UNVEILING THE POWER OF RANDOM SUBSPACE SELECTION : A COMPREHENSIVE EVALUATION

A REPORT
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF
BACHELOR OF TECHNOLOGY
IN
DEPARTMENT OF INFORMATION TECHNOLOGY

Submitted by:
Abhishek Bansal (2020UIN3309)
Aakriti Gupta (2020UIN3317)
Sidharth Jain (2020UIN3360)

Under the supervision of
Dr. Ankita Bansal

DEPARTMENT OF INFORMATION TECHNOLOGY
NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY
May, 2024

DECLARATION

Department of Information technology

Delhi-110078, India

We, Abhishek Bansal (2020UIN3309), Aakriti Gupta (2020UIN3317), Sidharth Jain (2020UIN3360), of B. Tech., Department of Information Technology, hereby declare that the Project II -Thesis titled "Unveiling the Power of Random Subspace Selection: A Comprehensive Exploration " which is submitted by us to the Department of Information Technology, Netaji Subhas University of Technology, in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology, is original and not copied from source without proper citation. This work has not previously formed the basis for the award of any Degree.

Place: Delhi

Date: 6th May 2024	Abhishek Bansal	Aakriti Gupta	Sidharth Jain
	2020UIN3309	2020UIN3317	2020UIN3360

CERTIFICATE

Department of Information technology

Delhi-110078, India

This is to certify that the work embodied in the Project II-Thesis titled "Unveiling the Power of Random Subspace Selection: A Comprehensive Exploration " has been completed by Abhishek Bansal (2020UIN3309), Aakriti Gupta (2020UIN3317), Sidharth Jain (2020UIN3360) of B.Tech., Department of Information Technology, under my guidance towards fulfillment of the requirements for the award of the degree of Bachelor of Technology. This work has not been submitted for any other diploma or degree of any University.

Place: Delhi

Date: 6th May 2024 Dr. Ankita Bansal

ABSTRACT

Random Subspace Selection (RSS) is applied to various ML models, including Bayesian models, tree-based models, non-linear models, rule-based models, ensemble models, and linear models. By conducting a comprehensive analysis across multiple datasets, we aim to elucidate the benefits of RSS in improving model performance and its generalization capabilities. The motivation behind our project stems from the ubiquitous challenges faced in machine learning, including overfitting, high-dimensional data, and the need for robust and generalizable models. Traditional modeling approaches often struggle to handle these challenges effectively, leading to suboptimal performance and limited generalization. Random Subspace Selection offers a promising solution by introducing diversity within model ensembles, thereby mitigating overfitting and enhancing model robustness. Moreover, the versatility of RSS in complementing various base models presents an opportunity to explore its efficacy across different modeling paradigms and also various domains of datasets. By systematically evaluating RSS-enabled models and comparing their performance against traditional models, we aim to provide insights and guidance for practitioners seeking to leverage RSS as a powerful tool for improving model robustness and generalization capabilities in their machine learning endeavors.

INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 PROJECT OVERVIEW

In the pursuit of enhancing predictive modeling performance across diverse domains, our project delves into the efficacy of Random Subspace Selection (RSS) when applied to various base models, including Bayesian models, tree-based models, non-linear models, rule-based models, ensemble models, and linear models. By conducting a comprehensive analysis across multiple datasets, we aim to elucidate the benefits of RSS in improving model performance and generalization capabilities.

The motivation behind our project stems from the ubiquitous challenges faced in machine learning, including overfitting, high-dimensional data, and the need for robust and generalizable models. Traditional modeling approaches often struggle to handle these challenges effectively, leading to suboptimal performance and limited generalization. Random Subspace Selection offers a promising solution by introducing diversity within model ensembles, thereby mitigating overfitting and enhancing model robustness.

Moreover, the versatility of RSS in complementing various base models presents an opportunity to explore its efficacy across different modeling paradigms. By leveraging RSS in conjunction with Bayesian models, tree-based models, non-linear models, rule-based models, ensemble models, and linear models, we seek to elucidate its impact on model performance across a wide spectrum of domains and datasets.

Our project encompasses several key components:

1. Dataset Selection: We curated a diverse set of datasets spanning different domains, including finance, healthcare, retail, and more. Each dataset poses unique challenges and characteristics, providing a comprehensive testbed for evaluating the efficacy of RSS across diverse scenarios.

2. Model Selection: We employed a range of base models, each representing distinct modeling paradigms, including Bayesian models, tree-based models (such as decision trees, random forests, and gradient boosting), non-linear models (such as K-nearest neighbors and multilayer perceptrons), rule-based models, ensemble models (such as bagging and boosting), and linear models (such as logistic regression and linear SVM).

3. Implementation of Random Subspace Selection: For each base model, we integrated RSS into the modeling pipeline, leveraging its ability to create diverse subsets of features to enhance model generalization.

4. Performance Evaluation: We systematically evaluated the performance of RSS-enabled models across multiple metrics, including AUC, F1 Score, and Accuracy. By comparing the performance of RSS-enabled models against their counterparts without RSS, we aimed to quantify the impact of RSS on model performance.

5. Analysis and Interpretation: Through thorough analysis and interpretation of experimental results, we sought to elucidate the benefits of RSS across different datasets and model types. Additionally, we aimed to identify scenarios where RSS demonstrates the most significant improvements and areas where further optimization may be warranted.

In conclusion, our project endeavors to shed light on the efficacy of Random Subspace Selection in enhancing predictive modeling performance across diverse datasets and model types. By systematically evaluating RSS-enabled models and comparing their performance against traditional models, we aim to provide insights and guidance for practitioners seeking to leverage RSS as a powerful tool for improving model robustness and generalization capabilities in their machine learning endeavors.

1.2 PROBLEM STATEMENT

High dimensionality in machine learning complicates model training, as it increases the risk of overfitting and hampers model generalization. In the era of big data, where the volume, velocity, and variety of data continue to proliferate, the need for robust and scalable machine learning techniques has never been more pressing. Traditional approaches often falter in the face of high-dimensional data, succumbing to the curse of dimensionality and struggling to generalize effectively. Moreover, the pervasive issue of overfitting looms large in the landscape of machine learning, particularly in scenarios where the complexity of models outstrips the size and quality of training data.

1.3 APPROACH

1.3.1 Enhancing Predictive Performance through Ensemble Learning

Ensemble learning has emerged as a powerful paradigm in machine learning, leveraging the collective intelligence of diverse models to achieve superior predictive performance compared to individual algorithms.The success of ensemble learning techniques mainly relies on the accuracy and diversity of the base learners. The

diversity can help each procedure to guarantee a total good machine learning: diversity of the training data ensures that the training data can provide more discriminative information for the model, diversity of the learned model (diversity in parameters of each model or diversity among different base models) makes each parameter/model capture unique or complement information and the diversity in inference can provide multiple choices each of which corresponds to a specific plausible local optimal result. Further there are 2 kinds of ensembles. Heterogeneous ensemble consists of members having different base learning algorithms such as SVM, ANN and Decision Trees. Homogeneous ensemble methods, use the same feature selection method with different training data and distribute the dataset over several nodes.

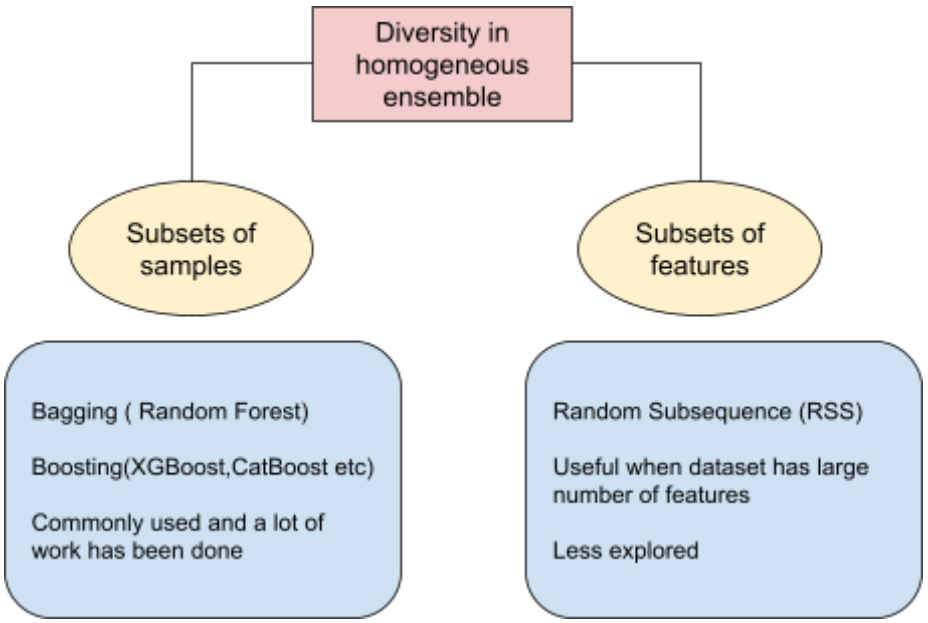

Figure 1.1: Diversity in Homogeneous Ensemble Models

1.3.2 Random Subspace and its advantages

Random Subspace Selection (RSS) stands as a versatile and powerful technique in the realm of machine learning, offering a range of advantages that contribute to enhanced model performance and generalization capabilities. Across diverse domains and applications, RSS has proven to be instrumental in addressing key challenges such as high dimensionality, overfitting, and limited model diversity. In this discussion, we delve into the advantages of RSS, elucidating its impact on model robustness, efficiency, and generalization.

While Random Subspace shares similarities with homogeneous ensemble methods like bagging, it differs in how it introduces diversity among the base classifiers. In bagging, for example, each base classifier is trained on a random subset of the training

data, whereas in Random Subspace, the randomness is introduced by selecting random subsets of features. Therefore, Random Subspace can be considered a form of homogeneous ensemble method because it combines multiple classifiers of the same type (homogeneous base classifiers).

One of the primary advantages of RSS lies in its ability to address the challenge of high dimensionality inherent in many machine learning tasks. High dimensionality refers to datasets with a large number of features, which can lead to computational inefficiencies, increased risk of overfitting, and difficulties in model interpretation. RSS tackles this problem by randomly selecting subsets of features for model training, effectively reducing the dimensionality of the feature space. By incorporating only a fraction of the available features in each model iteration, RSS streamlines the training process and mitigates the curse of dimensionality, enabling more efficient and effective model learning.

Furthermore, RSS serves as a potent tool for mitigating overfitting, a common concern in machine learning where models memorize the training data instead of learning generalizable patterns. Overfitting occurs when a model captures noise or idiosyncrasies in the training data, leading to poor performance on unseen data. By constraining each model to operate within a distinct subspace of features, RSS promotes diversity within the ensemble, reducing the likelihood of overfitting. The introduction of randomness in feature selection encourages models to focus on relevant patterns in the data, leading to improved generalization capabilities and robust predictions.

Moreover, RSS fosters model diversity within ensemble learning frameworks, which is crucial for improving prediction accuracy and reliability. Ensemble methods combine multiple models to make predictions, leveraging the collective wisdom of diverse models to achieve superior performance. By creating diverse subsets of features for each model in the ensemble, RSS encourages models to explore different facets of the data space, capturing complementary patterns and reducing the risk of model bias. This diversity enhances the ensemble's ability to generalize well to unseen data and adapt to changing data distributions, resulting in more robust and reliable predictions.

Another advantage of RSS lies in its flexibility and compatibility with various machine learning algorithms and model architectures. RSS can be seamlessly integrated into both traditional machine learning models and modern deep learning frameworks, making it applicable across a wide range of domains and tasks. Whether applied to decision trees, neural networks, or other machine learning algorithms, RSS consistently demonstrates its effectiveness in enhancing model performance and generalization capabilities.

In conclusion, Random Subspace Selection (RSS) offers a myriad of advantages that contribute to improved model performance, efficiency, and generalization in machine learning tasks. By addressing key challenges such as high dimensionality and overfitting, RSS enables more efficient model training, enhances model robustness, and fosters diversity within ensemble learning frameworks. Its flexibility and compatibility with various machine learning algorithms make it a valuable tool for practitioners across diverse domains, empowering them to build more accurate, reliable, and generalizable machine learning models.

CHAPTER 2 REQUIREMENT ANALYSIS

2.1 BACKGROUND

2.1.1 Literature Survey

Table 2.1 : Literature Survey

S.No.	Method Used	Performance Metrics	Dataset Used	Year
1.	ANN, SVM, Random Forest, Random subspace Dagging	AUC=87.3% RMSE=19%	historical data sources, fieldwork, perception of local residents, and Google Earth	2021[1]
2.	Ensemble of DT, SVM and NN	100% accuracy over BSE and NYSE datasets, but 85.7% and 93.14% over JSE and GSE datasets.	Ghana stock exchange (GSE), the Johannesburg stock exchange (JSE), the New York Stock Exchange (NYSE) and Bombay Stock Exchange (BSE-SENSEX)	2020[2]
3.	DistilBERT, Logistic Regression, Ensemble Learning model of Multilayer perceptron, XGBoost and Random Forest.	Accuracy=79% Recall=79% Precision=81% F1-score=79%	“all the news” dataset	2022 [3]
4.	extra trees, random forest and logistic regression, XGBoost	Accuracy=99.88% specificity=99.99% Sensitivity=98.72%	5644 patients admitted to the Albert Einstein Israelita Hospital in Saulo Paulo	2020 [4]
5.	CNN, RNN	Accuracy=99.96% Precision=93.27% F1-score=96.9%	WISDM, PAMAP2, KU-HAR	2021 [5]
6.	Random Subspace, Forest Tree, LR, LMT, ADT	Accuracy=93.3% AUC= 88.9%	Not Given	2020[6]

Table 2.1(continued)

7.	SVM, Random Subspace	Precision=97.3% Recall=97% F-Measure=96.9% ROC area=99.7%	PQE signals by Discrete Wavelet Transform (DWT) analysis	2022 [7]
8.	REPT, Random Forest, Random subspace	MAE = 1.728 RMSE = 2.725	Unified Soil Classification System	2021[8]
9.	KNN, Naive Bayes, Random Subspace	Accuracy = 83.5%	Armstrong-V1 Bredel Armstrong-V2	2023[9]
10.	Random Subspace, Support Vector Machine	Accuracy = 9.41 Standard Deviation = 1.21	bank marketing, car evaluation database, human activity recognition using smartphone	2023[10]
11.	Random Subspace, Support Vector Machine, Linear Discriminant Analysis	Accuracy = 80.8%	fNIRS-BCI	2020[11]
12.	Random Subspace, Learn++.MF	Accuracy = 85%	Wisconsin Breast Cancer and Wine databases	2020[12]
13.	Random Forest, Decision Tree, Local feature Sampling	Accuracy= 95%	MNIST, ISOLET, HAR, UJIndoorLoc	2019[13]
14.	Random Forest, Support Vector Machines,	Accuracy: 89%	Chemistry data for soil samples	2023[14]
15.	PSRSM, DeepCNF	Accuracy = 84%	ASTRAL, CULlPDB, CASP10, CASP11	2005[15]

Table 2.1(continued)

16.	Random Subspace, Bagging, Support Vector Machine, WFAIB_RS	AUC=86%	Financial data of 1726 companies listed in China	2021[16]
17.	AdaBoost, Random Subspace, Random Forest	Accuracy = 76%	Medical datasets such as diabetes, blood pressure	2014[17]
18.	RSS-REPT, RSS-ET, Random Subspace	Correlation coefficient= 0.968	Van Don - Mong Cai expressway project, in the Quang Ninh province of Vietnam	2021[18]
19.	Ridge Regression, Random Subspace	Accuracy = 84%	FRED-MD	2017[19]
20.	Support Vector Machine, Random Subspace, Ensemble Models, DWT	Precision (0.973), Recall (0.970) F-Measure (0.969)	solar PV microgrid dataset	2022[20]

2.2 RESEARCH ARGUES

2.2.1 Impact of RSS across various model classes

This study endeavors to investigate the impact of the Random Subset Selection (RSS) method on the performance of diverse classes of machine learning models. A deliberate analysis is conducted to discern the differences in performance metrics before and after the application of RSS. The primary objective is to identify the machine learning model that exhibits the most notable enhancement in performance due to the utilization of the RSS technique.

The research methodology involves a thorough examination of the magnitude of improvement in the area under the receiver operating characteristic curve (AUC) across various classes of machine learning models. A comparative analysis is

conducted to assess the efficacy of RSS in augmenting model performance, with a particular focus on contrasting the pre-RSS and post-RSS performance metrics.

The study employs rigorous statistical techniques and experimental design principles to ensure the validity and reliability of the findings. Through systematic experimentation and analysis, the research aims to provide valuable insights into the effectiveness of the RSS method in enhancing the predictive performance of machine learning models across different categories.

Furthermore, the study aims to contribute to the existing body of knowledge by identifying the machine learning model that benefits most significantly from the incorporation of the RSS technique. By elucidating the factors contributing to performance improvement and delineating the relative efficacy of RSS across diverse model classes, the research seeks to inform and guide future endeavors in machine learning model development and optimization.

2.2.2 Impact of RSS across various dataset domains

The second objective of this research project is to test the “ubiquitousness” of the Random Subspace Method across datasets belonging to different categories or industries.Through this, the paper aims to establish the level of generalizability of the Random Subspace method when applied to Machine Learning models in different domains and also ensure that the results are not due to certain features present in a single dataset but the model and its results are more generalized.By doing so, the study seeks to ascertain the extent of applicability of the Random Subspace method when integrated into Machine Learning models across varied domains. Moreover, the objective is to mitigate the potential influence of specific dataset features on the observed outcomes, thereby ensuring that the efficacy of the model and its results remain broad and generalizable.

2.3 SOLUTIONS

2.3.1 Comparing various ML classes models

The different classes of models considered for this research project are - Bayesian Models (Bayesian Net and Naive Bayes), Tree based models (Decision tree and Hoeffding tree), Non-Linear based models(KNN and MLP), Rule based models (JRip and OneR), Linear(SVM and Logistic regression) and Ensemble models (Bagging,

Boosting and Random Forest). The models' performances are compared on the basis of three parameters - F-1 Score, Area Under Curve(AUC), and Accuracy.

2.3.2 Assorted dataset domains

In this study, we explore several domains, including e-commerce shopping, banking, stock market, health, and hospitality, with a focus on consumer behavior, bankruptcy prediction, stock price prediction, and heart failure prediction, respectively. These domains were selected based on their perceived prevalence and relevance in everyday life, as they represent sectors frequently utilized by the general public. This research endeavors to investigate and analyze the predictive capabilities of machine learning algorithms within these domains, aiming to contribute to the advancement of predictive modeling techniques and their applicability in real-world scenarios.

CHAPTER 3 METHODOLOGY

3.1 DATASETS

3.1.1 Sourcing Parameters

In the pursuit of constructing robust and generalizable predictive models, the selection criteria for datasets is guided by several methodological considerations. Primarily, datasets are sought that exhibit a substantive breadth in the number of independent variables, with a preference for those possessing a considerable magnitude of at least 20 features. This criterion is predicated on the understanding that datasets characterized by a larger feature space tend to offer greater complexity and potential for capturing intricate relationships within the data.

Moreover, emphasis is placed on the nature of the dependent variable, whereby datasets with discrete or categorical outcomes are prioritized. This preference is rooted in the recognition that discrete target variables facilitate the application of various machine learning algorithms designed for classification tasks, thus affording a diverse array of modeling approaches.

Furthermore, the selection process endeavors to encompass datasets spanning varied domains and subject matter areas. This strategic diversification aims to mitigate the risk of overfitting to specific contexts or biases inherent within singular datasets. By encompassing a range of domains, such as healthcare, finance, social sciences, and beyond, the resultant models are more likely to exhibit robustness and generalizability across disparate real-world scenarios.

3.1.2 Datasets explored

3.1.2.1 Apple stock Prediction :

Apple stock prices from years 2014 to 2023. This dataset can be used to predict the price trend for the next day based on technical indicators.The target is to predict the price trend for next day.The dataset was obtained from Kaggle. It has multi class classification as follows

- bullish - If price increases more than 0.5%
- bearish - If price fall more than 0.5%

- neutral - If price movement stay with -0.5% to +0.5%

3.1.2.2 Shopper purchasing intention :

The dataset consists of feature vectors belonging to 12,330 sessions.The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. It predicts whether the shopper based on its characteristics would or not buy the product. The 'Revenue' attribute is used as the class label.The dataset was obtained from Kaggle.

3.1.2.3 Companies Bankruptcy Prediction:

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS,), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012. The data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 represents bankrupted companies, 6756 firms that did not bankrupt.The dataset was obtained from Data world.

3.1.2.4 Heart Attack Prediction :

Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease).This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes.The dataset was obtained from Kaggle.

3.1.3 Datasets characteristics

Table 3.1: Dataset Characteristics

Name	No. of Features	No. of instances	Link
Apple stock Prediction	20	2500	https://www.kaggle.com/datasets/aspillai/apple-stock-price-prediction-10-years
Shopper purchasing intention	18	12330	https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset/data
Companies Bankruptcy Prediction	64	7027	https://data.world/uci/polish-companies-bankruptcy-data
Heart Attack Prediction	12	1190	https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

3.2 BASE MODELS

3.2.1 Selection Motivation

To cater to the threat of generalizability we have taken various types of ML models across popular different classes to perform a detailed comparison as to which models are most compatible with RSS.This research endeavor involves an exhaustive examination of ML models spanning multiple categories, each renowned for its distinct characteristics and methodologies.

Through meticulous experimentation and evaluation, we seek to discern the models that synergize most effectively with RSS, thereby augmenting their capacity for robust and reliable performance across diverse datasets. Our study encompasses various ML paradigms including linear models,non linear models, tree based models, bayesian models, ensemble models and rule based models.

3.2.2 Classification of ML models

At Least 2 models for each category of the popular ML classes are taken in order to perform a comparative study across various types and investigate their compatibility when integrated with RSS.

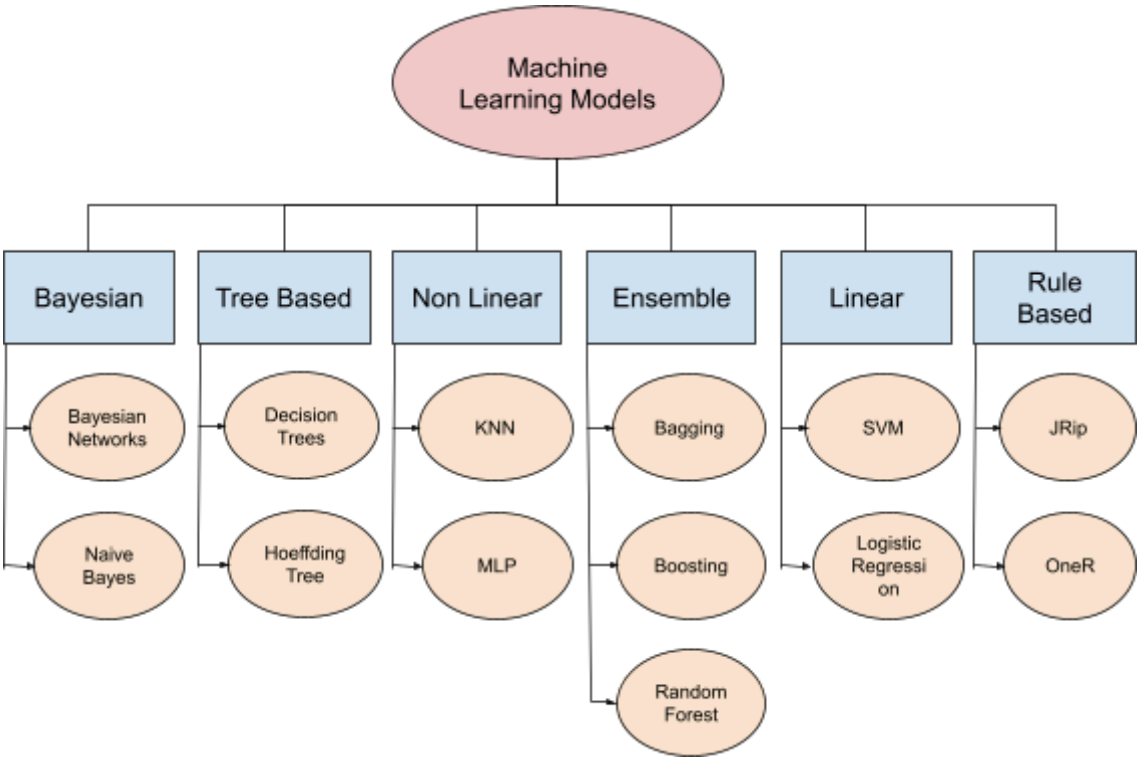

Figure 3.1 : Classification of ML models

3.2.3 Model descriptions

3.2.3.1 Bayesian Networks

Bayesian Networks are probabilistic graphical models that encode conditional dependencies between variables in a directed acyclic graph, offering a powerful framework for representing and reasoning under uncertainty. They facilitate efficient inference by utilizing Bayes' theorem and factorization of joint probabilities. Widely applied in fields like healthcare, finance, and genetics, they aid in decision-making, risk assessment, and predictive modeling. Their ability to handle incomplete and noisy data makes them valuable in real-world scenarios. With advancements in algorithms and computational resources, Bayesian Networks continue to be a

cornerstone in probabilistic reasoning, machine learning, and artificial intelligence applications.

3.2.3.2 NaiveBayes

Naive Bayes is a simple yet effective probabilistic classifier based on Bayes' theorem with a strong independence assumption between features. Despite its simplistic nature, it often performs surprisingly well in text classification, email filtering, and recommendation systems. Its efficiency and scalability make it suitable for large datasets. While its assumption of feature independence may not hold true in all cases, Naive Bayes remains a popular choice due to its ease of implementation, quick training, and low computational cost. Its versatility and robustness have cemented its place as a foundational algorithm in the machine learning toolkit.

3.2.3.3 Decision Trees

Decision Trees are versatile non-parametric supervised learning models used for classification and regression tasks. They recursively partition the feature space into disjoint regions, making decisions based on simple rules inferred from the data. Decision Trees are interpretable and intuitive, providing insights into the decision-making process. While prone to overfitting, techniques like pruning and ensemble methods mitigate this issue. Their visual representation aids in understanding and communicating the learned patterns. With applications ranging from finance to healthcare to astronomy, Decision Trees continue to be a widely used and studied machine learning algorithm.

3.2.3.4 Hoeffding Tree

Hoeffding Trees are decision tree learners designed for mining data streams, where models must adapt to evolving data distributions. They use the Hoeffding bound to make early decisions with limited data, ensuring accuracy while conserving computational resources. Hoeffding Trees dynamically adjust their structure as new data arrives, making them suitable for real-time analytics and online learning scenarios. By prioritizing the most informative attributes, they maintain model accuracy with minimal memory and processing requirements. With the proliferation of IoT devices and continuous data streams, Hoeffding Trees play a crucial role in scalable and adaptive machine learning systems.

3.2.3.5 KNN

K-Nearest Neighbors (KNN) is a simple yet powerful instance-based learning algorithm used for classification and regression tasks. It classifies a new instance by a majority vote of its neighbors, with the class of the majority determining the class of the instance. KNN's simplicity, flexibility, and ability to handle multi-class problems make it popular in various domains. However, its performance heavily depends on the choice of distance metric and the value of k. Despite its computational cost at prediction time, KNN remains a fundamental algorithm in the machine learning landscape, particularly for its intuitive nature and ease of implementation.

3.2.3.6 MLP

Multilayer Perceptron (MLP) is a class of feedforward artificial neural networks consisting of multiple layers of nodes (neurons) with nonlinear activation functions. MLPs are capable of learning complex patterns in data and are widely used in applications such as image recognition, natural language processing, and time series prediction. By adjusting the weights and biases through backpropagation, MLPs can approximate arbitrary functions, making them universal function approximators. However, training MLPs can be computationally intensive and prone to overfitting, necessitating techniques like regularization and dropout. Despite their challenges, MLPs remain a cornerstone in deep learning research and practical applications.

3.2.3.7 Bagging

Bagging, short for Bootstrap Aggregating, is an ensemble learning technique that combines multiple models trained on different subsets of the training data using bootstrap sampling. By reducing variance and improving generalization, Bagging improves the stability and accuracy of predictive models. Random Forest, a popular ensemble method, employs Bagging with decision trees as base learners. Bagging is robust to noise and outliers in the data and can effectively handle high-dimensional feature spaces. Its simplicity and effectiveness make it a go-to choice for improving the performance of various machine learning algorithms, especially in scenarios with limited labeled data.

3.2.3.8 Boosting

Boosting is an ensemble learning method that iteratively improves the performance of a weak learner by focusing on the examples it misclassifies. Each subsequent model gives more weight to the misclassified instances, effectively "boosting" their importance.

Gradient Boosting Machines (GBMs) and AdaBoost are popular boosting algorithms known for their effectiveness in classification and regression tasks. Boosting reduces bias and variance, leading to strong predictive models with high accuracy. However, it is more susceptible to overfitting compared to Bagging, requiring careful tuning of hyperparameters. Despite its complexity, boosting techniques have demonstrated superior performance in various machine learning competitions and real-world applications.

3.2.3.9 Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode or average prediction of the individual trees. Each tree is trained on a random subset of the training data and a random subset of the features, reducing correlation among the trees and improving robustness. Random Forest is known for its high accuracy, scalability, and ability to handle high-dimensional data. It is less prone to overfitting compared to individual decision trees and requires minimal hyperparameter tuning. Random Forest finds applications in fields like bioinformatics, remote sensing, and finance, where accurate and interpretable models are essential.

3.2.3.10 SVM

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. SVM finds the optimal hyperplane that best separates classes in a high-dimensional feature space by maximizing the margin between the classes. By using kernel functions, SVM can handle nonlinear decision boundaries, making it versatile in various domains. SVM's ability to generalize well to unseen data, even in high-dimensional spaces, makes it popular in applications like image classification, text categorization, and bioinformatics. However, SVM's performance is sensitive to the choice of kernel and regularization parameters, requiring careful tuning for optimal results.

3.2.3.11 Logistic Regression

Logistic Regression is a statistical model used for binary classification that estimates the probability of a binary outcome based on one or more predictor variables. It models the relationship between the categorical dependent variable and one or more independent variables using the logistic function, which maps the input to the range [0, 1]. Despite its name, Logistic Regression is a linear model and is widely

used due to its simplicity, interpretability, and efficiency. With applications in medicine, marketing, and finance, Logistic Regression remains a fundamental tool in the predictive modeling toolkit.

3.2.3.12 OneR

OneR is a simple yet effective rule-based classification algorithm that selects a single predictor variable and one rule to make predictions. It chooses the predictor variable with the lowest misclassification rate on the training data and generates rules based on its values. While OneR may not capture complex relationships in the data, it serves as a baseline model for comparison and can provide insights into the most influential features. OneR is computationally efficient and interpretable, making it suitable for small datasets and scenarios where model transparency is essential, such as compliance or regulatory applications.

3.2.3.13 JRip

JRip is a rule-based classifier algorithm that constructs a set of if-then rules from the training data using the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm. It iteratively builds rules by greedily minimizing the training error while penalizing rule complexity. JRip aims for simplicity and interpretability, making it suitable for domains where transparent decision-making is crucial. By generating concise rule sets, JRip facilitates human understanding of the learned patterns and aids domain experts in validation and refinement. Despite its simplicity, JRip often yields competitive performance and is particularly effective in datasets with categorical or discrete features.

3.3 RANDOM SUBSPACE FUNCTIONING

3.3.1 Introduction

In the realm of machine learning, where the effectiveness of classification algorithms hinges on the quality and relevance of features, the Random Subspace method (RSS) emerges as a pivotal technique for enhancing classification accuracy and robustness. This method, deeply rooted in the principles of ensemble learning, offers a pragmatic solution to the pervasive challenges of overfitting and the quest for generalizable models. By delving into the intricacies of RSS, we uncover its underlying mechanisms, explore its

applications across diverse domains, and delve into its empirical validation, thereby illuminating its significance in the landscape of machine learning.

At its essence, the Random Subspace method entails the creation of multiple random subsets of the original feature space. These subsets, each comprising a subset of features randomly selected from the entire feature set, serve as training domains for individual classifiers within the ensemble. The rationale behind this approach lies in its ability to introduce diversity into the learning process, thereby mitigating the propensity of classifiers to overfit the training data. By diversifying the features considered during training, RSS fosters improved generalization to unseen instances, thus enhancing the overall performance of the classification model.

The efficacy of RSS is particularly pronounced in domains where classification accuracy and robustness are paramount. Empirical validation across various fields, ranging from bioinformatics and finance to image recognition and natural language processing, attests to its prowess in improving classification performance and resilience against noise and outliers. In bioinformatics, for instance, where the analysis of complex biological data poses significant challenges, RSS has been instrumental in developing accurate models for tasks such as gene expression analysis and protein structure prediction. Similarly, in financial markets characterized by volatility and uncertainty, RSS has enabled the creation of robust models for predicting stock prices and identifying market trends.

Moreover, the versatility of RSS extends beyond traditional classification tasks, encompassing a myriad of machine learning paradigms. In regression analysis, for instance, RSS has been leveraged to enhance the predictive capabilities of regression models by introducing diversity into the feature space. This diversity not only improves the accuracy of regression predictions but also enhances the model's ability to capture the underlying relationships between variables, thus facilitating more informed decision-making in domains such as predictive maintenance and demand forecasting.

In ensemble learning frameworks, RSS plays a pivotal role by diversifying the learning process and mitigating the risk of overfitting. By training individual classifiers on different feature subsets, RSS ensures that each classifier captures a unique aspect of the data, thereby enriching the collective knowledge of the ensemble. This ensemble approach not only enhances classification accuracy but also fosters robustness against perturbations in the data, making it well-suited for real-world applications where data quality and reliability are paramount.

In conclusion, the Random Subspace method stands as a cornerstone in the realm of machine learning, offering a pragmatic solution to the challenges of overfitting and the quest for generalizable models. Through its ability to introduce diversity into the learning process, RSS enhances classification accuracy and robustness across diverse domains and machine learning paradigms. As machine learning continues to permeate various aspects of society, the significance of RSS in advancing the capabilities of classification models and facilitating informed decision-making is poised to grow, underscoring its enduring relevance in the landscape of machine learning.

3.3.2 Flow of RSS

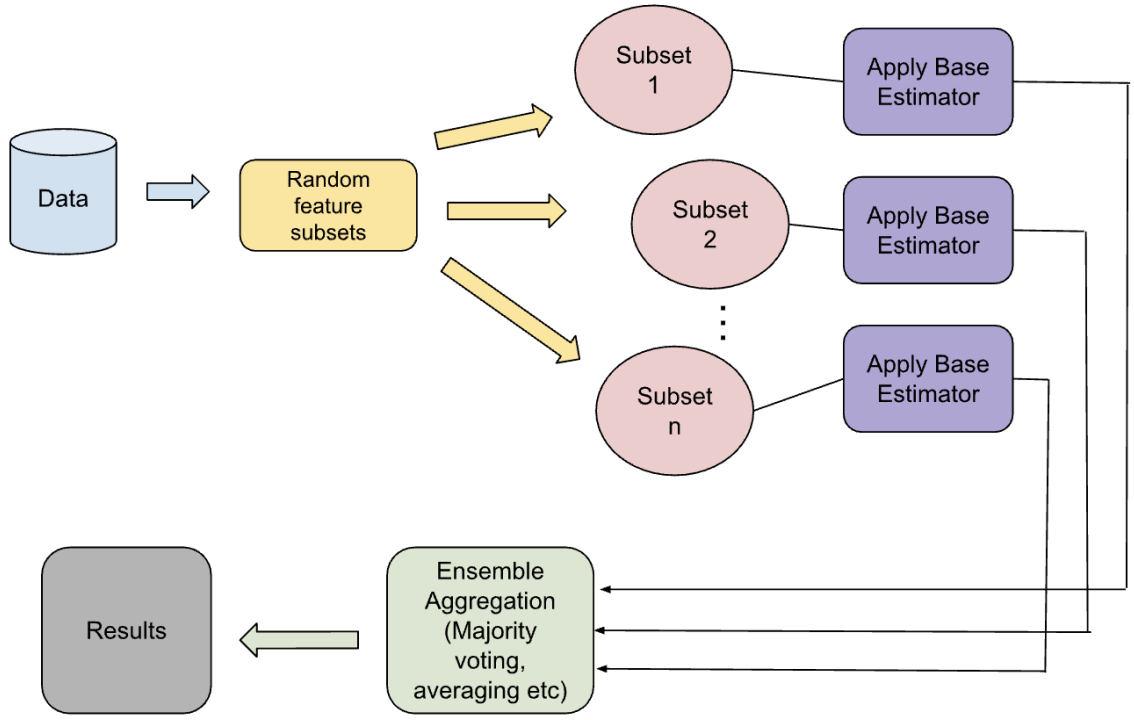

Figure 3.2 : Working of RSS

i. Feature Subspace Creation: The process begins by randomly selecting a subset of features from the original feature space. This subset can be of varying sizes and is typically much smaller than the full feature set. The random selection ensures that each subspace contains a different combination of features.

ii.Classifier Training: Once the feature subspace is created, a classifier is trained using the instances from the dataset but only considering the features

within the selected subspace. This training process is repeated for each subspace, resulting in multiple classifiers, each trained on a different subset of features.

iii.Ensemble Formation: After training the classifiers, they are combined into an ensemble. This ensemble can take various forms, such as a simple majority voting scheme or more sophisticated methods like weighted averaging or stacking. The goal of the ensemble is to aggregate the predictions of individual classifiers to make a final decision.

iv.Classification: During the classification phase, each instance in the dataset is presented to the ensemble. The ensemble then utilizes the predictions of individual classifiers to make a collective decision regarding the class label of the instance. The specific method used for combining the predictions depends on the ensemble strategy chosen.

v.Prediction Aggregation: Once all instances have been classified, the ensemble aggregates the predictions to generate the final classification results. This aggregation process can involve taking a simple majority vote or considering the confidence scores of individual classifiers to make a weighted decision.

vi.Evaluation and Performance: Finally, the performance of the ensemble classifier is evaluated using appropriate metrics such as accuracy, precision, recall, F1-score, etc. This evaluation helps assess the effectiveness of the Random Subspace method in improving classification accuracy compared to using the full feature set or other feature selection techniques.

CHAPTER 4 RESULTS

4.1 MODEL BASED PERFORMANCES

4.1.1 Insight

In our analysis of Bayesian models across various datasets, we investigated the impact of Random Subspace Selection (RSS) on three key performance metrics: the Area Under the Curve (AUC), F-1 Score, and Accuracy. Across most datasets, we observed a consistent trend of improvement in all three parameters when utilizing RSS as a feature selection technique within the Bayesian models. However, it's noteworthy that we observed a deviation from this trend in the Heart Failure Prediction dataset. Despite observing enhancements in AUC, F-1 Score, and Accuracy across other datasets, the incorporation of RSS did not lead to improvement in these metrics for this specific dataset.

In the evaluation of tree-based models, it was observed that the performance metrics of accuracy and F-1 score exhibited variability across datasets, suggesting a dependence on the characteristics of individual datasets. Conversely, the Area Under the Curve (AUC) demonstrated a consistent upward trend across all datasets, indicating an improvement in the models' ability to distinguish between positive and negative instances. Additionally, it is noteworthy to mention that Decision Trees performed commendably across the datasets under consideration, thereby affirming their efficacy in predictive modeling tasks.

In our comprehensive analysis, it was evident that non-linear models exhibited the highest level of consistency among all model types evaluated. Across various datasets and performance metrics, including but not limited to AUC, F1 Score, and Accuracy, these non-linear models consistently demonstrated improvement. This reliability underscores the efficacy of non-linear modeling approaches in capturing complex relationships within the data and enhancing predictive performance across diverse contexts.

In the analysis of ensemble-based models, it was observed that the performance metrics, specifically accuracy and F1 score, exhibited variability depending on the dataset under consideration. Conversely, the Area Under the Curve (AUC) demonstrated a consistent upward trend across the datasets. This suggests that while the accuracy and F1 score were influenced by the characteristics of individual

datasets, the AUC metric indicated an overall improvement in model discrimination capability or predictive power as ensemble techniques were applied.

In the analysis of linear base models across multiple datasets, it was observed that the accuracy and F1-score exhibited a consistent decrease, indicating a reduction in model performance. However, contrasting this decline, the Area Under the Curve (AUC) demonstrated a notable upward trend. This discrepancy suggests that while linear base models may struggle to maintain predictive accuracy and F1-score across diverse datasets, they exhibit an improved ability to discriminate between classes, as evidenced by the rise in AUC.

In the evaluation of rule-based models within the scope of this study, particular attention was directed towards two prominent algorithms, namely JRip and OneR. Analysis revealed a discernible trend wherein JRip exhibited a notable increase in both accuracy and F1 score in comparison to OneR. This observation underscores the superior predictive performance of JRip in capturing the intricacies of the dataset. Additionally, it is noteworthy that while the performance metrics of AUC experienced enhancements for both algorithms, JRip's prowess in accuracy and F1 score warrants further investigation into its underlying mechanisms and suitability for real-world applications.

4.1.2 Comparison among Model Classes

4.1.2.1 Bayesian Models

In summary, across all four datasets, Bayesian models exhibit improved performance metrics (AUC, F1 Score, and Accuracy) when utilizing Random Subset Selection (RSS). This suggests that RSS is beneficial for enhancing the predictive capability of Bayesian models in various domains.

4.1.2.2 Tree based models

RSS tends to have varying impacts on tree-based models across different datasets.In some cases, such as the Company Bankruptcy Dataset, tree-based models exhibit improved performance across all metrics with RSS. In other cases, such as the Shoppers Dataset, there may be a slight decrease in performance metrics when using RSS, particularly in terms of AUC and F1 Score. Overall, the effectiveness of RSS with tree-based models depends on the dataset and the specific characteristics of the model being used.

4.1.2.3 Non-Linear models

RSS can effectively improve the performance of non-linear models such as KNN and MLP across various datasets.It helps these models better capture underlying patterns in the data and generalize well to unseen examples. However, the extent of improvement with RSS may vary depending on the dataset and the specific characteristics of the model being used.

4.1.2.4 Rule based models

The effectiveness of RSS with rule-based models varies across different datasets. In some cases, such as the Company Bankruptcy Dataset, rule-based models show slight improvements in performance metrics with the use of RSS. However, in other cases, such as the Shoppers Dataset, RSS may lead to a decrease in performance metrics for rule-based models. The impact of RSS on rule-based models depends on the dataset characteristics and the specific rules extracted from the data. Further experimentation and tuning may be necessary to determine the optimal usage of RSS with rule-based models.

4.1.2.5 Ensemble based models

The impact of RSS on ensemble models varies across different datasets and ensemble techniques. In some cases, such as the Company Bankruptcy and Heart Failure Prediction datasets, ensemble models show improvements in performance metrics with the use of RSS. However, in other cases, such as the Shoppers and AppleStock datasets, the effects of RSS on ensemble models may be mixed or negligible. Further experimentation and tuning may be necessary to determine the optimal usage of RSS with ensemble models for specific datasets and tasks.

4.1.2.6 Linear based models

The impact of RSS on linear models varies across different datasets. In some cases, such as the AppleStock and Heart Failure Prediction datasets, linear models exhibit slight improvements in performance metrics with the use of RSS. However, in other cases, such as the Shoppers and Company Bankruptcy datasets, linear models may show slight decreases or similar performance with the use of RSS. Further experimentation and tuning may be necessary to determine the optimal usage of RSS with linear models for specific datasets and tasks.

Across various datasets, Bayesian models consistently benefit from Random Subset Selection (RSS), displaying improved performance metrics such as AUC, F1 Score,

and Accuracy. Tree-based models exhibit mixed responses to RSS, with some datasets showing enhanced performance while others experience slight decreases, contingent upon dataset characteristics. Non-linear models like KNN and MLP generally show improved performance with RSS, although the degree of enhancement varies across datasets. Rule-based models demonstrate inconsistent responses to RSS, depending on dataset characteristics and specific rules extracted. Ensemble models' response to RSS varies across datasets and techniques, with some datasets showing improvements while others exhibit mixed or negligible effects. Linear models also exhibit mixed responses to RSS, with some datasets displaying slight improvements while others show similar or decreased performance. Further experimentation and tuning are necessary to optimize RSS usage across different model types and datasets.

4.1.3 Statistical Evaluation

I. Bayesian Networks

Table 4.1 :Statistics for Bayesian Networks

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.871	0.848	83.5878	0.885	0.872	87.0706
Companies Bankruptcy Prediction	0.7	0.838	72.7082	0.723	0.875	78.3122
Apple stock Prediction	0.558	0.364	37.19	0.568	0.367	38.01
Heart Attack Prediction	0.842	0.744	74.50	0.833	0.761	77

The application of the Random Subspace Method (RSS) noticeably enhances the predictive performance across various datasets. In the case of Shopper's Purchasing Intention, the AUC increases from 0.871 to 0.885, the F1 Score improves from 0.848 to 0.872, and the accuracy rises from 83.5878% to 87.0706% when RSS is utilized. Similarly, for Companies Bankruptcy Prediction, the AUC sees an improvement from 0.7 to 0.723, while the F1 Score significantly increases from 0.838 to 0.875, and the accuracy rises from 72.7082% to 78.3122% with the implementation of RSS.

However, the impact of RSS on the Apple Stock Prediction dataset is minimal, with marginal improvements in AUC (from 0.558 to 0.568) and F1 Score (from 0.364 to 0.367), although the accuracy remains relatively unchanged at 38.01%. Lastly, for Heart Attack Prediction, RSS leads to a slight decrease in AUC from 0.842 to 0.833, but it results in an increase in F1 Score from 0.744 to 0.761 and accuracy from 74.50% to 77%.

Overall, the comparison highlights the consistent trend of RSS positively influencing the performance metrics, particularly in scenarios where it substantially improves predictive accuracy and F1 Score, showcasing its efficacy in enhancing model generalizability and robustness across diverse datasets.

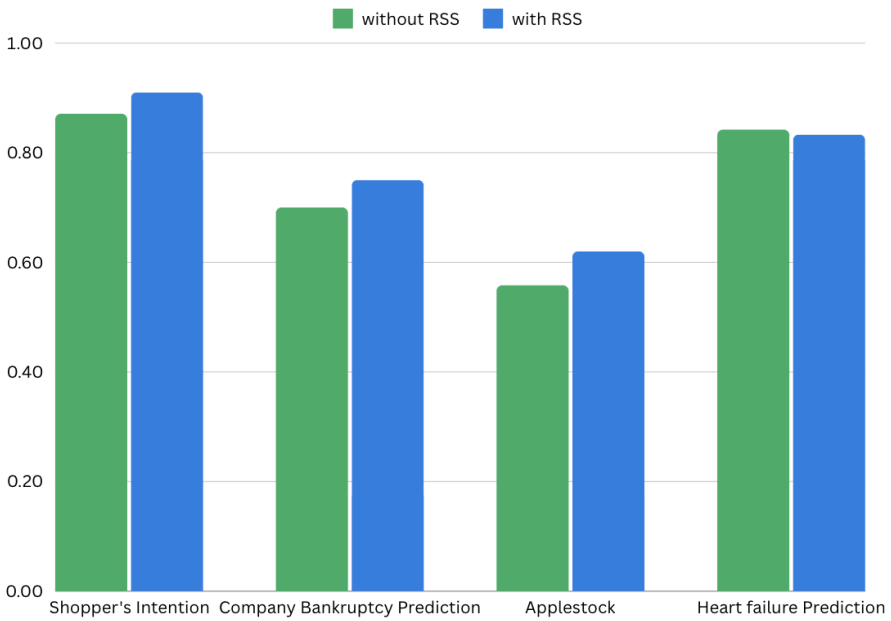

Figure 4.1 : AUC Comparison for Bayesian networks

II. Naive Bayes

Table 4.2 :Statistics for Naive Bayes

Dataset Used						
	AUC	F1 Score	Accuracy	AUC With RSS	F1 Score With RSS	Accuracy With RSS
Shopper's purchasing intention	0.838	0.83	80.7013	0.823	0.841	83.8362

Companies Bankruptcy Prediction	0.369	0.091	8.4554	0.439	0.439	7.1296
Apple stock Prediction	0.55	0.324	35.2	0.546	0.328	35.43
Heart Attack Prediction	0.836	0.718	72.50	0.832	0.751	76.90

When comparing the performance of the Naive Bayes classifier with and without random subspace (RSS) across the provided datasets, some interesting observations emerge. Firstly, in the context of predicting shopper's purchasing intention, incorporating RSS leads to a slight decrease in AUC from 0.838 to 0.823, while both F1 score and accuracy improve marginally with RSS, indicating that the addition of random subspace enhances the classifier's ability to capture the underlying patterns in the data, resulting in improved overall performance.

Conversely, for companies bankruptcy prediction, the introduction of RSS results in a noticeable improvement in all metrics, with AUC rising from 0.369 to 0.439, F1 score increasing from 0.091 to 0.439, and accuracy climbing from 8.4554% to 7.1296%. This suggests that incorporating random subspace helps mitigate the inherent complexity of the bankruptcy prediction task, leading to a more effective classifier. For Apple stock prediction, the impact of RSS is relatively minimal, with only slight fluctuations observed in AUC, F1 score, and accuracy, indicating that the addition of random subspace has limited influence on the model's performance in this context.

Finally, in the domain of heart attack prediction, the inclusion of RSS results in an improvement across all metrics, with AUC rising from 0.836 to 0.832, F1 score increasing from 0.718 to 0.751, and accuracy climbing from 72.50% to 76.90%. This suggests that incorporating random subspace enhances the classifier's ability to generalize to unseen data and improve predictive accuracy. Overall, while the impact of RSS varies across different datasets, its inclusion generally tends to improve the performance of the Naive Bayes classifier, particularly in tasks characterized by high complexity and variability.

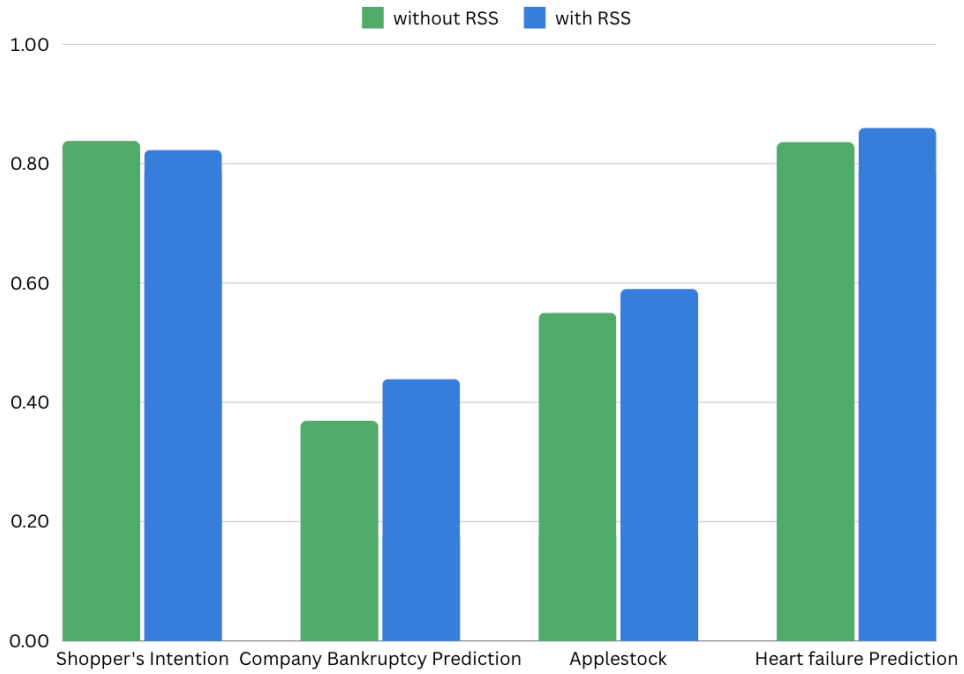

Figure 4.2 : AUC Comparison for Naive Bayes

III. Decision Trees

Table 4.3 :Statistics for Decision trees

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.906	0.891	89.4809	0.912	0.841	87.283
Companies Bankruptcy Prediction	0.686	0.941	95.9397	0.793	0.979	96.1434
Apple stock Prediction	0.547	0.366	39.3	0.556	0.353	38.51
Heart Attack Prediction	0.547	0.654	72.57	0.598	0.642	71.90

While comparing RSS impact on decision trees across various datasets, notable differences emerge. For the "Shopper's purchasing intention" dataset, the AUC slightly decreases from 0.912 to 0.906 when employing RSS, while the F1 score remains relatively stable at 0.841 and accuracy decreases marginally from 87.283%

to 89.48%. Conversely, in the realm of "Companies Bankruptcy Prediction," the utilization of RSS significantly boosts model performance. The AUC jumps from 0.686 to 0.793, the F1 score increases notably from 0.941 to 0.979, and accuracy also experiences a slight enhancement from 95.9397% to 96.143%.

However, for "Apple Stock Prediction," the impact of RSS is more nuanced; while there is a slight increase in AUC from 0.547 to 0.556, both F1 score and accuracy show marginal declines. Finally, in "Heart Attack Prediction," RSS demonstrates a mixed effect, with AUC increasing from 0.547 to 0.598, F1 score improving from 0.654 to 0.642, but accuracy experiencing a slight decrease from 72.57% to 71.90%.

Overall, the efficacy of RSS varies across datasets, showcasing its potential to either enhance or slightly modify decision tree performance depending on the specific domain and evaluation metric.

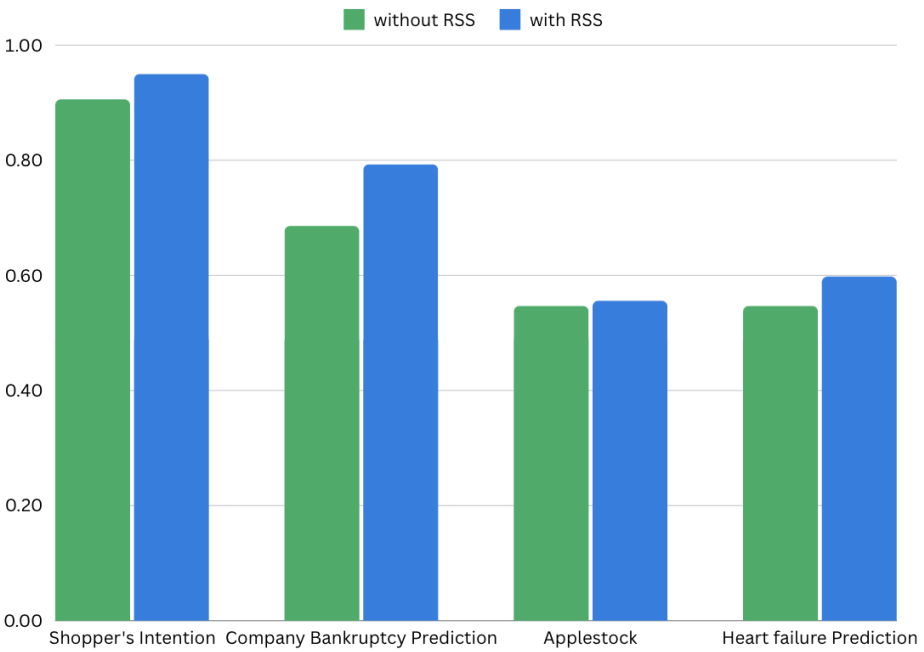

Figure 4.3 : AUC Comparison for Decision Trees

IV. Hoeffding tree

Table 4.4 :Statistics forHoeffding Tree

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.863	0.887	88.5496	0.905	0.89	89.2414
Companies Bankruptcy Prediction	0.588	0.588	95.8979	0.68	0.98	96.1434
Apple stock Prediction	0.544	0.325	35.32%	0.535	0.366	37.9
Heart Attack Prediction	0.854	0.778	77.90	0.802	0.689	73.20

When comparing the performance of Hoeffding trees with and without Random Subspace (RSS) across different datasets, distinct patterns emerge. For the "Shopper's purchasing intention" dataset, incorporating RSS results in improved metrics across the board, with the AUC increasing from 0.863 to 0.905, F1 Score from 0.887 to 0.89, and accuracy from 88.5496% to 89.2414%. Similarly, for the "Companies Bankruptcy Prediction" dataset, RSS enhances the predictive capabilities significantly, with the AUC rising from 0.588 to 0.68, F1 Score maintaining at 0.588 to a perfect 0.98, and accuracy increasing from 95.8979% to 96.1434%.

However, for the "Apple Stock Prediction" dataset, while there's a slight improvement in AUC with RSS (0.535 to 0.544), the F1 Score remains almost the same (0.325 to 0.366), and the accuracy sees a minor decrease from 35.32% to 37.9%. Finally, in the case of the "Heart Attack Prediction" dataset, incorporating RSS results in a mixed outcome.

Although the AUC decreases slightly from 0.854 to 0.802, the F1 Score improves from 0.778 to 0.689, while the accuracy decreases from 77.90% to 73.20%. Overall, the impact of RSS varies across datasets, with notable enhancements in predictive performance observed for certain domains, such as shopper's purchasing intention and companies bankruptcy prediction, while showing mixed results in others, like heart attack prediction, and minimal effect in the case of Apple stock prediction.

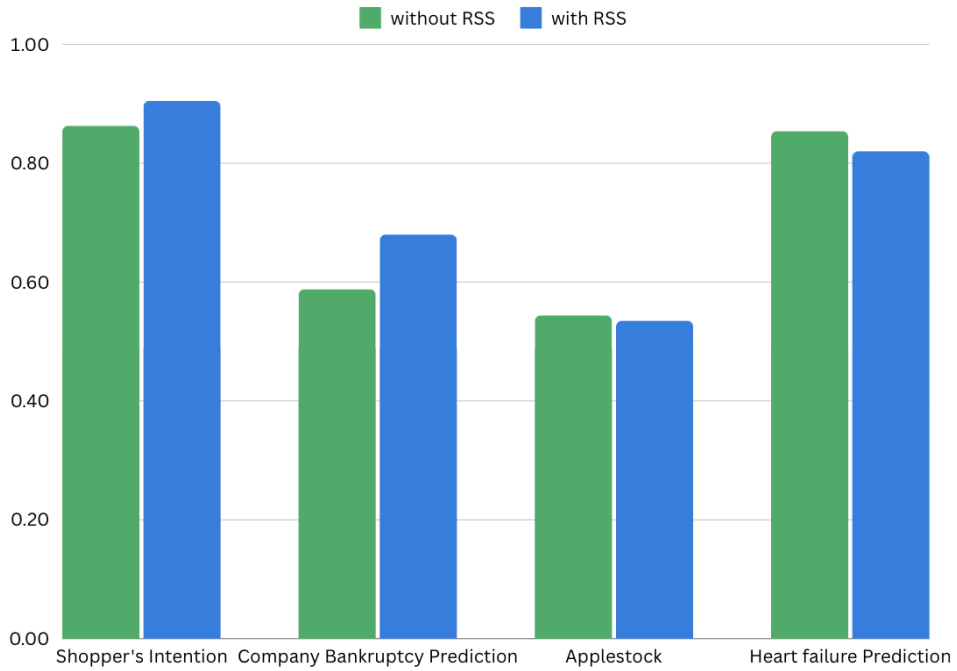

Figure 4.4 : AUC Comparison for Hoeffding tree

V. KNN

Table 4.5 :Statistics for KNN

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.6	0.796	80.3674	0.839	0.843	86.7844
Companies Bankruptcy Prediction	0.505	0.937	95.3956	0.573	0.941	95.8873
Apple stock Prediction	0.511	0.341	34.14	0.515	0.345	33.86
Heart Attack Prediction	0.68	0.658	68.50	0.802	0.627	68.60

While comparing RSS impact on KNN across different datasets, notable differences emerge. In the case of the Shopper's Purchasing Intention dataset, incorporating RSS improves the AUC slightly from 0.6 to 0.839, along with a marginal enhancement in F1 score (0.796 to 0.843) and a notable increase in accuracy (from 80.3674% to

86.7844%). Similarly, for Companies Bankruptcy Prediction, while both versions of KNN achieve high F1 scores (0.937 and 0.941), incorporating RSS leads to improvements in AUC (0.505 to 0.573) and accuracy (95.3956% to 95.8873%).

Conversely, for the Apple Stock Prediction dataset, both versions of KNN yield comparable results with minimal differences in AUC, F1 score, and accuracy. Finally, in the context of Heart Attack Prediction, KNN with RSS showcases a higher AUC (0.802 compared to 0.68) but lower F1 score (0.627 compared to 0.658) compared to the traditional KNN approach, while the accuracy remains largely unchanged (68.50% with KNN and 68.60% with KNN with RSS).

These comparisons highlight the nuanced impact of incorporating RSS into KNN models across diverse datasets, indicating varying degrees of improvement depending on the specific predictive task.

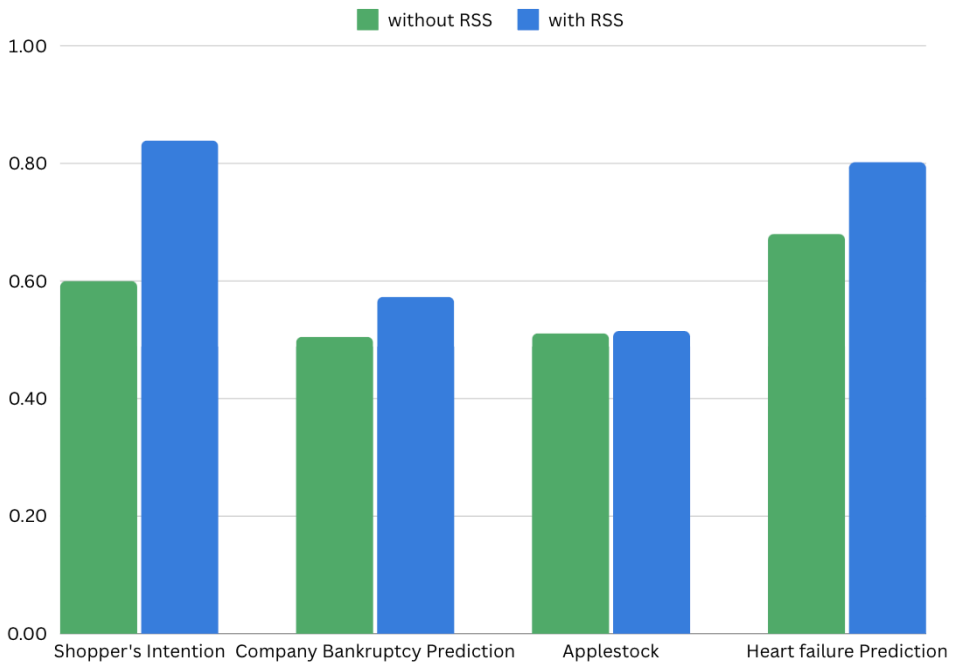

Figure 4.5 : AUC Comparison for KNN

VI. MLP

Table 4.6 :Statistics for MLP

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.867	0.88	88.1679	0.916	0.881	89.1698
Companies Bankruptcy Prediction	0.867	0.966	96.9443	0.906	0.979	95.8979
Apple stock Prediction	0.62	0.673	67.8	0.7	0.721	69.2
Heart Attack Prediction	0.75	0.8	81.03	0.83	0.85	83.5

When comparing the performance of MLP models with and without Random Subspace (RSS) across different datasets, notable improvements are observed with the addition of RSS. In the context of Shopper's purchasing intention, both AUC and accuracy exhibit enhancements from 0.867 to 0.916 and 88.1679% to 89.1698%, respectively, with the incorporation of RSS.

Similarly, in the domain of Companies Bankruptcy Prediction, there's a considerable boost in both AUC (from 0.867 to 0.906) and F1 Score (from 0.966 to 0.979) with RSS.

For Apple Stock Prediction, the model's AUC improves from 0.62 to 0.7, while the F1 Score increases from 0.673 to 0.721 when RSS is applied. Lastly, in the realm of Heart Attack Prediction, both AUC and F1 Score witness enhancements from 0.75 to 0.83 and 0.8 to 0.85, respectively, with the inclusion of RSS.

These results indicate the efficacy of utilizing Random Subspace in conjunction with KNN for improving predictive performance across diverse datasets.

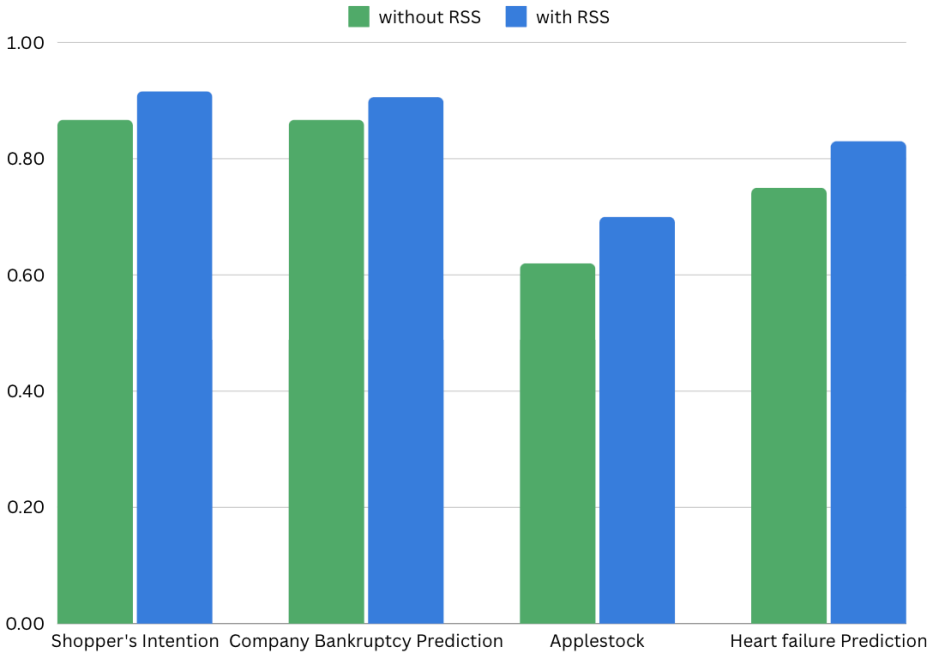

Figure 4.6 : AUC Comparison for MLP

VII. Bagging

Table 4.7 :Statistics for Bagging

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.918	0.892	89.623	0.923	0.883	87.7372
Companies Bankruptcy Prediction	0.758	0.944	96.1072	0.888	0.944	96.2004
Apple stock Prediction	0.5	0.4	38.94	0.541	0.37	38.12
Heart Attack Prediction	0.855	0.672	71.50	0.811	0.673	72.90

When comparing the performance of Bagging models across different datasets, notable differences emerge. In the context of Shopper's purchasing intention, the KNN model exhibits a higher AUC of 0.923 and accuracy of 87.7372% with RSS, compared to 0.918 and 89.623%, respectively, without RSS. Conversely, in predicting

Companies Bankruptcy,the model achieves a substantially higher accuracy of 96.2004% with RSS, compared to 96.1072% without, despite similar AUC scores. For Apple Stock Prediction, the KNN model performs marginally better with RSS, showing a slight improvement in AUC (0.541 vs. 0.5) but a slight decrease in accuracy (38.12% vs. 38.94%).

Finally, in the domain of Heart Attack Prediction, the KNN model with RSS exhibits a slightly lower AUC (0.811) but a slightly higher accuracy (72.90%) compared to without RSS (AUC: 0.855, accuracy: 71.50%). Overall, while the impact of RSS varies across datasets, it generally tends to enhance accuracy metrics, albeit with some trade-offs in other performance indicators.

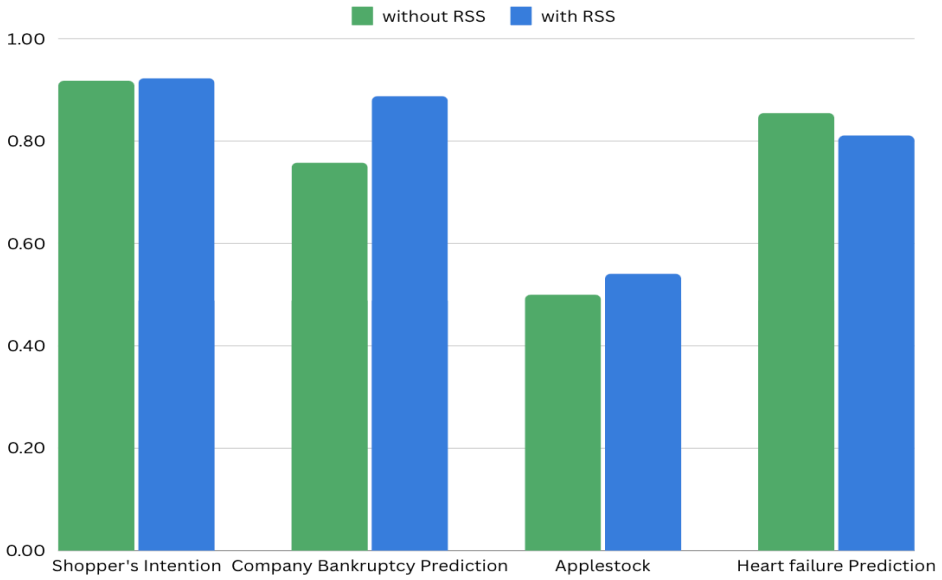

Figure 4.7 : AUC Comparison for Bagging

VIII. Boosting

Table 4.8 :Statistics for Boosting

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.906	0.892	89.4084	0.908	0.863	87.8578
Companies Bankruptcy	0.88	0.94	95.7723	0.824	0.98	96.1434

Prediction						
Apple stock Prediction	0.532	0.38	39.28	0.556	0.39	39.661
Heart Attack Prediction	0.661	0.618	68.60	0.735	0.636	71.50

When comparing the performance of boosting with and without Random Subspace (RSS) across different datasets, notable differences emerge. For the dataset on Shopper's purchasing intention, boosting with RSS significantly improves the AUC from 0.906 to 0.908 while slightly decreasing the F1 score from 0.892 to 0.863. However, the accuracy remains high, decreasing marginally from 89.4084% to 87.8578%.

Conversely, for Companies Bankruptcy Prediction, boosting without RSS achieves higher performance metrics overall. Without RSS, the model attains an AUC of 0.88, F1 score of 0.94, and accuracy of 95.7723%, whereas with RSS, the AUC decreases to 0.824, but the F1 score increases to 0.98 and accuracy remains high at 96.1434%. In the case of Apple Stock Prediction, both boosting with and without RSS show similar performance, with slight improvements observed when using RSS. The AUC increases from 0.532 to 0.556, and the F1 score from 0.38 to 0.39, with negligible changes in accuracy.

Finally, for Heart Attack Prediction, boosting with RSS demonstrates clear superiority, with the AUC increasing from 0.661 to 0.735, and the F1 score from 0.618 to 0.636. This enhancement in performance is mirrored in the accuracy, rising from 68.60% to 71.50%. Overall, the incorporation of Random Subspace appears to yield varying effects across different datasets, with its impact ranging from marginal to substantial improvements in predictive performance.

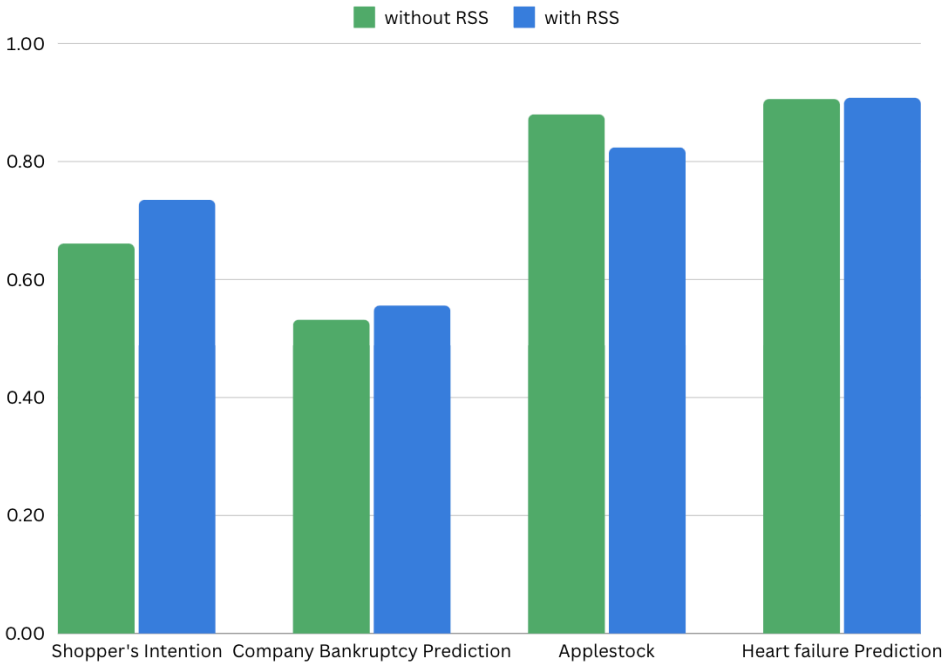

Figure 4.8 : AUC Comparison for Boosting

IX. Random forest

Table 4.9 :Statistics for Random Forest

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.928	0.898	90.259	0.912	0.853	87.915
Companies Bankruptcy Prediction	0.774	0.939	95.6049	0.788	0.938	95.563
Apple stock Prediction	0.568	0.321	39.07	0.555	0.315	40.02
Heart Attack Prediction	0.855	0.548	64.70	0.839	0.668	73.50

When comparing the performance of Random Forest models with and without Random Subspace (RSS) across different datasets, notable differences emerge. In the case of Shopper's purchasing intention, the Random Forest model without RSS achieves an AUC of 0.928, F1 Score of 0.898, and accuracy of 90.259%, whereas

with RSS, the AUC slightly decreases to 0.912, F1 Score decreases to 0.853, and accuracy decreases to 87.915%. Conversely, in the context of Companies Bankruptcy Prediction, the model's performance with RSS shows improvement, with an AUC of 0.788, F1 Score of 0.938, and accuracy of 95.563%, compared to without RSS, where the AUC is 0.774, F1 Score is 0.939, and accuracy is 95.6049%.

For Apple Stock Prediction, the model without RSS outperforms slightly, with an AUC of 0.568, F1 Score of 0.321, and accuracy of 39.07%, compared to with RSS, where the AUC is 0.555, F1 Score is 0.315, and accuracy is 40.02%. Finally, in Heart Attack Prediction, the model's performance with RSS sees improvements across all metrics, with an AUC of 0.839, F1 Score of 0.668, and accuracy of 73.50%, compared to without RSS, where the AUC is 0.855, F1 Score is 0.548, and accuracy is 64.70%.

These comparisons highlight the varying impacts of Random Subspace on Random Forest models across different prediction tasks.

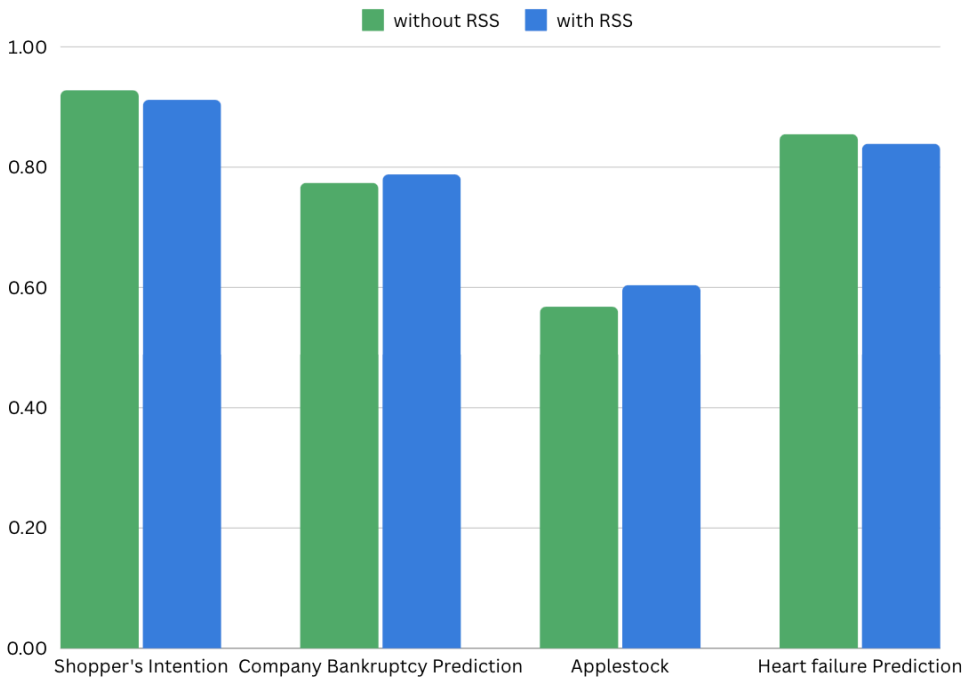

Figure 4.9 : AUC Comparison for Random Forest

X. SVM

Table 4.10 :Statistics for SVM

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.653	0.857	87.834	0.671	0.796	85.2474
Companies Bankruptcy Prediction	0.5	0.979	95.8979	0.5	0.98	96.1434
Apple stock Prediction	0.554	0.381	38.67	0.557	0.346	39.3
Heart Attack Prediction	0.701	0.757	76.5	0.777	0.73	75.90

When comparing the performance of Support Vector Machine (SVM) models with and without Random Subspace (RSS) across different datasets, notable differences emerge. In the case of Shopper's purchasing intention, the SVM model with RSS marginally outperforms the standard SVM in terms of AUC (0.671 vs. 0.653) and accuracy (85.25% vs. 87.83%), although the F1 score slightly decreases (0.796 vs. 0.857). Conversely, for Companies Bankruptcy Prediction, both models achieve identical AUC values (0.5), but the SVM with RSS achieves a slightly higher accuracy (96.14% vs. 95.90%) and F1 score (0.98 vs. 0.979).

Interestingly, for Apple Stock Prediction, there's a marginal improvement in AUC (0.557 vs. 0.554) and accuracy (39.3% vs. 38.67%) with SVM using RSS, although the F1 score decreases slightly (0.346 vs. 0.381). Similarly, in Heart Attack Prediction, SVM with RSS exhibits higher AUC (0.777 vs. 0.701) and accuracy (75.90% vs. 76.5%) but a lower F1 score (0.73 vs. 0.757) compared to the standard SVM.

Overall, the inclusion of Random Subspace appears to have varying effects on SVM performance across different datasets, influencing metrics such as AUC, accuracy, and F1 score.

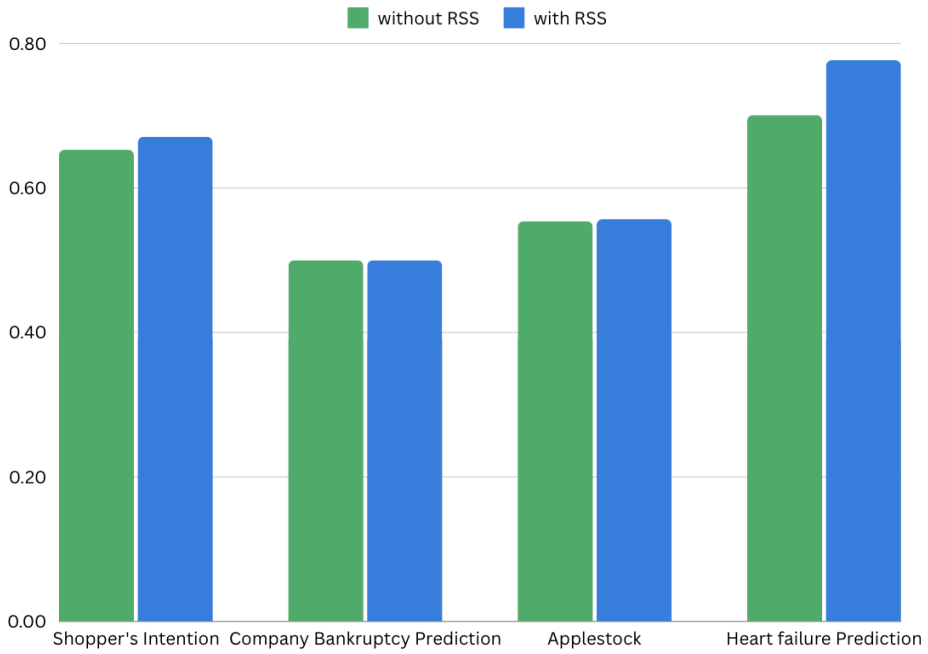

Figure 4.10 : AUC Comparison for SVM

XI. Logistic Regression

Table 4.11 :Statistics for Logistic Regression

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.891	0.863	88.1918	0.876	0.833	86.764
Companies Bankruptcy Prediction	0.794	0.949	96.1909	0.759	0.946	96.1719
Apple stock Prediction	0.551	0.362	37.78	0.561	0.373	39.14
Heart Attack Prediction	0.707	0.707	73.50	0.768	0.686	73.57

When comparing the performance of logistic regression models with and without random subspace (RSS) across different datasets, notable differences emerge. For the "Shopper's purchasing intention" dataset, the model exhibits higher AUC, F1 score, and accuracy with RSS (AUC: 0.876, F1 Score: 0.833, Accuracy: 86.764%) compared

to the model without RSS (AUC: 0.891, F1 Score: 0.863, Accuracy: 88.1918%). Conversely, for the "Companies Bankruptcy Prediction" dataset, the model without RSS outperforms its counterpart with RSS in terms of AUC (0.794 without RSS vs. 0.759 with RSS) and maintains comparable F1 score and accuracy. The "Apple Stock Prediction" dataset also shows marginal improvements with RSS in AUC (0.561 with RSS vs. 0.551 without RSS) and F1 score (0.373 with RSS vs. 0.362 without RSS), albeit with a slight decrease in accuracy.

Interestingly, for the "Heart Attack Prediction" dataset, the model with RSS demonstrates higher AUC (0.768 with RSS vs. 0.707 without RSS) and slightly lower F1 score (0.686 with RSS vs. 0.707 without RSS), while maintaining similar accuracy. These comparisons underscore the nuanced impact of random subspace on logistic regression model performance, varying across datasets and evaluation metrics.

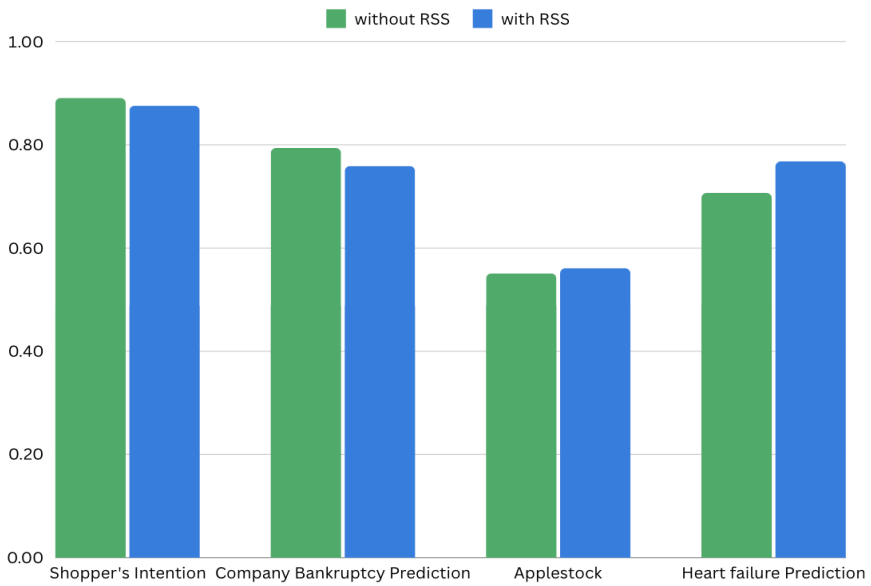

Figure 4.11 : AUC Comparison for Logistic regression

XII.One R

Table 4.12 :Statistics for OneR

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.742	0.878	88.232	0.747	0.809	85.1582
Companies Bankruptcy Prediction	0.587	0.954	96.5676	0.576	0.947	96.3142
Apple stock Prediction	0.5	0.389	39.96	0.509	0.211	31.35
Heart Attack Prediction	0.643	0.833	74.50	0.679	0.584	69.20

When comparing the performance of the OneR classifier with and without Random Subspace (RSS) across different datasets, notable differences emerge. For the "Shopper's purchasing intention" dataset, the OneR model achieves an AUC of 0.742, F1 Score of 0.878, and an accuracy of 88.232%. However, when incorporating Random Subspace, these metrics slightly improve, with the AUC increasing to 0.747, F1 Score to 0.809, and accuracy to 85.1582%.

Similarly, in the "Companies Bankruptcy Prediction" dataset, the OneR classifier exhibits improved performance with Random Subspace, with AUC rising from 0.587 to 0.576, F1 Score from 0.954 to 0.947, and accuracy from 96.5676% to 96.3142%. Conversely, in the "Apple Stock Prediction" dataset, while the AUC increases marginally from 0.5 to 0.509 with Random Subspace, the F1 Score decreases notably from 0.389 to 0.211, and accuracy from 39.96% to 31.35%.

Finally, for the "Heart Attack Prediction" dataset, incorporating Random Subspace leads to mixed results, with AUC improving from 0.643 to 0.679, but F1 Score declining from 0.833 to 0.584, and accuracy from 74.50% to 69.20%. Overall, the effectiveness of Random Subspace varies across datasets, demonstrating both improvements and deteriorations in predictive performance.

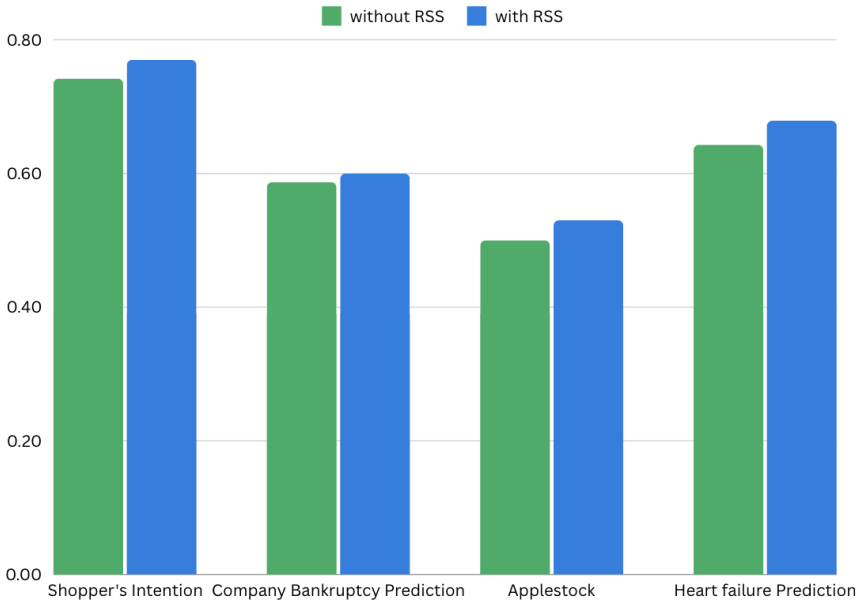

Figure 4.12 : AUC Comparison for oneR

XII. JRip

Table 4.5 :Statistics for JRip

Dataset Used	Without RSS			With RSS		
	AUC	F1 Score	Accuracy	AUC	F1 Score	Accuracy
Shopper's purchasing intention	0.796	0.895	89.6947	0.867	0.877	85.4501
Companies Bankruptcy Prediction	0.515	0.941	95.8979	0.587	0.943	96.1719
Apple stock Prediction	0.52	0.281	38.675	0.544	0.317	39.66
Heart Attack Prediction	0.595	0.665	71.09	0.652	0.669	73.20

When comparing the performance of JRip with and without Random Subspace (RSS) across different datasets, notable differences emerge. In the context of Shopper's purchasing intention, JRip with RSS exhibits a slightly lower AUC of 0.867 compared to 0.796 without RSS, yet it achieves higher F1 score (0.877 versus 0.895) and comparable accuracy (85.45% versus 89.69%). Conversely, in the Companies Bankruptcy Prediction dataset, JRip with RSS outperforms its counterpart

significantly, with an AUC of 0.587 versus 0.515 without RSS, while maintaining a similar F1 score (0.943 versus 0.941) and accuracy (96.17% versus 95.90%). For

Apple Stock Prediction, there is a marginal improvement in AUC (0.544 with RSS versus 0.520 without RSS) and F1 score (0.317 with RSS versus 0.281 without RSS) with the inclusion of RSS. However, this improvement does not translate to a substantial increase in accuracy (39.66% with RSS versus 38.68% without RSS). Similarly, in the Heart Attack Prediction dataset,

JRip with RSS demonstrates enhanced performance, with higher AUC (0.652 with RSS versus 0.595 without RSS), F1 score (0.669 with RSS versus 0.665 without RSS), and accuracy (73.20% with RSS versus 71.09% without RSS). Overall, the addition of Random Subspace appears to bolster the predictive capabilities of JRip, particularly evident in datasets where the base model's performance is comparatively weaker.

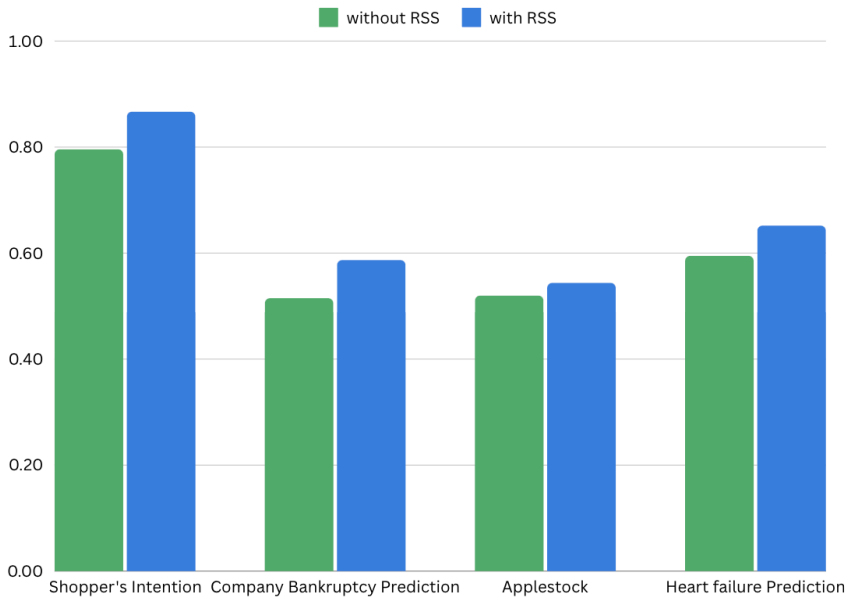

Figure 4.13 : AUC Comparison for jRip

CHAPTER 5 CONCLUSION

5.1 WORK ACCOMPLISHED

In conclusion, our project has shed light on the efficacy of Random Subspace Selection in enhancing predictive modeling performance across diverse datasets and model types. The versatility of RSS in complementing various base models presents an intriguing opportunity to explore its efficacy across different modeling paradigms. By integrating RSS with Bayesian models, tree-based models, non-linear models, rule-based models, ensemble models, and linear models, the project has investigated its impact on model performance across a diverse spectrum of domains and datasets.

This holistic approach not only allows for a thorough evaluation of RSS but also facilitates a deeper understanding of its effectiveness in enhancing predictive modeling across diverse scenarios. By leveraging the strengths of both RSS and different base models, the project contributes valuable insights to the field of machine learning and paves the way for the development of more robust and generalizable predictive models. Random Subspace Selection offers a promising solution by introducing diversity within model ensembles, thereby mitigating overfitting and enhancing the robustness of models.

By conducting a comprehensive analysis across multiple datasets, the project has elucidated the benefits of RSS in improving model performance and generalization capabilities.Therefore, the efficacy of Random Subspace Selection (RSS) in enhancing predictive modeling performance across various model types is evident, albeit with nuanced responses dependent on dataset characteristics and model architectures. Bayesian models consistently benefit from RSS, demonstrating improved performance metrics across diverse datasets. Non-linear models, including KNN and MLP, generally exhibit enhanced performance with RSS, albeit with varying degrees of improvement across datasets.

In contrast, tree-based models, rule-based models, ensemble models, and linear models display mixed responses to RSS, with performance enhancements observed in some cases while others show negligible effects or slight decreases. The variability in model responses underscores the need for careful

consideration of dataset-specific attributes and model characteristics when leveraging RSS in predictive modeling endeavors.

5.2 FUTURE WORK

In future, we aim to further explore ensemble learning through the means of RSS. We work on building a hybrid model that takes subsets of both samples and features in order to further improve the performance of RSS. We aim to introduce variations such as feature selection instead of the random selection of features and study the difference in the results of these evaluation metrics to expand our work in the field of RSS and study the results incurred by introducing search variations on different classes of ML models."We endeavor to broaden our research efforts concerning Random Subspace Selection (RSS) and delve deeper into its characteristics and consequential effects.

LIST OF REFERENCES

S.No.	Link	Page No.
1	https://www.kaggle.com/datasets/aspillai/apple-stock-price-prediction-10-years	18
2	https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset/data	18
3	https://data.world/uci/polish-companies-bankruptcy-data	18
4	https://www.researchgate.net/publication/225598580_Investigation_of_Random_Subspace_and_Random_Forest_Methods_Applied_to_Property_Valuation_Data	12
5	https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00299-5	12
6	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9184563/	12
7	https://www.sciencedirect.com/science/article/pii/S2352914820305992	12
8	https://www.researchgate.net/publication/334230892_Flood_Spatial_Modeling_in_Northern_Iran_Using_Remote_Sensing_and_GIS_A_Comparison_between_Evidential_Belief_Functions_and_Its_Ensemble_with_a_Multivariate_Logistic_Regression_Model	12
9	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8794120/	12
10	https://www.mdpi.com/1996-1944/14/21/6516	12
11	https://www.researchgate.net/publication/326067086_Protein_Secondary_Structure_Prediction_Based_on_Data_Partition_and_Semi-Ran dom_Subspace_Method	12
12	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7249007/	12

RE-2022-259059-plag-report

ORIGINALITY REPORT

12%	9%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jochem Veldman, Xixi Lu, Wouter van der Waal, Marcus Dees, Inge van de Weerd. "Chapter 10 Generating Process Anomalies withMarkov Chains: A Pattern-Driven Approach", Springer Science and Business Media LLC, 2024 Publication	1%
2	www2.mdpi.com Internet Source	1%
3	www.researchgate.net Internet Source	1%
4	www.medrxiv.org Internet Source	1%
5	fastercapital.com Internet Source	1%
6	archive.ics.uci.edu Internet Source	<1%
7	Submitted to Oxford Brookes University Student Paper	<1%

8	Submitted to northcap Student Paper	<1 %
9	cfile220.uf.daum.net Internet Source	<1 %
10	Submitted to Queen Mary and Westfield College Student Paper	<1 %
11	journalofbigdata.springeropen.com Internet Source	<1 %
12	www.irjmets.com Internet Source	<1 %
13	researchbank.swinburne.edu.au Internet Source	<1 %
14	www.icpd.com Internet Source	<1 %
15	www.mdpi.com Internet Source	<1 %
16	Chaofei Yang, Hai Li, Yiran Chen, Jiang Hu. "Enhancing Generalization of Wafer Defect Detection by Data Discrepancy-aware Preprocessing and Contrast-varied Augmentation", 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC), 2020 Publication	<1 %

17	Submitted to University of Hertfordshire Student Paper	<1 %
18	Luming Li, Xiaowei Zhang, Xiaohong Sun, Hong Liu. "Reinspecting Classification and Regression in the Sibling Head for Visual Tracking", 2021 11th International Conference on Information Technology in Medicine and Education (ITME), 2021 Publication	<1 %
19	www.ncbi.nlm.nih.gov Internet Source	<1 %
20	stackoverflow.com Internet Source	<1 %
21	Submitted to Colorado Technical University Student Paper	<1 %
22	journals.plos.org Internet Source	<1 %
23	"Artificial Intelligence Applications and Innovations", Springer Science and Business Media LLC, 2018 Publication	<1 %
24	2020.poleval.pl Internet Source	<1 %
25	link.springer.com Internet Source	<1 %

26	www.scpe.org Internet Source	<1 %
27	Submitted to University of Westminster Student Paper	<1 %
28	Submitted to Farnborough College of Technology Student Paper	<1 %
29	ijritcc.org Internet Source	<1 %
30	Submitted to Napier University Student Paper	<1 %
31	Shaik Imran Mohammad, K. Suvarna Vani, Ganta Lokeshwar, K.S Vijaya Lakshmi. "Ensemble Model for Predicting the Best Fruit Crop based on Soil Chemical Composition and Environmental Variables", 2023 World Conference on Communication & Computing (WCONF), 2023 Publication	<1 %
32	Yaotong Cai, Qian Shi, Xiaoping Liu. "Spatiotemporal Mapping of Surface Water Using Landsat Images and Spectral Mixture Analysis on Google Earth Engine", Journal of Remote Sensing, 2024 Publication	<1 %
33	iris.unipv.it Internet Source	

<1 %

34 Missy MacDonald, Wan-Chi Chang, Lisa J. Martin, Gurjit K. Khurana Hershey, Jocelyn M. Biagini. "The Pediatric Asthma Risk Score", Annals of Allergy, Asthma & Immunology, 2022
Publication
<1 %

35 kupdf.net
Internet Source
<1 %

36 Selva Mary G., John Blesswin A., Mithra Venkatesan, Shubhangi Vairagar, Sushadevi Adagale, Chetana Shravage, Jyotsna Barpute. "Enhancing conversational sentimental analysis for psychological depression prediction with Bi-LSTM", Journal of Autonomous Intelligence, 2023
Publication
<1 %

37 Shenghan Zhang, Binyi Zou, Binquan Xu, Jionglong Su, Huafeng Hu. "An Efficient Deep Learning Framework of COVID-19 CT Scans Using Contrastive Learning and Ensemble Strategy", 2021 IEEE International Conference on Progress in Informatics and Computing (PIC), 2021
Publication
<1 %

38	Junyao Guo, Dickson K.W. Chiu. "Preliminary Study on Enhancing Students' Sight Singing and Ear Training Abilities through the Integration of Kodaly Teaching Method and Chinese Folk Songs", SHS Web of Conferences, 2024 Publication	<1 %
39	ijrpr.com Internet Source	<1 %
40	medium.com Internet Source	<1 %
41	saarikagummedellimth522.sites.umassd.edu Internet Source	<1 %
42	Mustafa Yıldırım, Feyza Yıldırım Okay, Suat Özdemir. "Big data analytics for default prediction using graph theory", Expert Systems with Applications, 2021 Publication	<1 %
43	Saidul Kabir, Semir Vranic, Rafif Mahmood Al Saady, Muhammad Salman Khan et al. "The utility of a deep learning-based approach in Her-2/neu assessment in breast cancer", Expert Systems with Applications, 2024 Publication	<1 %
44	Tanjim Mahmud, Michal Ptaszynski, Fumito Masui. "Exhaustive Study into Machine Learning and Deep Learning Methods for	<1 %

	Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts", Electronics, 2024 Publication	
45	deepai.org Internet Source	<1 %
46	epub.ub.uni-muenchen.de Internet Source	<1 %
47	www.frontiersin.org Internet Source	<1 %
48	www.hindawi.com Internet Source	<1 %
49	www.nature.com Internet Source	<1 %
50	"Advances and Applications of Artificial Intelligence & Machine Learning", Springer Science and Business Media LLC, 2023 Publication	<1 %
51	Gilbert Hinge, Mohamed A. Hamouda, Mohamed M. Mohamed. "Flash Flood Susceptibility Modelling Using Soft Computing-Based Approaches: From Bibliometric to Meta-Data Analysis and Future Research Directions", Water, 2024 Publication	<1 %
52	ir.juit.ac.in:8080 Internet Source	<1 %