# ECE 9603 Assignment 1: Forecasting

1. **Dataset selection:**

* Cardiovascular disease dataset: 70000 samples
* 11 features: id, age (days), gender, height, weight, ap\_hi (systolic BP), ap\_lo (diastolic BP), cholesterol, gluc (glucose), smoke, alco, active, cardio (target)
* Target is binary: 0 (no CVD) and 1 (CVD present)
* <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>
* 13 columns across 70000 rows.
* 11 input features + 1 target variable

Summary: This is a classification problem using cross-sectional forecasting as the data is from different individuals across different times.

1. **Data Exploration and Transformation:**

* This dataset consists of 70,000 samples of the various quantities of an individuals’ overall health. The features included in the dataset are: id, age (days), gender, height, weight, ap\_hi (systolic BP), ap\_lo (diastolic BP), cholesterol, gluc (glucose), smoke, alco, active, cardio (target).

1. **Forecasting Approaches:**

**Logistic Regression:**

* Binary classification algorithm modeling probability using logistic function: P(y=1) = 1/(1 + e^(-wx))
* Learns linear decision boundary by maximizing log-likelihood
* **Why appropriate:** Provides interpretable coefficients showing each feature's contribution to CVD risk, establishing a strong baseline for comparison. Clinically relevant as risk scores often use linear combinations of risk factors.

**Support Vector Machine (SVM):**

* Finds optimal hyperplane maximizing margin between classes
* RBF kernel maps data to higher-dimensional space enabling non-linear decision boundaries
* **Why appropriate:** Effective for complex, non-linear relationships between medical features. Robust to outliers due to support vector focus. Works well with medium-to-large datasets.

**Random Forest:**

* Ensemble of decision trees using bootstrap aggregation and feature randomness
* Each tree votes; majority vote determines classification
* **Why appropriate:** Handles non-linear feature interactions naturally (e.g., age × blood pressure interaction). Robust to feature scaling. Provides feature importance rankings identifying key CVD risk factors. Minimal hyperparameter tuning required for good performance.

• Problem Description (3 points)

o Detailed description of the forecasting problem.

o Include background information and explain relevance.

• Data for Modelling (5 points)

o Describe dataset (size, attributes, ranges, context, quantity).

o Identify which attributes/parts were used.

o Describe transformations and explain their purpose.

• Background (3 points)

o Overview of selected algorithms.

o Explain how they work.

o Justify why they are appropriate for your problem.

• Methodology (4 points)

o Describe how algorithms were applied.

o Note algorithm-specific transformations and why they were required.

o Describe evaluation procedure (hold-out or cross-validation).

• Results (5 points)

o Present results in figures/tables with appropriate metrics.

o Compare algorithms on both transformed and untransformed data.

o Discuss how transformations affected performance and why.

## Problem Description

The forecasting problem being examined is about the diagnosis of CVD (cardio-vascular disease). CVD is the single leading cause of death globally, accounting for 31% of all deaths worldwide (WHO). Means of early detection can help to limit the impact of CVD and increase the ability to make preventative interventions. This is a cross-sectional classification problem in which the goal is to classify individuals as having or not having CVD. This prediction is being made across different patients at what could be viewed as the same point in time for each respective patient. The features of these input patients can be found in table [1].

## Data For Modeling

The dataset being used contains the medical examination results (blood pressure, cholesterol, glucose), and patient-reported lifestyle factors (smoking, alcohol consumption, physical activity) from 70,000 patients. There are 11 different input features and 1 target variable, and they can be found in table [1]. The target variable is a binary decision about whether a given input has CVD or not. This makes this a problem of classification. The class distribution through the data set is perfectly balanced (50% have CVD, 50% do not have CVD). The features have been quantified in the following manner:

* **Demographic Considerations:** Age (years), Gender (1=female, 2=male)
* **Physical Measurements:** Height (cm), Weight (kg), BMI (derived: weight/height^2)
* **Examination Results:** Systolic BP (ap\_hi), Diastolic BP (ap\_lo), cholesterol (1=normal, 2=above normal, 3=well above normal), glucose (1=normal, 2=above normal, 3=well above normal)
* **Lifestyle:** smoking status (binary), alcohol intake (binary), physical activity (binary)

## Background

Logistic regression, SVM (support vector machine), and neural network (multi-layer perceptron) are the algorithms that are going to be applied to the data set. Logistic regression learns a linear decision boundary by maximizing log-likelihood. It is an appropriate algorithm to use in this case because it provides interpretable coefficients showing each individual features’ contribution to CVD risk. This will establish a strong baseline for comparison between the features. The next algorithm applied is SVM which works by finding the optimal hyperplane that maximizes the margin between the two classes. The RBF (radial basis function) kernel maps data to higher-dimensional space enabling non-linear decision boundaries. This approach is appropriate because it is resilient to outliers due to the nature of the support vector. It is also effective for complex, non-linear relationships between medical features. The final approach being used is neural network which work by using backpropagation to adjust weights and minimize classification error. The architecture being used will be 2 hidden layers (64 and 32 neurons) with ReLU activation function being applied. This approach is appropriate because it can learn complex non-linear patterns in large datasets.

## Methodology

## Results

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Number | Actual Feature | Feature Type | Feature Name (var name) | Data Type |
| 1 | Age | Objective | age | Int (days) |
| 2 | Height | Objective | height | Int (cm) |
| 3 | Weight | Objective | weight | Float (kg) |
| 4 | Gender | Objective | gender | Categorical code |
| 5 | Systolic Blood Pressure | Examination | ap\_hi | Int |
| 6 | Diastolic Blood Pressure | Examination | ap\_lo | Int |
| 7 | Cholesterol | Examination | cholesterol | 1: normal  2: above normal 3: well above normal |
| 8 | Glucose | Examination | gluc | 1: normal  2: above normal 3: well above normal |
| 9 | Smoking | Subjective | smoke | Binary |
| 10 | Alcohol Intake | Subjective | alco | Binary |
| 11 | Physical Activity | Subjective | active | Binary |
| 12 | Presence or Absence of Cardiovascular Disease | Target Variable | cardio | Binary |

### **1. Problem Description (3 points)**

"This study addresses cardiovascular disease (CVD) prediction using patient medical examination data. CVD is the leading cause of death globally, accounting for 31% of all deaths worldwide (WHO). Early detection through risk assessment is crucial for preventive interventions. The dataset contains objective medical examination results (blood pressure, cholesterol, glucose) and patient-reported lifestyle factors (smoking, alcohol consumption, physical activity) from 70,000 patients. The goal is to classify patients as having or not having cardiovascular disease."

### **2. Data for Modelling (5 points)**

**Dataset Description:**

* **Size:** 70,000 patient records (after cleaning: ~66,000)
* **Source:** Kaggle cardiovascular disease dataset
* **Features:** 11 input features + 1 target variable
* **Class Distribution:** Perfectly balanced (50% CVD, 50% no CVD)

**Feature Descriptions:**

* **Demographic:** Age (years), Gender (1=female, 2=male)
* **Physical Measurements:** Height (cm), Weight (kg), BMI (derived: weight/height²)
* **Examination Results:** Systolic BP (ap\_hi), Diastolic BP (ap\_lo), Cholesterol (1=normal, 2=above normal, 3=well above normal), Glucose (1=normal, 2=above normal, 3=well above normal)
* **Lifestyle:** Smoking status (binary), Alcohol intake (binary), Physical activity (binary)

**Data Quality:**

* No missing values in original dataset
* Removed outliers outside physiologically plausible ranges (e.g., BP > 220 or < 80)
* Final dataset: [insert your number] samples

**Transformations Applied:**

Standardization (Z-score normalization):

* **Purpose:** Centers features at mean=0 with standard deviation=1
* **Reasoning:** Blood pressure ranges ~80-180 mmHg while age ranges ~30-65 years. Without standardization, features with larger numerical ranges dominate distance-based algorithms (k-NN, SVM) and gradient-based optimization (neural networks). Standardization ensures all features contribute equally to model training.
* **Application:** Applied to Logistic Regression, SVM, and Neural Network

Min-Max Scaling:

* **Purpose:** Scales features to [0, 1] range
* **Reasoning:** Neural networks with sigmoid or tanh activation functions perform better with input values in a bounded range. Min-Max scaling also prevents numerical instability in gradient computations.
* **Application:** Applied specifically to Neural Network

Feature Engineering:

* **BMI Calculation:** Created BMI = weight/(height/100)²
* **Reasoning:** BMI is a standard clinical measure combining height and weight into a single obesity indicator, providing more interpretable and predictive information than raw measurements alone.

**Attributes Selected:** All features were retained as each provides unique clinical information relevant to cardiovascular disease prediction. Age, blood pressure, and cholesterol are established cardiovascular risk factors, while lifestyle behaviors (smoking, alcohol, activity) are modifiable risk factors.

### **3. Background (3 points)**

**Logistic Regression:**

* Binary classification algorithm modeling probability using logistic function: P(y=1) = 1/(1 + e^(-wx))
* Learns linear decision boundary by maximizing log-likelihood
* **Why appropriate:** Provides interpretable coefficients showing each feature's contribution to CVD risk, establishing a strong baseline for comparison. Clinically relevant as risk scores often use linear combinations of risk factors.

**Support Vector Machine (SVM):**

* Finds optimal hyperplane maximizing margin between classes
* RBF kernel maps data to higher-dimensional space enabling non-linear decision boundaries
* **Why appropriate:** Effective for complex, non-linear relationships between medical features. Robust to outliers due to support vector focus. Works well with medium-to-large datasets.

**Random Forest:**

* Ensemble of decision trees using bootstrap aggregation and feature randomness
* Each tree votes; majority vote determines classification
* **Why appropriate:** Handles non-linear feature interactions naturally (e.g., age × blood pressure interaction). Robust to feature scaling. Provides feature importance rankings identifying key CVD risk factors. Minimal hyperparameter tuning required for good performance.

**Neural Network (Multi-Layer Perceptron):**

* Layers of interconnected neurons with non-linear activation functions
* Backpropagation adjusts weights to minimize classification error
* **Architecture:** 2 hidden layers (64 and 32 neurons) with ReLU activation
* **Why appropriate:** Can learn complex non-linear patterns in large datasets (70,000 samples sufficient for training). Automatically discovers feature interactions without manual feature engineering.

### **4. Methodology (4 points)**

**Data Preprocessing:**

1. Outlier removal based on physiological plausibility
2. Feature engineering (BMI calculation, age conversion to years)
3. Train-test split (80/20) with stratification to maintain class balance

**Transformation Strategy:**

* **Logistic Regression:** Tested on both original and standardized data to assess sensitivity to feature scaling
* **SVM:** Applied standardization (required for RBF kernel to function properly with mixed-scale features)
* **Random Forest:** Tested on both original and standardized data (tree-based methods theoretically scale-invariant)
* **Neural Network:** Applied Min-Max scaling to [0,1] range (optimal for ReLU activation and gradient stability)

**Model Training:**

* All models trained using scikit-learn with default parameters (Assignment 1 focus on comparison, not optimization)
* Random state=42 for reproducibility
* Neural Network: 500 max epochs with early stopping to prevent overfitting

**Evaluation Procedure:**

* **Hold-out validation:** 80% training, 20% testing (stratified split)
* **Metrics:** Accuracy, Precision, Recall, F1-Score, AUC-ROC
* **Rationale:** AUC-ROC chosen as primary metric as it's threshold-independent and clinically relevant for risk assessment tools. Balanced dataset allows accuracy as secondary metric.

### **5. Results (5 points)**

[Insert your actual results table and figures here]

**Model Performance Comparison:**

| **Model** | **Transformation** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| --- | --- | --- | --- | --- | --- | --- |
| Logistic Regression | Original | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| Logistic Regression | Standardized | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| SVM | Original | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| SVM | Standardized | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| Random Forest | Original | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| Random Forest | Standardized | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |
| Neural Network | MinMax | 0.xxx | 0.xxx | 0.xxx | 0.xxx | 0.xxx |

**Key Findings:**

Impact of Transformations:

1. **SVM showed dramatic improvement** with standardization (expected performance increase of 5-15% in AUC-ROC)
   * **Original data:** [your result] AUC-ROC
   * **Standardized data:** [your result] AUC-ROC
   * **Explanation:** RBF kernel calculates distances between samples. Without standardization, blood pressure (range ~100) dominates over binary features (0-1), causing the kernel to focus on BP while ignoring other features. Standardization equalizes all feature contributions, enabling the kernel to capture complex multi-feature patterns.
2. **Logistic Regression showed modest improvement** with standardization
   * **Explanation:** While logistic regression is theoretically scale-invariant (coefficients adjust to scale), standardization improves numerical stability and convergence speed during optimization. The performance gain reflects more stable gradient descent.
3. **Random Forest showed minimal/no change** between original and standardized
   * **Explanation:** Tree-based algorithms make split decisions based on feature thresholds, not distances or magnitudes. A split on "age > 50" is equivalent whether age is in years, months, or standardized units. This confirms the theoretical scale-invariance of tree-based methods.
4. **Neural Network performed well** with Min-Max scaling
   * **Explanation:** Scaling to [0,1] range kept activations in the effective range of ReLU functions, preventing vanishing/exploding gradients. This enabled stable learning of non-linear patterns.

**Best Performing Model:** [Identify your best model, likely SVM or Random Forest with ~72-75% AUC-ROC]

**Feature Importance Analysis (from Random Forest):** Top 3 most important features: [list them - likely age, systolic BP, cholesterol]

* These align with established cardiovascular risk factors in medical literature, validating model interpretability.

**Clinical Implications:** The models achieve [XX%] accuracy, suggesting potential as screening tools. However, [XX]% false negative rate indicates need for combining with clinical judgment rather than standalone diagnosis.