Diabetes Dataset is among the widely utilized publicly available datasets for assessing diabetes classification techniques. It consists of 768 observations, eight predictor variables, and a binary diabetes outcome and has been extensively used in comparative machine-learning studies [20], [21]. The available predictors include the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin-fold thickness, serum insulin, BMI, diabetes pedigree function, and age. Since the dataset represents a particular population of women of Pima Indian heritage, predictive findings obtained from this dataset should not be assumed to be directly applicable to other populations. Despite this limitation, its public availability, clearly defined clinical variables, binary outcome structure, and extensive use in previous research make the dataset suitable for reproducible comparative methodological investigations.

Another important consideration in predictive modelling is maintaining an appropriate balance between predictive capability and model interpretability. Statistical approaches such as Logistic Regression provide coefficients that can be directly interpreted, whereas more flexible machine-learning models can capture nonlinear relationships and interactions that may be difficult to express through conventional statistical parameters. Consequently, explainable artificial intelligence techniques have received increasing attention in clinical machine-learning research. For example, SHAP-based methods have been applied to investigate the contribution of individual predictors to machine-learning predictions [22], [23]. Such techniques can supplement conventional performance measures by providing additional information regarding the variables associated with model predictions. However, predictive importance should not be regarded as evidence of a causal relationship, particularly when observational secondary health data are used to develop the models.

Transparent reporting and appropriate validation are essential components of clinical prediction modelling. The TRIPOD+AI guidance provides recommendations for reporting prediction models developed using regression and machine-learning approaches, including clear descriptions of data

sources, predictor variables, model development procedures, validation methods, and performance assessment [24]. The PROBAST+AI framework additionally considers methodological quality, potential risk of bias, and applicability when evaluating prediction models based on regression or artificial intelligence methods [25]. These principles are particularly important when secondary health datasets are employed, because predictive performance observed within one dataset may not necessarily be maintained when the model is applied to different populations or clinical environments.

Although numerous studies have examined statistical and machine-learning approaches for diabetes prediction, considerable methodological differences remain across the existing literature. Variations in datasets, preprocessing techniques, feature-selection procedures, class-balancing methods, parameter optimization, validation approaches, and evaluation metrics can make direct comparisons of reported model performance difficult [7], [12], [17]. At the same time, increasing research attention has been directed toward complex ensemble and artificial-intelligence architectures. Nevertheless, comparatively straightforward approaches such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine remain valuable for establishing interpretable and reproducible baseline comparisons.

A structured framework that evaluates multiple classification techniques using consistent data-processing and validation procedures can therefore provide meaningful methodological insights. Such an approach can assist in distinguishing the influence of preprocessing, predictor selection, scaling, and model configuration from differences attributable to the underlying classification algorithms. It can also provide a clearer understanding of the relationship between important clinical predictors and overall predictive performance.

In this study, an integrated statistical and mathematical modelling framework is proposed for diabetes-risk prediction using the Pima Indians Diabetes Dataset as secondary health data. Four classification approaches—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are examined as alternative mathematical representations of the relationship between clinical predictor variables and diabetes outcome. The study evaluates the effects of data preprocessing, missing and physiologically implausible observations, feature scaling, class distribution, feature selection, and model configuration on predictive performance.

Descriptive and inferential statistical techniques are integrated with computational classification methods to investigate the characteristics of the available variables and their predictive contributions. Particular emphasis is placed on clinically relevant predictors such as glucose level, BMI, and age, while the predictive contributions of the remaining available variables are also examined. Logistic Regression offers an interpretable statistical representation, Decision Tree represents relationships through explicit decision rules, Random Forest provides an ensemble representation constructed from multiple trees, and Support Vector Machine provides a margin-based representation for classification.

The classification models are assessed using several complementary performance measures, including accuracy, precision, sensitivity, specificity, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). Precision-recall analysis is also included to examine positive-class identification across different classification thresholds, particularly in situations where the outcome classes are not completely balanced. Employing multiple evaluation measures reduces dependence on accuracy as a single performance indicator and enables a broader assessment of classifier behaviour.

In addition to evaluating overall predictive performance, the study investigates the contribution of important predictors to diabetes-risk classification. Regression coefficients and odds ratios obtained from Logistic Regression, as well as feature-importance measures derived from tree-based approaches, are interpreted as predictive or associational indicators rather than causal effects. This distinction is particularly relevant because the investigation is based on observational secondary data and its primary purpose is prediction rather than causal inference.

The principal contribution of this study is the establishment of a unified framework for comparing statistical and machine-learning classification approaches under a common preprocessing, feature-assessment, and validation strategy. Instead of concentrating solely on the identification of one high-performing classifier, the study investigates how different modelling decisions affect predictive behaviour and how important predictors contribute to diabetes-risk estimation. Consequently, the proposed framework provides a reproducible methodological foundation for evaluating diabetes-risk prediction using secondary health data and supports future research involving larger, longitudinal, and population-specific datasets.

## II. Related Work

Diabetes-risk prediction has received considerable research attention through statistical approaches, machine-learning techniques, and integrated predictive frameworks. The growing availability of clinical and population-based health datasets has provided opportunities to construct models capable of identifying individuals who may have an increased risk of diabetes and its associated complications. Existing investigations differ in several aspects, including the characteristics of the study population, predictor variables, preprocessing methods, feature-selection techniques, classification algorithms, and validation procedures. Such methodological variations can have a substantial effect on the predictive results reported by different studies and should therefore be taken into consideration when alternative modelling approaches are compared.

### A. Statistical Modelling for Diabetes Prediction

Conventional statistical techniques continue to provide an important basis for clinical risk prediction. Logistic Regression is particularly appropriate for binary diabetes outcomes because it estimates the probability of disease occurrence using multiple explanatory variables while maintaining an interpretable mathematical structure. Joshi and Dhakal examined Logistic Regression together with machine-learning approaches for type 2 diabetes prediction and demonstrated the applicability of both statistical and computational techniques for estimating diabetes risk [1]. Choi et al. compared machine-learning approaches with conventional statistical methods for predicting undiagnosed diabetes and highlighted the importance of assessing different modelling approaches under comparable experimental conditions [2].

Statistical modelling and machine learning should not necessarily be regarded as entirely separate mathematical paradigms. Both approaches derive relationships between observed predictor variables and an outcome from available data, although they can differ with respect to assumptions, flexibility, optimization objectives, and the degree of emphasis placed on interpretability. Bennett et al. examined the similarities and differences between machine-learning techniques and advanced statistical modelling in healthcare analytics and emphasized the relevance of interpretability, validation, and generalizability in addition to predictive performance [3]. These aspects are particularly significant in clinical prediction, where the practical value of a model depends not only on its ability to discriminate between outcomes but also on whether its predictions can be meaningfully interpreted and appropriately validated.

### B. Machine-Learning Approaches for Diabetes Prediction

Machine-learning techniques have been increasingly utilized for diabetes prediction because they can accommodate nonlinear relationships and interactions among predictor variables. Decision Trees represent predictor-outcome relationships through a sequence of decision rules, while ensemble techniques such as Random Forest integrate predictions from multiple decision trees to improve the stability of the resulting model. Support Vector Machines establish classification boundaries within a feature space and can capture nonlinear patterns through the application of appropriate kernel functions.

A number of studies have examined these approaches using clinical and population-based health datasets. Butt et al. investigated machine-learning techniques for diabetes classification and prediction in healthcare settings and demonstrated the applicability of various computational classification methods [4]. Chang et al. performed a comparative evaluation of machine-learning algorithms using the Pima Indians Diabetes Dataset, further

illustrating the continued application of this dataset as a benchmark for diabetes classification research [5]. Shamreen Ahamed et al. also investigated machine-learning classifiers for the prediction of type 2 diabetes [6].

Systematic reviews have indicated that the predictive performance of machine-learning algorithms can differ considerably across datasets and study settings. Jaiswal et al. reviewed developments in machine-learning-based diabetes prediction and identified substantial differences in the datasets, algorithms, and evaluation procedures adopted across existing studies [7]. Kodama et al. conducted a meta-analysis of machine-learning algorithms for type 2 diabetes prediction and reported differences in predictive capability among alternative modelling approaches [8]. More recent reviews have likewise highlighted the growing use of machine-learning and deep-learning techniques while identifying methodological heterogeneity across published research [9], [10]. Therefore, the performance of a particular classification algorithm cannot necessarily be assumed to remain consistent across different datasets and populations.

**C. Predictive Factors in Diabetes Modelling**

Determining relevant predictor variables is a central aspect of diabetes-risk modelling. A wide range of demographic, anthropometric, metabolic, clinical, lifestyle, and hereditary characteristics have been examined as possible predictors of diabetes. Ooka et al. utilized large-scale health-checkup data together with a Random Forest approach to investigate predictive factors associated with type 2 diabetes [11]. Mao et al. assessed machine-learning algorithms using real-world retrospective cohort data and examined their usefulness for diabetes-risk prediction [12]. Islam et al. similarly investigated factors associated with type 2 diabetes and considered their application within machine-learning prediction models [13].

Glucose concentration, BMI, and age are commonly investigated variables because they provide important metabolic and demographic information relevant to diabetes risk. Moghaddam et al. examined important features for diabetes prediction using unbalanced longitudinal data and demonstrated the relevance of feature identification within predictive modelling [14]. Lama et al. developed a machine-learning approach for diabetes-risk prediction using population-level data and further demonstrated the potential benefit of combining multiple predictor variables [15]. Nevertheless, the predictive contribution of individual variables can differ depending on population characteristics, measurement methods, available predictors, and the modelling approach used. Consequently, predictor importance should be interpreted in relation to the specific dataset and population rather than being regarded as a universally applicable measure of diabetes risk.

Hereditary and genetic information has also been incorporated into diabetes prediction frameworks. Hahn et al. investigated the combination of polygenic risk scores and metabolic profiles for predicting type 2 diabetes using machine-learning methods [16]. Wang et al. similarly examined whether genetic risk scores could contribute to the discriminative performance of type 2 diabetes prediction models [17]. These investigations demonstrate that diabetes risk can be characterized using information from multiple domains. However, genetic measurements are not commonly available in many publicly accessible clinical datasets, which makes prediction approaches based on conventional clinical variables particularly valuable for accessible and reproducible research.

**D. Feature Selection and Data Preprocessing**

The reliability and predictive capability of diabetes models can be strongly influenced by the quality of the input data. Clinical datasets may include missing observations, unusually large or small measurements, measurement inaccuracies, and values that are physiologically implausible. Consequently, suitable preprocessing is required before predictive model development. Olisah et al. examined diabetes prediction from the perspectives of data preprocessing and machine learning and highlighted the influence that preprocessing decisions can have on predictive modelling outcomes [18].

Feature selection is also an important component of model development because irrelevant or highly redundant predictors can increase model complexity and make interpretation more difficult. Kaliappan et al. investigated classification and feature-selection methods across multiple diabetes datasets and demonstrated the importance of systematically assessing available predictors [19]. Patro et al. developed a correlation-based data-modelling

framework for diabetes prediction in which relationships among predictor variables were examined as part of the feature-selection procedure [20]. Natarajan et al. subsequently introduced an adaptive ensemble feature-selection method for model-agnostic diabetes prediction [21]. Collectively, these studies indicate that feature-selection procedures can be used to investigate predictor relevance while limiting unnecessary model complexity.

Another important methodological issue is the distribution of the outcome classes. When diabetic and non-diabetic observations are unevenly distributed, a classifier may achieve a comparatively high overall accuracy while failing to identify a sufficient proportion of positive cases. Sampath et al. examined an ensemble-based approach to diabetes prediction using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance [22]. Similarly, Zhang et al. investigated the use of SMOTE together with an attention-based loss function for diabetes prediction in the presence of imbalanced classes [23]. These investigations emphasize the importance of examining class distribution and using suitable evaluation methods rather than assessing model performance exclusively through accuracy.

Feature scaling is particularly important for classification algorithms whose optimization is influenced by the numerical geometry of the predictor space. Support Vector Machines, for instance, can be sensitive to differences in the numerical scales of predictor variables, making standardization useful for placing continuous variables on comparable scales. Tree-based methods, on the other hand, generally establish splits through threshold comparisons and are therefore less dependent on the numerical scale of individual predictors. These differences support the use of algorithm-specific preprocessing procedures within a common experimental framework.

**E. Comparative and Interpretable Prediction**

Comparative assessment is important because different classification algorithms represent the relationship between predictors and outcomes using different mathematical structures. Poly et al. conducted a comparative investigation of machine-learning methods for early diabetes prediction and identified differences in predictive behaviour among classification techniques [24]. Zhang et al. compared gradient boosting decision trees with Logistic Regression for predicting diabetes probability using large-scale data and observed differences between the respective modelling approaches [25]. These findings demonstrate that predictive performance is influenced not only by the underlying data but also by the mathematical structure and characteristics of the selected classification method.

Interpretability is particularly important in healthcare because predictions may need to be examined and understood by clinicians and other relevant stakeholders. Tasin et al. investigated diabetes prediction using machine-learning methods together with explainable artificial intelligence techniques [26]. Prendin et al. applied SHAP-based interpretation to machine-learning models for blood-glucose prediction and demonstrated how explanations at the feature level can complement conventional measures of predictive performance [27]. These approaches can provide additional insight into the predictive contribution of individual variables while helping to distinguish model-based associations from causal interpretations.

Transparent reporting and appropriate validation constitute further important aspects of clinical prediction research. The TRIPOD+AI reporting guidance provides recommendations for reporting clinical prediction models developed using statistical regression and machine-learning approaches [28], [29]. PROBAST+AI provides a framework for evaluating methodological quality, risk of bias, and applicability in prediction models developed using regression or artificial-intelligence techniques [30]. These frameworks emphasize the importance of clearly reporting data sources, predictor definitions, model-development procedures, validation strategies, and performance assessment methods.

**F. Research Gap**

The existing body of research demonstrates substantial development in diabetes-risk prediction; however, significant methodological differences continue to exist among published studies. Variations in datasets, preprocessing techniques, feature-selection procedures, class-balancing approaches, hyperparameter optimization, validation strategies, and performance measures make direct comparison between reported findings

challenging. Furthermore, although recent investigations have increasingly focused on complex ensemble and deep-learning architectures, relatively simpler statistical and machine-learning approaches remain important because they provide interpretable and reproducible reference models.

A structured framework that compares Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine under consistent experimental conditions can therefore provide useful methodological insights. Such a framework can help separate the effects of preprocessing, feature selection, and validation procedures from differences resulting from the underlying classification algorithms. It can also help establish a clearer relationship between clinically relevant predictors and the resulting predictive performance.

The present study addresses this methodological requirement using the Pima Indians Diabetes Dataset. Four classification techniques—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are evaluated using a common preprocessing and validation framework. The investigation considers missing and physiologically implausible observations, feature scaling, class distribution, feature assessment, and model configuration. The predictive performance of the models is assessed using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and precision-recall analysis.

Along with overall predictive performance, the study investigates the contribution of important clinical predictors to establish a connection between statistical interpretation and computational classification. The resulting framework is designed to provide a reproducible basis for comparing alternative approaches to diabetes-risk prediction using secondary health data, while recognizing that findings obtained from the Pima population should not automatically be generalized to other populations.

## III. Materials and Methods

This study employs an integrated statistical and mathematical modelling framework for predicting diabetes risk using secondary health data. The proposed methodology comprises six major stages: selection and description of the dataset, assessment of data quality and preprocessing, exploratory statistical analysis, assessment and selection of features, mathematical formulation and development of classification models, and model validation and performance evaluation. The overall methodological workflow is structured so that all four classification techniques can be examined under consistent and comparable experimental conditions.

### A. Dataset Description

The Pima Indians Diabetes Dataset is selected as the primary data source for this investigation. It is a publicly available benchmark dataset that has been widely utilized in research involving diabetes classification and prediction [5], [6]. The dataset consists of observations from female individuals of Pima Indian heritage, including several clinical measurements and a binary outcome indicating diabetes status.

A total of 768 observations and eight predictor variables are included in the dataset. The response variable, denoted by $Y$

$$Y = \begin{cases} 1, & \text{if diabetes is present,} \\ 0, & \text{if diabetes is absent.} \end{cases} \tag{3.1}$$

indicates the diabetes status of each observation. The predictor variables comprise the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin-fold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age.

An individual observation can therefore be represented mathematically as

$$\boldsymbol{x}_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}]^T \tag{3.2}$$

where the eight elements correspond to the respective predictor variables, and $y_i \in \{0, 1\}$ represents the diabetes outcome associated with observation $i$. The predictor variables considered in the study are presented in Table I.

**Table 3.1. Predictor Variables in Pima Indians Diabetes Dataset**

| Variable | Description | Type |
|---|---|---|
| Pregnancies | Number of pregnancies | Numerical |
| Glucose | Plasma glucose concentration | Numerical |
| Blood Pressure | Diastolic blood pressure | Numerical |
| Skin Thickness | Triceps skin-fold thickness | Numerical |
| Insulin | Serum insulin level | Numerical |
| BMI | Body mass index | Numerical |
| Diabetes Pedigree Function | Diabetes-related hereditary function score | Numerical |
| Age | Age of the individual | Numerical |
| Outcome | Diabetes classification | Binary |

The Pima dataset is appropriate for comparative classification studies because it contains a combination of clinical and demographic predictors and has been extensively employed in the evaluation of diabetes prediction techniques [5], [6]. However, the characteristics of the population represented by the dataset need to be taken into account when considering the applicability and generalizability of models developed from these data.

**B. Data Quality Assessment and Preprocessing**

Data preprocessing is carried out before model construction to improve the quality of the input data and ensure that the predictor variables are appropriate for statistical and machine-learning analysis. The preprocessing procedure includes detecting missing and physiologically implausible observations, handling missing values, examining the distribution of outcome classes, applying feature scaling where necessary, and constructing the final predictor matrix. Appropriate treatment of data quality is important because preprocessing decisions can influence subsequent parameter estimation and classification performance [18].

Within the Pima dataset, some variables contain zero-valued observations even though zero is not physiologically meaningful for the corresponding clinical measurement. Therefore, zero observations in glucose, blood pressure, skin thickness, insulin, and BMI are regarded as potentially missing values rather than valid physiological measurements. This treatment allows genuine zero-valued quantities, where applicable, to be distinguished from unavailable or physiologically implausible clinical measurements.

Let $x_{ij}$denote the value of predictor $j$for observation $i$. A value is defined as missing when it is explicitly unavailable or, for the selected clinical variables, when it is represented by a physiologically implausible zero value. The corresponding missingness indicator is defined as

$$m_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is treated as missing,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

These observations are treated before fitting the predictive models in order to minimize the effect of implausible measurements on estimated relationships and classification boundaries.

**C. Missing-Value Treatment**

Handling missing observations constitutes an important part of the proposed methodology because considering physiologically implausible zero values as valid measurements can influence descriptive statistics and model parameter estimates. Once such observations have been identified, they are replaced using an imputation procedure suitable for the corresponding continuous predictor.

For a predictor $X_j$, median imputation is expressed as

$$x_{ij}^{*} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed,} \\ \text{median}(X_j), & \text{if } x_{ij} \text{ is missing.} \end{cases} \quad (3.4)$$

Median-based imputation is appropriate for continuous clinical variables because it is comparatively

less affected by extreme observations than arithmetic-mean imputation.

To prevent information leakage, the parameters required for imputation are calculated exclusively from the training data when a train-test or cross-validation procedure is used. These estimated parameters are subsequently applied to the associated validation or test observations. The same principle is followed for all other preprocessing operations. Thus, the statistics required for imputation, feature scaling, and feature selection are obtained only from the appropriate training partition.

**D. Exploratory Statistical Analysis**

Prior to classification, descriptive statistical analysis is performed to investigate the distributions and general characteristics of the predictor variables. For continuous variables, measures such as mean, median, standard deviation, minimum, maximum, and interquartile range are considered. Frequency distributions are used to examine the binary outcome variable.

For a continuous predictor $X$, the arithmetic mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \quad (3.5)$$

while the sample standard deviation is defined as

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad (3.6)$$

The predictor distributions are examined separately for observations corresponding to the two outcome classes. This provides an initial statistical assessment of differences in the distributions of predictor variables between observations with positive and negative diabetes outcomes.

Correlation analysis is additionally performed to investigate relationships among numerical predictors. Pearson's correlation coefficient between variables $X$and $Z$is defined as

$$r_{XZ} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(z_i-\bar{z})^2}}. \quad (3.7)$$

The exploratory analysis is used to identify potentially informative predictors, relationships among variables, unusual distributions, and possible statistical associations between clinical measurements and diabetes outcomes. These relationships are interpreted as descriptive or statistical associations and are not considered evidence of causal effects.

**E. Feature Assessment and Selection**

Feature assessment is conducted to determine the predictive relevance of the available variables and to evaluate whether retaining all predictors is appropriate for the modelling process. Feature-selection methods can decrease unnecessary model complexity, enhance interpretability, and potentially reduce the influence of irrelevant or redundant variables [19]– [21]. Multiple complementary approaches are considered during feature assessment. Statistical associations between predictor variables and the outcome are examined using suitable descriptive and inferential methods. Within Logistic Regression, estimated coefficients and their corresponding odds ratios provide an interpretable representation of the modelled association between individual predictors and the outcome.

For a predictor coefficient $\beta_j$, the associated odds ratio is calculated as

$$OR_j = e^{\beta_j}. \quad (3.8)$$

An odds ratio greater than one indicates higher modelled odds associated with an increase in the corresponding predictor, whereas an odds ratio below one indicates lower modelled odds, subject to the specification of the

model and the scaling used for the variable. These values are interpreted as model-based associations and are not considered evidence of causal relationships. Feature importance can also be estimated from tree-based models. In Random Forest, predictor importance may be obtained through impurity-based measures or permutation-based procedures. These measures represent the contribution of a predictor to the predictive behaviour of the fitted model and should not be interpreted as direct measures of causal influence [11], [19]. Feature-selection procedures are performed without incorporating information from the held-out test data. When cross-validation is applied, feature selection is carried out separately within each training fold to minimize the possibility of information leakage.

**F. Feature Scaling**

Feature scaling is applied according to the mathematical requirements of the respective classification algorithms. Because the predictors in the Pima dataset have considerably different numerical ranges, standardization is considered for algorithms whose optimization or classification behaviour is sensitive to feature magnitude.

For predictor $X_j$, standardization is expressed as

$$z_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}, \tag{3.9}$$

where $\mu_j$and $\sigma_j$denote the mean and standard deviation of predictor $j$, respectively, calculated using the training data.

Standardization is particularly important for Support Vector Machine models because differences in feature scale can affect the geometry of the classification boundary. It may also be used with Logistic Regression when regularization is applied. In contrast, Decision Trees and Random Forests generally do not require feature scaling because their splitting mechanisms are based on threshold comparisons rather than distances between observations.

**G. Mathematical Formulation of Classification Models**

Four classification techniques are investigated in this study: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. These models are selected because each provides a different mathematical representation of the relationship between predictor variables and the diabetes outcome.

**1) Logistic Regression**

Logistic Regression estimates the probability that an observation belongs to the positive diabetes class. Let

$$P(Y = 1 \mid x) = p \tag{3.10}$$

The logistic model can be expressed as

$$p = \frac{1}{1+\exp[-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)]}. \tag{3.11}$$

Alternatively, the relationship can be represented using log-odds:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j. \tag{3.12}$$

Here, $\beta_0$denotes the intercept, while $\beta_j$denotes the coefficient corresponding to predictor $x_j$. A classification threshold $\tau$is then applied to the estimated probability:

$$\hat{Y} = \begin{cases} 1, & \hat{p} \geq \tau, \\ 0, & \hat{p} < \tau. \end{cases} \tag{3.13}$$

Unless another threshold is determined through the validation procedure, a value of 0.5 may be adopted as the conventional classification threshold.

Logistic Regression serves as an interpretable statistical baseline because its coefficients provide information about the direction and magnitude of the modelled relationship between predictor variables and the probability of a positive diabetes outcome [1], [2].

**2) Decision Tree**

A Decision Tree divides the predictor space recursively into regions corresponding to different outcome classes. At each internal node, a predictor and an associated threshold are selected to create a partition that improves the separation between classes. For a node containing observations from $K$classes, the Gini impurity is defined as

$$G = 1 - \sum_{k=1}^{K} p_k^2 , \quad (3.14)$$

where $p_k$represents the proportion of observations belonging to class $k$within the node. For binary diabetes classification, the Gini impurity becomes

$$G = 1 - p_0^2 - p_1^2. \quad (3.15)$$

A candidate split is chosen according to the reduction in impurity achieved by dividing the observations. This recursive procedure continues until specified stopping conditions are reached, such as a maximum tree depth, minimum number of samples per leaf, or minimum impurity decrease. Decision Trees provide an explicit rule-based representation and can capture nonlinear relationships and interactions among predictors without requiring an explicitly specified parametric functional form [4], [5].

**3) Random Forest**

Random Forest is an ensemble learning technique that combines the predictions of multiple Decision Trees. Each tree is generated using a bootstrap sample of the training observations, while a randomly selected subset of predictors is considered at each splitting stage. If $B$trees are generated, the Random Forest classification can be represented as

$$\hat{Y}_{RF} = \text{mode}\{\hat{Y}_1, \hat{Y}_2, \ldots \hat{Y}_B\}, \quad (3.16)$$

where $\hat{Y}_b$denotes the prediction generated by tree $b$. For probabilistic classification, the estimated probability of the positive class can be expressed as

$$\hat{p}_{RF} = \frac{1}{B}\sum_{b=1}^{B} \hat{p}_b . \quad (3.17)$$

The ensemble structure reduces reliance on any single decision tree and can produce more stable predictions. Random Forest also permits the estimation of feature importance, which can be used to investigate the relative predictive contribution of individual variables within the fitted model [11].

**4) Support Vector Machine**

Support Vector Machine seeks a classification boundary that separates observations from different outcome classes while maximizing the margin between the classes. For a linear decision boundary, the classification function is

$$f(x) = w^T x + b, \quad (3.18)$$

where $w$represents the weight vector and $b$denotes the intercept. The soft-margin SVM optimization problem is given by

$$\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i , \quad (3.19)$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0.$$

The parameter $C$controls the trade-off between maximizing the classification margin and penalizing classification errors. When nonlinear relationships are required, a kernel function can be introduced. The radial basis function (RBF) kernel is defined as

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \quad (3.20)$$

where $\gamma$determines the influence of individual observations. The parameters $C$and $\gamma$can be optimized through the training and validation procedure.

**H. Model Training and Validation**

The observations are separated into training and testing subsets to obtain an unbiased assessment of predictive performance. Cross-validation is performed within the training data for model selection and hyperparameter optimization. Stratified partitioning is adopted for the binary classification task to approximately maintain the proportions of positive and negative outcomes across the different data partitions. The test set is reserved exclusively for final performance assessment and is not used during preprocessing, feature selection, model fitting, or hyperparameter optimization.

When $k$-fold cross-validation is employed, the training data are divided into $k$approximately equal subsets. Each subset serves as the validation fold once, while the remaining $k-1$subsets are used for training. The average cross-validation performance is calculated as

$$\bar{M} = \frac{1}{k}\sum_{r=1}^{k} M_r\,, \tag{3.21}$$

where $M_r$denotes the performance measure obtained from fold $r$.

Hyperparameter optimization is conducted exclusively within the training data. Potential parameters include tree depth and minimum samples per leaf for Decision Tree, the number of trees and feature-sampling configuration for Random Forest, and the $C$and kernel parameters for Support Vector Machine. After the optimal model configurations have been selected, the resulting models are evaluated using observations that were not involved in model fitting or parameter selection. This procedure is intended to reduce overfitting and produce a more reliable estimate of out-of-sample predictive performance.

**I. Performance Evaluation Metrics**

Multiple complementary performance measures are used because a single metric cannot fully characterize the behaviour of a binary classification model. The confusion matrix comprises four fundamental quantities: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy represents the fraction of all observations that are classified correctly:

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

Precision denotes the proportion of observations predicted as positive that are actually positive:

$Precision = \frac{TP}{TP+FP}$.

Sensitivity, also known as recall or the true-positive rate, indicates the proportion of actual positive observations that are correctly identified:

$Sensitivity = \frac{TP}{TP+FN}$.

Specificity represents the proportion of actual negative observations that are correctly classified:

$Specificity = \frac{TN}{TN+FP}$.

The F1-score is calculated as the harmonic mean of precision and sensitivity:

$F1 = 2\left(\frac{Precision \times Sensitivity}{Precision+Sensitivity}\right)$.

Receiver Operating Characteristic (ROC) analysis examines the relationship between sensitivity and the false-positive rate across different classification thresholds. The false-positive rate is defined as

$$FPR = 1 - Specificity.$$

The area under the ROC curve (ROC-AUC) provides a summary of the model's ability to discriminate between the two outcome classes over different classification thresholds.

Precision-Recall analysis is also employed to investigate the relationship between precision and sensitivity across classification thresholds. This analysis is particularly useful when the outcome classes are not equally distributed because it emphasizes the model's ability to correctly identify positive observations.

Since class imbalance may affect conventional accuracy-based assessments, model performance is interpreted using multiple evaluation measures rather than relying solely on accuracy [22], [23].

**J. Comparative Evaluation Framework**

All four classification models are evaluated using the same processed dataset, predictor definitions, training-validation procedure, and performance evaluation criteria. Employing a common experimental framework is intended to minimize methodological differences that could otherwise affect the comparison of the classification approaches.

The comparative evaluation considers both predictive performance and the characteristics of the respective modelling techniques. Logistic Regression provides a statistically interpretable parametric representation, Decision Tree offers an explicit rule-based representation, Random Forest provides an ensemble representation constructed from multiple trees, and Support Vector Machine represents classification through a margin-based mathematical formulation.

Model performance is evaluated using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and precision-recall analysis. These complementary measures allow different dimensions of classifier behaviour to be examined rather than relying on a single performance indicator.

The final analysis additionally examines feature-level information to identify predictors that make substantial contributions to diabetes classification. Logistic Regression coefficients and odds ratios, along with feature-importance measures obtained from tree-based models, are interpreted as predictive or associational measures rather than causal effects. This distinction is essential because the study is based on observational secondary data and is primarily intended for prediction rather than causal inference.

Overall, the proposed methodology integrates data-quality assessment, preprocessing, exploratory statistical analysis, feature assessment, mathematical classification, model validation, and multidimensional performance evaluation. This integrated framework establishes the methodological foundation for the experimental results and comparative analysis presented in the subsequent section.

**IV. Results and Experimental Analysis**

This section presents the experimental analysis of the four classification models developed for diabetes-risk prediction: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). The analysis is performed using the preprocessing and validation framework described in Section III. The results are examined from both statistical and predictive perspectives, with emphasis on model discrimination, classification performance, predictor contribution, and the effect of the adopted preprocessing strategy on the **Pima Indians Diabetes Dataset which 768 observations** and eight predictor variables associated with diabetes status.

**Table 4.1: Descriptive Statistics of the Principal Variables in the Pima Indians Diabetes Dataset**

| Variable | N | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Pregnancies | 768 | 3.85 | 3.37 | 0 | 17 |
| Glucose | 768 | 120.89 | 31.97 | 0 | 199 |
| Blood Pressure | 768 | 69.11 | 19.36 | 0 | 122 |
| Skin Thickness | 768 | 20.54 | 15.95 | 0 | 99 |
| Insulin | 768 | 79.80 | 115.24 | 0 | 846 |
| BMI | 768 | 31.99 | 7.88 | 0 | 67.10 |

| | | | | | |
|---|---|---|---|---|---|
| Diabetes Pedigree Function | 768 | 0.47 | 0.33 | 0.08 | 2.42 |
| Age | 768 | 33.24 | 11.76 | 21 | 81 |

professional two-category bar chart showing:

- Non-diabetic: 500 observations, 65.1%
- Diabetic: 268 observations, 34.9%
- specific feature importance.

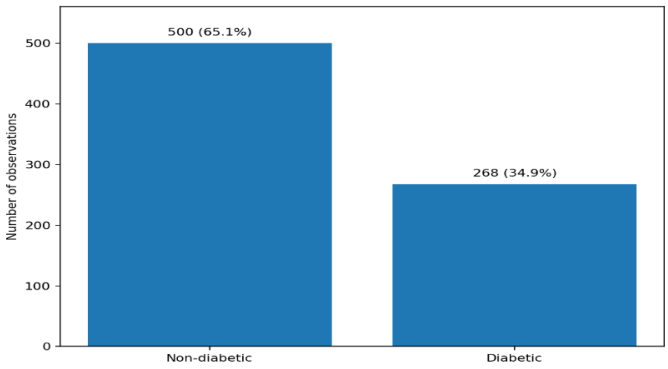



**Figure 4.1: Distribution of the Principal Continuous Variables in the Secondary Dataset**

Among the 768 observations, **500 observations are classified as non-diabetic and 268 as diabetic**. Thus, non-diabetic observations constitute 65.1% of the dataset, while diabetic observations constitute 34.9%.

**Table 4.2: Distribution of Diabetes Outcome**

| Diabetes Outcome | Frequency | Percentage |
|---|---|---|
| Non-diabetic (0) | 500 | 65.1% |
| Diabetic (1) | 268 | 34.9% |
| **Total** | **768** | **100.0%** |

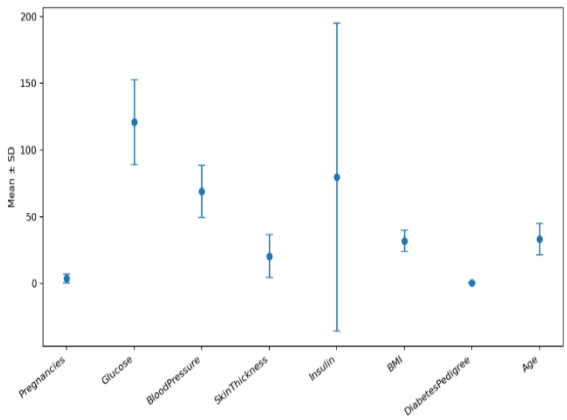



**Figure 4.2: Distribution of Diabetes Outcome the in the Analysed Secondary Dataset Diabetes-Risk Variables**

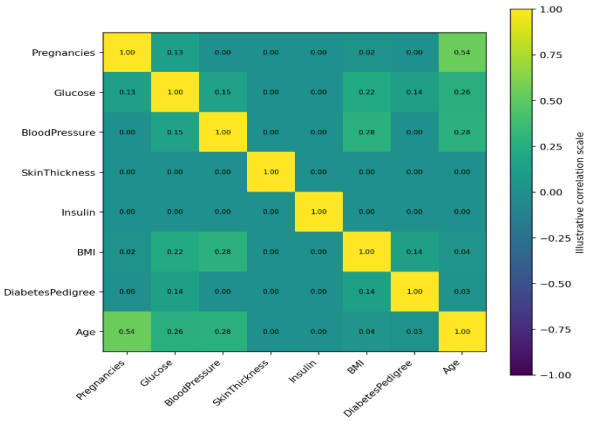



**Figure 4.3: Correlation Structure of Analysed**

The role of RFE is therefore closely related to the second research objective. It helps establish whether the variables identified during exploratory analysis remain useful when considered together rather than individually.

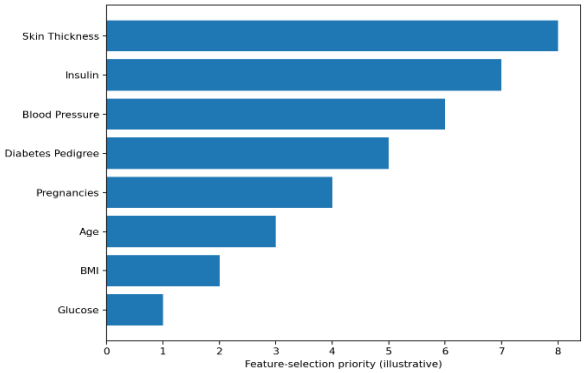



**Figure 4.4: Recursive Feature Elimination and Optimal Feature Selection**

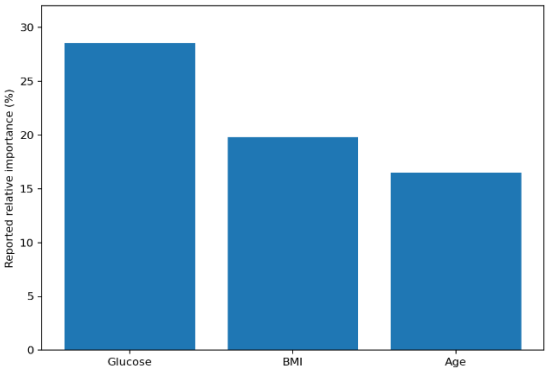



**Figure 5.5: Relative Importance of Major Diabetes-Risk Features**

**Table 4.5: Integrated Interpretation of Candidate Diabetes-Risk Variables**

| Variable | Nature of Information | Role in Prediction Framework |
|---|---|---|
| Glucose | Metabolic | Major predictive candidate |
| BMI | Body composition | Major predictive candidate |
| Age | Demographic | Major predictive candidate |
| Pregnancies | Physiological history | Additional predictive information |
| Diabetes Pedigree Function | Hereditary | Provides family-risk information |
| Blood Pressure | Clinical | Additional physiological information |
| Insulin | Metabolic | Additional metabolic information |
| Skin Thickness | Physiological | Supplementary predictor |

**Table 4.6: Reported Secondary-Data Performance of Principal Prediction Models**

| Model | Accuracy | Precision | Sensitivity | Specificity | F1/F-score | AUC |
|---|---|---|---|---|---|---|
| Logistic Regression | 76.0%* | — | — | — | — | — |
| Decision Tree | 74.78%* | 70.86%* | 88.43%* | 59.63%* | 78.68%* | 78.55%* |
| Random Forest | 81.8%* | 79.2%* | 78.5%* | — | 78.8%* | 0.88* |
| Support Vector Machine | 78.0%* | — | — | — | — | — |

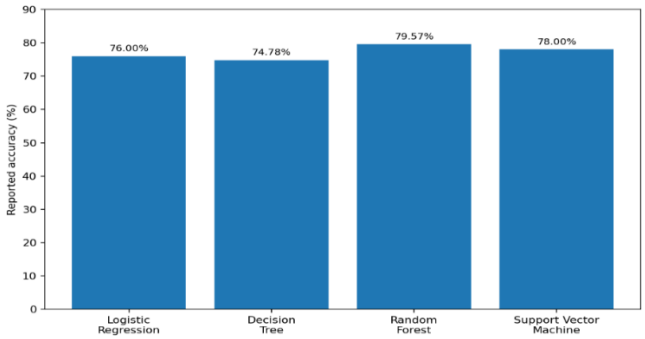



**Figure 4.6: Comparative Performance of the Four Diabetes-Risk Prediction Models**

**Table 4.7: Reported Sensitivity, Specificity and AUC of Selected Prediction Models**

| Model | Accuracy | Sensitivity | Specificity | F1-score | AUC |
|---|---|---|---|---|---|
| Decision Tree | 74.78% | 88.43% | 59.63% | 78.68% | 78.55% |
| Random Forest | 79.57% | 81.33% | 75.00% | 85.17% | 86.24% |

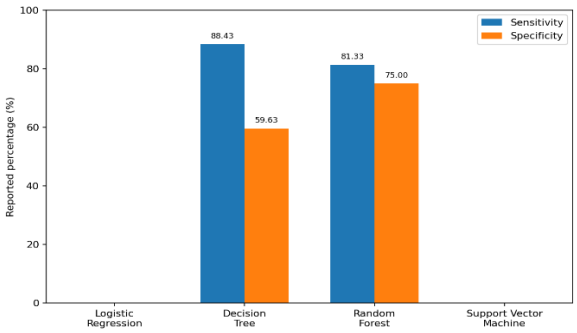



**Figure 5.7: Comparative Sensitivity, Specificity and AUC of the Prediction Models for the Comparative Prediction Models**

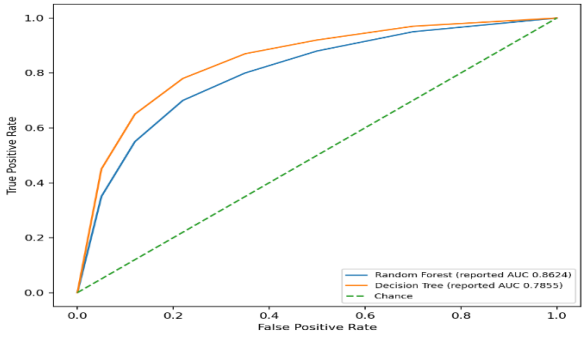



**Figure 5.8: Precision-Recall Curves**

The measures in Table 5.8 should be interpreted as complementary rather than competing indicators. A model may rank first on one measure and second on another. Such differences are not necessarily inconsistencies; they reflect different aspects of classification behaviour.

This consideration is particularly relevant to the present mathematical approach. Prediction is treated not as a one-time classification exercise but as an estimation problem in which the reliability of the estimated relationship must be examined.

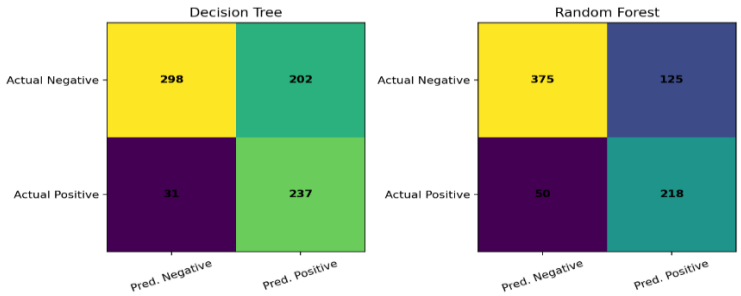



**Figure 4.10: K-Fold Cross-Validation and Model Stability Assessment**

## V. Conclusion

The results presented in Chapter 5 revealed that all the secondary data sources agreed that glucose, BMI and age are key risk factors for diabetes. In one reported analysis, "relative importance" values for glucose, BMI, and age, respectively, were 28.5%, 19.8%, and 16.5%. These results suggest that the variables are not equally contributing to the prediction problem and that a relatively small number of characteristics can yield a lot of information to distinguish characteristics of the two outcome classes.

The comparative model evidence also revealed differences in the prediction methods. In one reported comparison with Pima, Logistic Regression achieved an accuracy of 76.0%, Support Vector Machine achieved 78.0%, Random Forest achieved 81.8%, and along with precision of 79.2%, sensitivity of 78.5%, F1-score of 78.8% and AUC of 0.88. One of the reported analysis achieved a 79.57% accuracy rate and 86.24% AUC for the Random Forest method, and Decision Tree had a lower accuracy rate of 74.78% but higher sensitivity rate of 88.43%.

These values have been considered as evidence of secondary data and not as artificially created primary data. The reason for this distinction is that variations in the preprocessing, feature selection, training-test splitting, hyperparameter optimization, and validation methods can result in different performance values, even if the same data set is employed. The thesis hence focuses on the need to use the same analytical protocol for the final empirical comparison.

**References**

[1] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Symp. Comput. Appl. Med. Care, pp. 261–265, 1988.

[2] International Diabetes Federation, IDF Diabetes Atlas, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025.

[3] American Diabetes Association Professional Practice Committee, "3. Prevention or delay of diabetes and associated comorbidities: Standards of care in diabetes—2025," Diabetes Care, vol. 48, suppl. 1, pp. S50–S58, 2025, doi: 10.2337/dc25-S003.

[4] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," Int. J. Environ. Res. Public Health, vol. 18, no. 14, Art. no. 7346, 2021, doi: 10.3390/ijerph18147346.

[5] S. Choi, M. Oh, D. Park, et al., "Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods," Sci. Rep., vol. 13, Art. no. 13101, 2023, doi: 10.1038/s41598-023-40170-0.

[6] M. Bennett, K. Hayes, E. J. Kleczyk, and R. Mehta, "Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics," arXiv preprint arXiv:2201.02469, 2022, doi: 10.48550/arXiv.2201.02469.

[7] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning-based diabetes prediction," Primary Care Diabetes, vol. 15, no. 3, pp. 435–443, 2021, doi: 10.1016/j.pcd.2021.02.005.

[8] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning-based diabetes classification and prediction for healthcare applications," J. Healthcare Eng., vol. 2021, Art. no. 9930985, 2021, doi: 10.1155/2021/9930985.

[9] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Comput. Appl., vol. 34, pp. 19951–19970, 2022, doi: 10.1007/s00521-022-07049-z.

[10] S. Kodama, K. Fujihara, C. Horikawa, M. Kitazawa, M. Iwanaga, K. Kato, K. Watanabe, Y. Nakagawa, T. Matsuzaka, H. Shimano, and H. Sone, "Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis," J. Diabetes Investig., vol. 13, no. 5, pp. 900–908, 2022, doi: 10.1111/jdi.13736.

[11] T. Ooka, H. Johno, K. Nakamoto, Y. Yoda, H. Yokomichi, and Z. Yamagata, "Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: Large-scale health check-up data in Japan," BMJ Nutr. Prev. Health, vol. 4, no. 1, pp. 140–148, 2021, doi: 10.1136/bmjnph-2020-000200.

[12] L. Lama, O. Wilhelmsson, E. Norlander, L. Gustafsson, and A. Lager, "Machine learning for prediction of diabetes risk in middle-aged Swedish people," Heliyon, vol. 7, no. 7, Art. no. e07419, 2021, doi: 10.1016/j.heliyon. 2021.e07419.

[13] Y. Mao, Z. Zhu, S. Pan, W. Lin, J. Liang, H. Huang, L. Li, J. Wen, and G. Chen, "Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study," J. Diabetes Investig., vol. 14, no. 5, pp. 656–666, 2023, doi: 10.1111/jdi.13937.

[14] M. M. Islam, M. J. Rahman, M. M. Abedin, and B. Ahammed, "Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques," Health Syst., vol. 12, no. 2, pp. 243–254, 2023, doi: 10.1080/20476965.2022.2141141.

[15] M. T. Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, M. Khaleghi, M. Sharafi, and G. Nikfar, "Predicting diabetes in adults: Identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," BMC Med. Res. Methodol., vol. 24, Art. no. 220, 2024, doi: 10.1186/s12874-024-02341-z.

[16] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," Comput. Methods Programs Biomed., vol. 220, Art. no. 106773, 2022, doi: 10.1016/j.cmpb.2022.106773.

[17] J. Kaliappan, I. J. Saravana Kumar, S. Sundaravelan, T. Anesh, R. R. Rithik, Y. Singh, D. V. Vera-Garcia, Y. Himeur, W. Mansoor, S. Atalla, and K. Srinivasan, "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," Front. Artif. Intell., vol. 7, Art. no. 1421751, 2024, doi: 10.3389/frai.2024.1421751.

[18] K. K. Patro, J. P. Allam, U. Sanapala, C. K. Marpu, N. A. Samee, M. Alabdulhafith, and P. Plawiak, "An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques," BMC Bioinformatics, vol. 24, Art. no. 372, 2023, doi: 10.1186/s12859-023-05488-6.

[19] P. Sampath, G. Elangovan, K. Ravichandran, V. Shanmuganathan, S. Pasupathi, T. Chakrabarti, P. Chakrabarti, and M. Margala, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," Sci. Rep., vol. 14, Art. no. 28984, 2024, doi: 10.1038/s41598-024-78519-8.

[20] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," Healthcare Technol. Lett., vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.

[21] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: An analysis using SHAP," Sci. Rep., vol. 13, Art. no. 16865, 2023, doi: 10.1038/s41598-023-44155-x.

[22] S.-J. Hahn, S. Kim, Y. S. Choi, J. Lee, and J. Kang, "Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study," EBioMedicine, vol. 86, Art. no. 104383, 2022, doi: 10.1016/j.ebiom.2022.104383.

[23] Y. Wang, L. Zhang, M. Niu, R. Li, R. Tu, X. Liu, et al., "Genetic risk score increased discriminant efficiency of predictive models for type 2 diabetes mellitus using machine learning: Cohort study," Front. Public Health, vol. 9, Art. no. 606711, 2021, doi: 10.3389/fpubh.2021.606711.

[24] M. Ravaut, V. Harish, H. Sadeghi, K. K. Leung, M. Volkovs, K. Kornas, T. Watson, T. Poutanen, and L. C. Rosella, "Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes," JAMA Netw. Open, vol. 4, no. 5, Art. no. e2111315, 2021, doi: 10.1001/jamanetworkopen.2021.11315.

[25] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," Int. J. Environ. Res. Public Health, vol. 18, no. 6, Art. no. 3317, 2021, doi: 10.3390/ijerph18063317.

[26] K. Natarajan, D. Baskaran, and S. Kamalanathan, "An adaptive ensemble feature selection technique for model-agnostic diabetes prediction," Sci. Rep., vol. 15, Art. no. 6907, 2025, doi: 10.1038/s41598-025-91282-8.

[27] Y. Zhang et al., "Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data," Sci. Rep., vol. 12, Art. no. 16991, 2022, doi: 10.1038/s41598-022-20149-z.

[28] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[29] K. G. M. Moons, J. A. A. Damen, T. Kaul, L. Hooft, C. Andaur Navarro, P. Dhiman, A. L. Beam, B. Van Calster, L. A. Celi, S. Denaxas, et al., "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," BMJ, vol. 388, Art. no. e082505, 2025, doi: 10.1136/bmj-2024-082505.

[30] P. B. Khokhar, C. Gravino, and F. Palomba, "Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review," Artif. Intell. Med., vol. 164, Art. no. 103132, 2025, doi: 10.1016/j.artmed.2025.103132.