الجمهـــوريـــــة الجزائــــريـــــة الديمقـــراطيـــــة الشــعبيــــة
People's Democratic Republic of Algeria
وزارة التـــــعلـــــيـــــــم العــــــــالــــــــي والبـــحـــــــث العــلـمـــــــي
Ministry of Higher Education and Scientific Research

جامعة وهران 2 محمد بن أ حمد
University of Oran 2 Mohamed Ben Ahmed

معهد الصيانة والأمن الصناعي
Institute of Maintenance and Industrial Safety

Department of Industrial Safety and Environment

To award the master's degree

Sector: Industrial Safety and Environment
Specialty: Security Prevention and Intervention

Theme

Feature Selection for Decision Making in Risk Assessment

Realized by:

AZZOUZ AIRECHE
AID MOHAMED FAROUK

Before the jury composed of:

First and Last name	Grade	Establishment	Quality
……………………………………....	………….	………………………....	**President**
ZOUAIRI SAIM	………….	IMSI	**Supervisor**
……………………………………....	………….	………………………....	**Examinator**

Year 2023/2024

Dedication

First of all, we thank God Almighty for giving us strength throughout our academic journey till we were able to complete it.

We want to thank our parents and family who encouraged us to learn and work.

Many thanks to our supervisor, Dr. Zouairi Saim who helped us in writing this thesis, and all the professors who taught us during our stages.

We are also grateful to our colleagues, friends, Kaggle community and everyone who was present throughout this period

I would like to acknowledge and dedicate this work to my beloved mother “Lalibi.K” who encouraged me the most to complete my studies may God protect her for me

-Aid Med Farouk

Abstract

Risk assessment plays a crucial role in various domains in our current time, including finance, healthcare, and more and there are many methods that can be used in this process, one of which we will explain in this work is feature selection.

Firstly, in this thesis we define the risk assessment, and its role in enhancing the safety and health of humans, environment and we talk about its five steps and how risks are ranked.

Then, we explain what the feature selection process is and how he is based on selecting the most consistent, relevant, and non-redundant features, and its importance in improving model interpretability, reducing computational complexity, and How to Choose a right Feature Selection Method for Machine Learning.

After that, we mention all feature selection methods in both supervised and unsupervised techniques and their effectiveness which are Filter, Wrapper, Hybrid, and Embedded Methods and how to use each one of them in machine learning.

Finally, we study the importance of using feature selection in risk assessment to improve safety by reducing the rate of accidents using the coding language Python and other tools such as PyCharm, Pandas and Scikit-learn and we take the data base from Kaggle of a Brazilian industry then we find out which features should be considered of, to assess the risks and enhance safety and health of workers.

Keywords: Risk assessment, Feature selection, Machine learning, feature selection methods (Filter, Wrapper, Hybrid, and Embedded), using feature selection in risk assessment to improve safety

يلعب تقييم المخاطر دورًا حاسمًا في مجالات مختلفة في وقتنا الحالي، بما في ذلك التمويل والرعاية الصحية والمزيد، وهناك العديد من الأساليب التي يمكن استخدامها في هذه العملية، أحدها سنشرحه في هذا العمل وهو اختيار الميزات.

أولاً في هذه الأطروحة، قمنا بتعريف تقييم المخاطر ودوره في تعزيز سلامة وصحة الإنسان والبيئة ونتحدث عن خطواته الخمس وكيفية تصنيف المخاطر.

ثم نوضح ما هي عملية اختيار الميزة وكيف أنها تعتمد على اختيار الميزات الأكثر اتساقًا وذات صلة وغير زائدة عن الحاجة. وأهميتها في تحسين إمكانية تفسير النموذج، وتقليل التعقيد الحسابي، وكيفية اختيار طريقة اختيار الميزة الصحيحة للتعلم الآلي.

بعد ذلك نذكر جميع طرق اختيار الميزات في كل من التقنيات الخاضعة للإشراف وغير الخاضعة للإشراف وفعاليتها وهي أساليب التصفية والالتفاف والهجين والمضمن وكيفية استخدام كل منها في التعلم الآلي.

وأخيراً قمنا بدراسة أهمية استخدام اختيار الميزات في تقييم المخاطر لتحسين السلامة من خلال تقليل معدل الحوادث وأخذنا قاعدة البيانات من Scikit-learnو Pandasو PyCharmوأدوات أخرى مثل Pythonباستخدام لغة البرمجة البرازيلية الصناعة ثم نكتشف الميزات التي يجب أخذها في الاعتبار لتقييم المخاطر وتعزيز سلامة وصحة Kaggle العمال.

الكلمات المفتاحية: تقييم المخاطر، اختيار الميزة، التعلم الآلي، طرق اختيار الميزات (أساليب التصفية والالتفاف والهجين والمضمن)، أهمية استخدام اختيار الميزات في تقييم المخاطر لتحسين السلامة.

Table of contents

GENERAL INTRODUCTION

Risk assessment is a critical process in the field of safety. Accurate risk prediction enables informed decision-making, resource allocation, and mitigation strategies. However, achieving reliable risk assessment models requires thoughtful consideration of the features (variables) used for prediction, in this context we delve into Feature selection.

Feature selection is a fundamental step in building robust risk assessment models. By carefully choosing relevant features, we enhance prediction accuracy and facilitate meaningful decision-making.

In risk assessment, features can range from historical data (such as past incidents) to environmental factors (such as temperature or pollution levels). These features serve as inputs to our models, shaping their predictions and influencing subsequent decisions.

In today's data-rich environment, we often encounter high-dimensional datasets with numerous features. While having abundant data is advantageous, it also poses challenges.

Despite the significance of Feature selection, it remains underutilized in risk assessment because many practitioners are unaware of Feature selection techniques and their impact.

We'll explore various Feature selection methods, how do we choose a feature, its crucial role, how and when do we apply it and its impact on the model stability and risk assessment and highlight the importance of Feature selection in assessing the risks.

We'll also at the end simulate the process with a software and using python programming language.

1 CHAPTER I: The methods and goals of Risk Assessment

1.1 Introduction

Conducting a risk assessment stands as a crucial measure in safeguarding both employees and the business, while also ensuring compliance with legal standards. This process directs attention to the significant hazards within the work environment those capable of inflicting genuine harm. Often, simple precautions suffice to manage risks effectively, such as promptly addressing spills to prevent slips or keeping cupboard drawers closed to mitigate tripping hazards. Individuals, including workers, deserve protection from potential harm resulting from inadequate safety measures. The repercussions of accidents and health issues extend beyond personal suffering, impacting business operations through lost productivity, equipment damage, increased insurance expenses, or legal proceedings. It is a legal obligation to evaluate workplace risks systematically and develop strategies for risk control.

1.2 Definitions

1.2.1 Hazard

Is anything with the potential to cause harm in terms of human injury or ill health, such as work materials, equipment, work methods or practices, poor work design or exposure to harmful agents such as chemicals, noise or vibration.

1.2.2 Risk

Is the likelihood that somebody will be harmed by the hazard and how serious the harm might be. When considering risk, you should also consider the number of people at risk from the hazard.

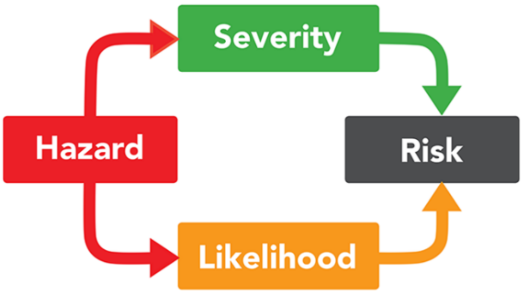


Figure 1.1:Relation between hazards and risks [2]

1.2.3 Control measures (or controls)

Are the precautions taken to ensure that a hazard will not injure anyone. When putting a control measure in place ensure that is does not create an additional hazard According to [1]
.

1.3 Everything related to Risk Assessment:

1.3.1 What is a Risk Assessment?

Risk assessment is the process of identifying hazards that could negatively affect an organization's ability to conduct business. These assessments help identify inherent business risks and prompt measures, processes and controls to reduce the impact of these risks on business operations.

Risk assessments help ensure the health and safety of employees and customers by identifying potential hazards. The goal of this process is to determine what measures should be implemented to mitigate those risks. For example, certain hazards or risks might determine the type of protective gear and equipment a worker needs.

Different industries present different types of hazards, and as such, risk assessments vary from industry to industry.

As a risk assessment is conducted, vulnerabilities and weaknesses that could make a business more hazardous are analyzed. Potential vulnerabilities could include construction deficiencies, security issues and process system errors. Companies can use a risk assessment framework (RAF) to prioritize and share the details of the assessment, including any risks to their IT infrastructure. The RAF helps an organization identify hazards and any business assets put at risk by these hazards, as well as potential fallout if these risks come to fruition. If a hazard has a large enough impact, then a mitigation strategy can be constructed.

In large enterprises, the chief risk officer or a chief risk manager usually conducts the risk assessment process, Risk assessments are also a major component of a risk analysis.

Risk analysis (as shown in Figure 1.2): a similar process of identifying and analyzing potential issues that could negatively affect key business initiatives or projects [3].

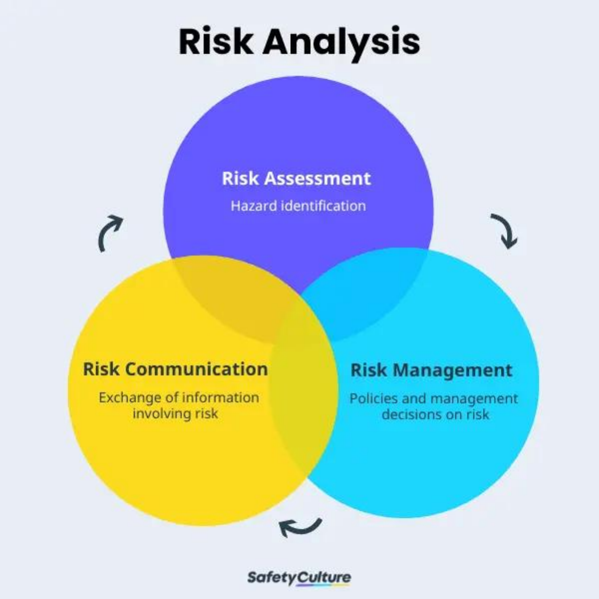


Figure 1.2: Risk Analysis Framework by SafetyCulture [4]

1.3.2 Why is risk assessment important?

A risk assessment ensures that you are able to identify all hazards in the workplace which may lead to an injury or illness. Once risks are identified, the business is then able to review the best measures to eliminate the risk completely or implement control measures to minimize the likelihood of an injury/illness occurring [5].

Risk assessments are highly important as they can assist to:

- create awareness of hazards and risks
- identify who may be at risk
- determine whether there are existing and adequate control measures in place

1.3.3 What is the purpose of a risk assessment?

The purpose of a risk assessment is to identify hazards in the workplace in order to implement control measures that can eliminate or minimize risks as much as possible. This, in turn, will help with providing a safer working environment.

Risk assessments should be completed in consultation with workers. This will assist in identifying hazards which may normally go unnoticed [5].

1.3.4 What are the benefits of a risk assessment?

The greatest benefit of a risk assessment is ensuring safety within your workplace. A comprehensive risk assessment may prevent, or in the very least minimize, workplace injuries or illnesses. Other benefits of risk assessments may include:

1.3.4.1. Cost-saving

The right control measures should reduce injuries or illnesses occurring in the workplace, which will cause a reduction in workers compensation claims, absenteeism, and reactive measures.

13.4.2. Employee loyalty

Placing a strong emphasis on risk assessments may convey to your employees that you take their safety seriously. This can result in greater loyalty from their end [5]

1.3.5 Three Types of risk assessment

While the exact details of risk assessments may vary greatly across different industries, HSE distinguishes three general risk assessment types [4]:

1.3.5.1. Large Scale Assessments

This refers to risk assessments performed for large scale complex hazard sites such as the nuclear, and oil and gas industry. This type of assessment requires the use of an advanced risk assessment technique called a Quantitative Risk Assessment (QRA).

1.3.5.2. Required specific assessments

This refers to assessments that are required under specific legislation or regulations, such as the handling of hazardous substances and manual handling.

1.3.5.3. General assessments

This type of assessment manages general workplace risks and is required under the management of legal health and safety administrations such as OSHA and HSE.

1.3.6 Examples of Risk Assessments:

There are various risk assessments used across different industries tailored to specific needs and control measures. Here are common examples of risk assessment:

Health and Safety Risk Assessment: A type of risk assessment used by safety managers to determine health and safety risks associated with the job, work environment, and current processes, in which hazards can be identified as biological, chemical, energy, environmental, and the like.

Workplace Risk Assessment : Performed by office managers
and school administrators, this tool helps ensure that a workplace is free from health and safety threats. This assessment also helps boost employee morale and productivity.
Fall Risk Assessment : Performed by the nursing staff of aged care units or centers to evaluate the possibility of falling and ensure that the facilities, equipment, and other factors are safe for elderly patients.

Construction Risk Assessment : A vital assessment used in the construction site to help safety teams implement corrective measures and stakeholders comply with safety regulations.
Manufacturing Risk Assessment :This type of risk assessment is conducted to evaluate the risk for manufacturing facilities, aiming to lay down mitigation strategies and prioritize urgent issues [6].

1.3.7 Who is responsible for carrying out risk assessments?

According to [7] it is the responsibility of the employer (or self-employed person) to carry out the risk assessment at work or to appoint someone with the relevant knowledge, experience and skills to do so.

The Management of Health and Safety at Work Regulations 1999 states that an employer must take reasonable steps 'for the effective planning, organization, control, monitoring and review of the preventive and protective measures.' So even if the task of risk management is delegated, it is ultimately the responsibility of the management within any business to ensure it is effectively completed.

Once hazards have been identified, the associated risks evaluated and steps taken to minimize the potential effects, the next step for an employer is to clearly and effectively communicate the risk assessment process and content to relevant parties.

The process of communication is more effectively achieved if the relevant persons are involved with the risk assessment process at every stage. The person carrying out an activity or task is often best placed to provide details on the associated hazards and risks and should participate fully in the completion of the risk assessment.

1.3.8 When to carry out a risk assessment?

A suitable and sufficient risk assessment must be carried out prior to a particular activity or task being carried out in order to eliminate, reduce or suitably control any associated risk to the health, safety and wellbeing of persons involved with (or affected by) the task/activity in question.

Once completed a risk assessment should be reviewed periodically (proportionate to the level of risk involved) and in any case when either the current assessment is no longer valid and/or if at any stage there has been significant changes to the specific activity or task.

Relevant risk assessments should be reviewed following an accident, incident or ill-health event in order to verify if the control measures and level of evaluated risk where appropriate or require amendment [7].

1.3.9 How to conduct a risk assessment?

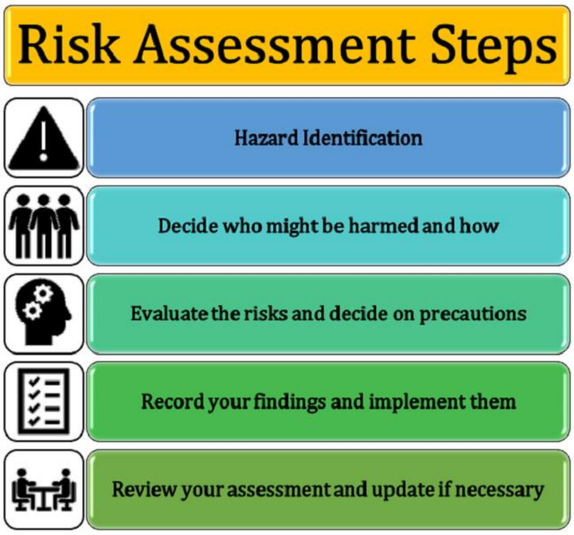

Risk Assessment Steps
Hazard Identification
Decide who might be harmed and how
Evaluate the risks and decide on precautions
Record your findings and implement them
Review your assessment and update if necessary

Figure 1.3:The 5 steps of Risk assessment [8]

According to [7] The HSE has recommended a five-step process for completing a risk assessment. This provides a useful checklist to follow to ensure that the assessment is suitably comprehensive. It involves:

- Identifying potential hazards
- Identifying who might be harmed by those hazards
- Evaluating risk (severity and likelihood) and establishing suitable precautions
- Implementing controls and recording findings

- Reviewing the assessment and re-assessing if necessary.

1.3.9.1. Identify potential hazards

It is important to firstly identify any potential hazards within a workplace that may cause harm to anyone that comes into contact with them. They may not always be obvious so some simple steps one can take to identify hazards are:

- Observation: Walking around the workplace and looking at what activities, tasks, processes or substances used could harm employees (or others)
- Looking back over past accidents and ill-health records as they may identify less obvious hazards
- Checking manufacturers' data sheets, instructions, information and guidance
- Consulting with employees (and others) who are carrying out the activities, tasks or processes.

It may be useful to group hazards into five categories, namely physical, chemical, biological, ergonomic and psychological.

1.3.9.2. Identify who might be harmed by those hazards

Next, identify who might be harmed by those potential hazards. It should also be noted how they could be affected, be it through direct contact or indirect contact. It is not necessary to list people by name, rather by identifying groups including:

- Employees
- Contractors

Some hazards may present a higher risk to certain groups including children, young people, new or expectant mothers, new employees, home workers, and lone workers.

1.3.9.3. Evaluate risk severity and establish precautions

After identifying any hazards and who might be affected, it is important to evaluate the severity the risk may present (should it occur) and establish suitable and effective controls to reduce this level of risk as far as is ‘reasonably practicable’. This means that everything possible is done to ensure health and safety considering all relevant factors including:

- Likelihood that harm may occur
- Severity of harm that may occur
- Knowledge about eliminating, reducing or controlling hazards and risks
- Availability of control measures designed to eliminate, reduce or suitably control or the risk
- Costs associated with available control measures designed to eliminate, reduce or suitably control or the risk

Assessing the severity of a risk requires an evaluation of the likelihood of an occurrence and how substantial the consequences that it may cause. Some factors affecting this evaluation include the duration and frequency of exposure, number of persons affected, competence of those exposed, the type of equipment and its condition, and availability of first-aid provision and/or emergency support.

1.3.9.4. Implement changes and record findings

If a workplace has five or more individuals, significate findings of the risk assessments are required to be kept either electronically or in writing. Recording findings on a risk assessment form is an easy way to keep track of the risks and control measures put in place to reduce the identified risk. The form includes:

- What hazards were found
- Person(s) or groups affected
- The controls put in place to manage risks and who is monitoring them
- Who carried out the assessment
- On what date the assessment was done.

It is sensible to ensure the risk assessment is proportionate to the activity or task being carried out and this can often be a straightforward process for generic tasks.

1.3.9.5. Review the assessment and reassess if necessary

Employers should periodically review the assessment and if necessary, re-assess any controls in place.

A good guide as to when one may need to review processes includes:

- After any significant change within the workplace or process in question
- After an accident or ill-health incident has occurred
- After near-misses have been reported.

Forgetting to review the risk assessment is easy, especially when trying to run a business. Don't wait until it's too late, set a date to review the risk assessment when conducting it and don't forget to add the date to the diary.

Significant changes can happen in businesses and when they do, make sure to review the risk assessment and amend it if needed. If the organization is planning changes that will happen in the future, ensure a risk assessment review is included.

1.4 Roles and responsibilities

Workers' safety and health is protected in Europe by an approach based on assessing and managing risks. In order to carry out effective workplace risk assessment, all those involved require a clear understanding of the legal context, concepts, the process of assessing the risks and the role to be played by the main actors involved in the process [9].

1.4.1 Workers' roles and responsibilities:

It is important that workers participate in the risk assessment. They know the problems and the details of what really happens when they perform their tasks or activities, so they should be involved in the assessment. Their practical knowledge or competence is also often needed to develop workable preventive measures.

Workers' participation is not only a right, but also fundamental to make the employers' occupational health and safety management effective and efficient.

Workers and/or their representatives have the right/duty to:

- be consulted on arrangements for the organization of the risk assessment and for the appointment of those undertaking the task.
- participate in the risk assessment.
- alert their supervisors or employers regarding perceived risks.
- report any changes in the workplace.
- be informed of the risks to their safety and health and of the measures necessary to eliminate or reduce these risks.
- be involved in the process of deciding on the preventive and protective measures to be put in place.
- ask the employer to put in place appropriate measures and to submit proposals to minimize hazards or to remove the danger at source.
- cooperate to help the employer to ensure that the working environment is safe.
- be trained/receive instructions on the measures to be put in place.
- take care as far as possible of their safety and health and that of other persons affected by their acts in accordance with the training and the instructions given by the employer.

In addition, it is important that workers' representatives are trained so that they understand risk assessment and their role in it.

1.4.2 Employers' roles and responsibilities

Employers should carefully prepare what they are going to do in order to meet their responsibilities to carry out a risk assessment and put in place the measures necessary for the safety and health of workers. It is recommended that they do this following the mentioned steps:

- commissioning, organizing and coordinating the assessment.
- appointing competent people to make the assessments; the person carrying out the risk assessment can be:
    - the employers themselves.
    - employees designated by the employers.

    - external assessors and service providers if there is a lack of competent personnel in the workplace.

- consulting workers' representatives on arrangements for the appointment of those who will make the assessments.
- providing the necessary information, training, resources and support to assessors who are the employer's own employees.
- ensuring adequate coordination between assessors (where relevant).
- involving management and encouraging the participation of the workforce.
- determining the arrangements to be made for reviewing and revising the risk assessment.
- ensuring that the preventive and protective measures take account of the results of the assessment.
- ensuring that the risk assessment is documented.
- monitoring the protective and preventive measures to ensure that their effectiveness is maintained.
- informing workers and/or their representatives of the results of the assessment and of the measures introduced (making the records available to them).

1.5 Analysis of Relevant Risks

According to [10] risk analysis is the phase where the level of the risk and its nature are assessed and understood. This information is the first input to decision makers on whether risks need to be treated or not and what is the most appropriate and cost-effective risk treatment methodology.

Risk analysis involves:

- thorough examination of the risk sources.
- their positive and negative consequences.
- the likelihood that those consequences may occur and the factors that affect them.
- assessment of any existing controls or processes that tend to minimize negative risks or enhance positive risks (these controls may derive from a wider set of standards,

controls or good practices selected according to an applicability statement and may also come from previous risk treatment activities.)

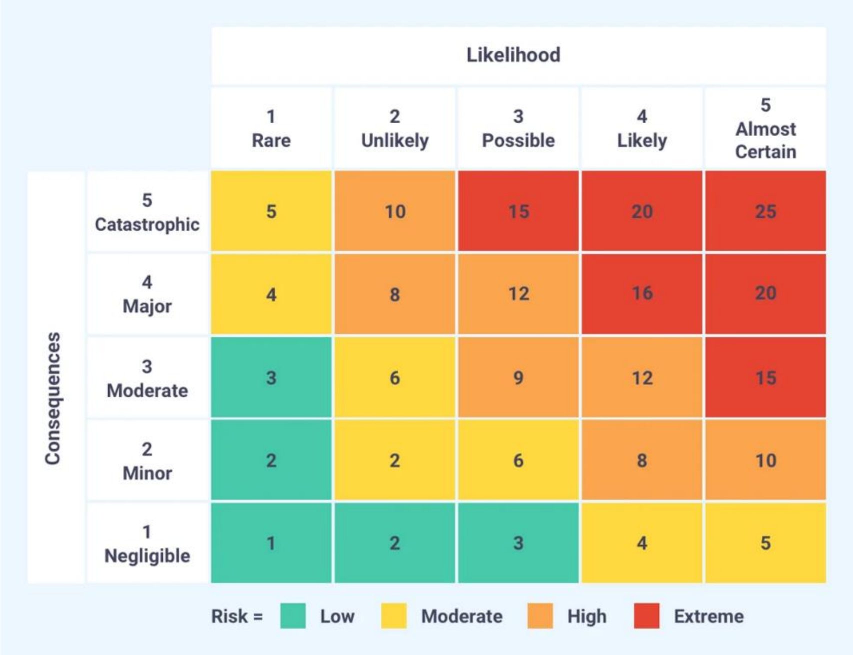
The level of risk can be estimated by using statistical analysis and calculations combining impact and likelihood. Any formulas and methods for combining them must be consistent with the criteria defined when establishing the Risk Management context. This is because an event may have multiple consequences and affect different objectives, therefore consequences and likelihood need to be combined to calculate the level of risk. If no reliable or statistically reliable and relevant past data is available (kept for e.g. an incident database), other estimates may be made as long as they are appropriately communicated and approved by the decision makers.

Information used to estimate impact and likelihood usually comes from:

- past experience or data and records (e.g. incident reporting)
- reliable practices, international standards or guidelines
- market research and analysis
- experiments and prototypes
- economic, engineering or other models
- specialist and expert advice

Risk analysis techniques include

- interviews with experts in the area of interest and questionnaires
- use of existing models and simulations



		Likelihood				
		1 Rare	2 Unlikely	3 Possible	4 Likely	5 Almost Certain
Consequences	5 Catastrophic	5	10	15	20	25
	4 Major	4	8	12	16	20
	3 Moderate	3	6	9	12	15
	2 Minor	2	2	6	8	10
	1 Negligible	1	2	3	4	5

Risk = 🟩 Low 🟨 Moderate 🟧 High 🟥 Extreme

Figure 1.4:Qualitative and quantitative risk analysis [11]

Risk analysis may vary in detail according to the risk, the purpose of the analysis, and the required protection level of the relevant information, data and resources. Analysis may be qualitative, semi-quantitative or quantitative or a combination of these. In any case, the type of analysis performed should, as stated above, be consistent with the criteria developed as part of the definition of the Risk Management context.

A short description of the above-mentioned types of analysis types is as follows:

1.5.1 Qualitative analysis

In qualitative analysis, the magnitude and likelihood of potential consequences are presented and described in detail. The scales used can be formed or adjusted to suit the circumstances, and different descriptions may be used for different risks.

Qualitative analysis may be used:

- as an initial assessment to identify risks which will be the subject of further, detailed analysis.
- where non-tangible aspects of risk are to be considered (e.g. reputation, culture, image etc.)
- where there is a lack of adequate information and numerical data or resources necessary for a statistically acceptable quantitative approach.

1.5.2 Semi-quantitative analysis

In semi-quantitative analysis the objective is to try to assign some values to the scales used in the qualitative assessment. These values are usually indicative and not real, which is the prerequisite of the quantitative approach.

Therefore, as the value allocated to each scale is not an accurate representation of the actual magnitude of impact or likelihood, the numbers used must only be combined using a formula that recognizes the limitations or assumptions made in the description of the scales used.

It should be also mentioned that the use of semi-quantitative analysis may lead to various inconsistencies due to the fact that the numbers chosen may not properly reflect analogies between risks, particularly when either consequences or likelihood are extreme.

1.5.3 Quantitative analysis

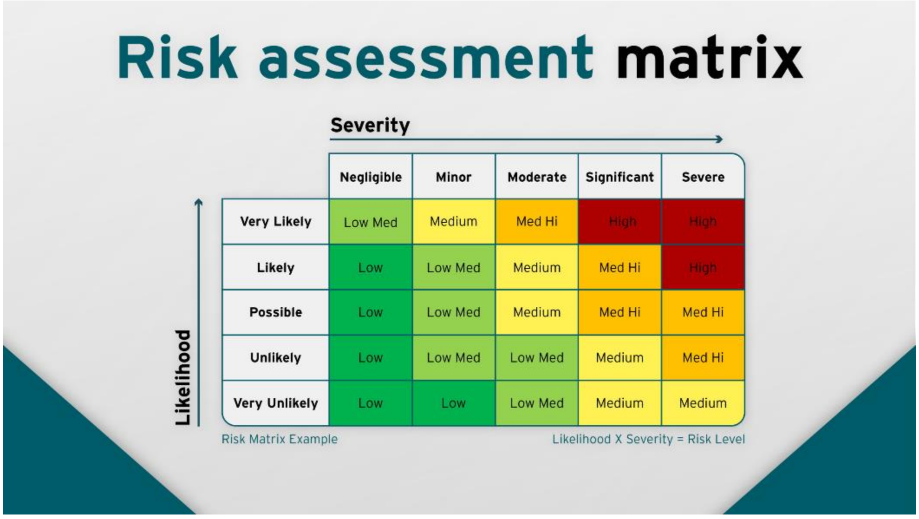
In quantitative analysis numerical values are assigned to both impact and likelihood. These values are derived from a variety of sources. The quality of the entire analysis depends on the accuracy of the assigned values and the validity of the statistical models used.

Impact can be determined by evaluating and processing the various results of an event or by extrapolation from experimental studies or past data. Consequences may be expressed in various terms of:

- Monetary
- Technical
- Operational
- Human
- Impact criteria.

As it is made clear from the above analysis, the specification of the risk level is not unique. Impact and likelihood may be expressed or combined differently, according to the type of risk and the scope and objective of the Risk Management process.

1.6 How are risks ranked or prioritized?



Risk assessment matrix

Severity

Likelihood

	Negligible	Minor	Moderate	Significant	Severe
Very Likely	Low Med	Medium	Med Hi	High	High
Likely	Low	Low Med	Medium	Med Hi	High
Possible	Low	Low Med	Medium	Med Hi	Med Hi
Unlikely	Low	Low Med	Low Med	Medium	Med Hi
Very Unlikely	Low	Low	Low Med	Medium	Medium

Risk Matrix Example

Likelihood X Severity = Risk Level

Figure 1.5: Risk assessment matrix

According to [12] ranking or prioritizing hazards is one way to help determine which risk is the most serious and thus which to control first. Priority is usually established by taking into account the employee exposure and the potential for incident, injury or illness. By assigning a priority to the risks, you are creating a ranking or an action list.

There is no one simple or single way to determine the level of risk. Nor will a single technique apply in all situations. The organization has to determine which technique will work best for each situation. Ranking hazards requires the knowledge of the workplace activities, urgency of situations, and most importantly, objective judgement.

For simple or less complex situations, an assessment can literally be a discussion or brainstorming session based on knowledge and experience. In some cases, checklists or a probability matrix can be helpful. For more complex situations, a team of knowledgeable personnel who are familiar with the work is usually necessary.

As an example, consider this simple risk matrix. Table 1 shows the relationship between probability and severity.

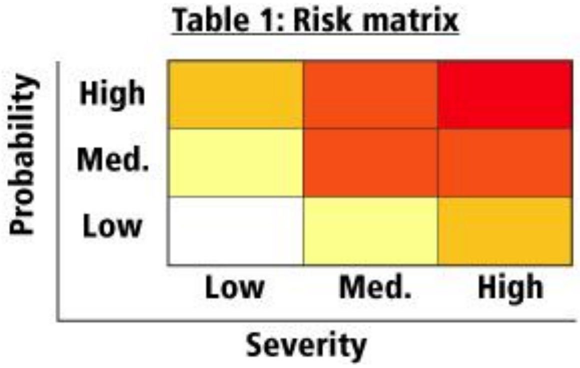


Table 1: Risk matrix

Figure 1.6:table (1) shows relationship between probability and severity

Severity ratings in this example represent:

- High: major fracture, poisoning, significant loss of blood, serious head injury, or fatal disease
- Medium: sprain, strain, localized burn, dermatitis, asthma, injury requiring days off work
- Low: an injury that requires first aid only; short-term pain, irritation, or dizziness

Probability ratings in this example represent:

- High: likely to be experienced once or twice a year by an individual
- Medium: may be experienced once every five years by an individual
- Low: may occur once during a working lifetime

The cells in Table 1 correspond to a risk level, as shown in Table 2.

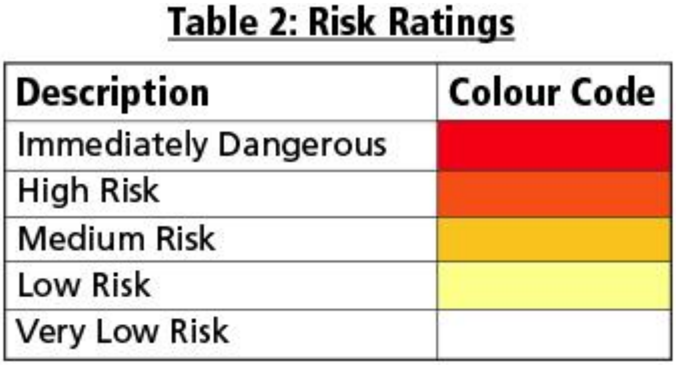


Table 2: Risk Ratings

Description	Colour Code
Immediately Dangerous	🟥
High Risk	🟧
Medium Risk	🟨
Low Risk	▨
Very Low Risk	⬜

Figure 1.7:table (2) shows risk ratings

These risk ratings correspond to recommended actions such as:

- Immediately dangerous: stop the process and implement controls
- High risk: investigate the process and implement controls immediately

- Medium risk: keep the process going; however, a control plan must be developed and should be implemented as soon as possible
- Low risk: keep the process going but monitor regularly. A control plan should also be investigated
- Very low risk: keep monitoring the process

Let's use an example: When painting a room, a step stool must be used to reach higher areas. The individual will not be standing higher than 1 meter (3 feet) at any time. The assessment team reviewed the situation and agreed that working from a step stool at 1 m is likely to:

- Cause a short-term injury such as a strain or sprain if the individual falls. A severe sprain may require days off work. This outcome is similar to a medium severity rating.
- Occur once in a working lifetime as painting is an uncommon activity for this organization. This criterion is similar to a low probability rating.

When compared to the risk matrix chart (Table 1), these values correspond to a low risk.

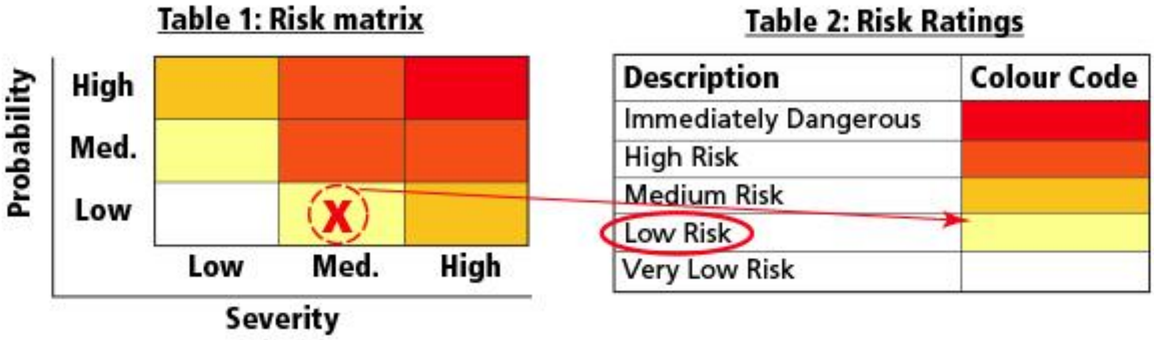


Table 1: Risk matrix

Probability	Low	Med.	High
High	🟨	🟧	🟥
Med.	▨	🟧	🟧
Low	⬜	X	🟨

Severity

Table 2: Risk Ratings

Description	Colour Code
Immediately Dangerous	🟥
High Risk	🟧
Medium Risk	🟨
Low Risk	▨
Very Low Risk	⬜

Figure 1.8:comparing Table (1) with table (2)

The workplace decides to implement risk control measures, including the use of a stool with a large top that will allow the individual to maintain stability when standing on the stool. They also determined that while the floor surface is flat, they provided training to the individual on the importance of making sure the stool's legs always rest on the flat surface. The training also included steps to avoid excess reaching while painting.

2 Chapter II: Explanation of Feature selection

2.1 Feature Selection definition

All machine learning workflows depend on feature engineering, which comprises feature extraction and feature selection that are fundamental building blocks of modern machine learning pipelines. Despite the fact that feature extraction and feature selection processes share some overlap, often, these terms are erroneously equated. Feature extraction is the process of using domain knowledge to extract new variables from raw data that make machine learning algorithms work. The feature selection process is based on selecting the most consistent, relevant, and non-redundant features.

The objectives of feature selection techniques include:

- simplification of models to make them easier to interpret by researchers/users
- shorter training times
- avoiding the curse of dimensionality
- enhanced generalization by reducing overfitting (formally, reduction of variance)

Dataset size reduction is more important nowadays because of the plethora of developed analysis methods that are at the researcher's disposal, while the size of an average dataset keeps growing both with respect to the number of features and samples [13].

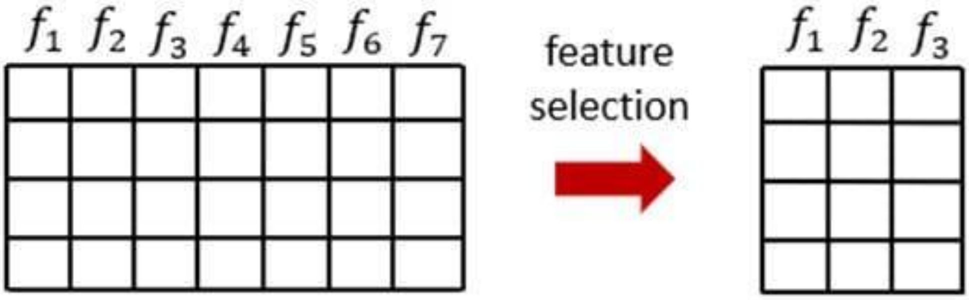


Figure 2.9:Illustration of feature selection and data size reduction in tabular data

2.2 What Makes Some Feature Representations Better Than Others?

Regardless of the technological approach to feature representation, there is a common question that haunts data scientists in most machine learning workflows: What makes some feature representations better than others?

This might seem like an insane question considering modern machine learning problems are using hundreds of thousands or even millions of features that are impossible to interpret by domain experts.

While there is no trivial answer to our target questions, there are some general principles that we can follow. In general, there are three key desired properties in feature representations:

- Disentangling of causal factors
- Easy to model
- Works well with regularization strategies

Fundamentally, solid representations include features that correspond to the underlying causes of the observed data. More specifically, this thesis links the quality of representations to structures in which different features and directions correspond to different causes in the underlying dataset so that the representation is able to disentangle one cause from another.

Another leading indicator of good representation is the simplicity of modelling. For a given machine learning problem/dataset, we can find many representations that separate the underlying causal factors, but they could be brutally hard to model [13].

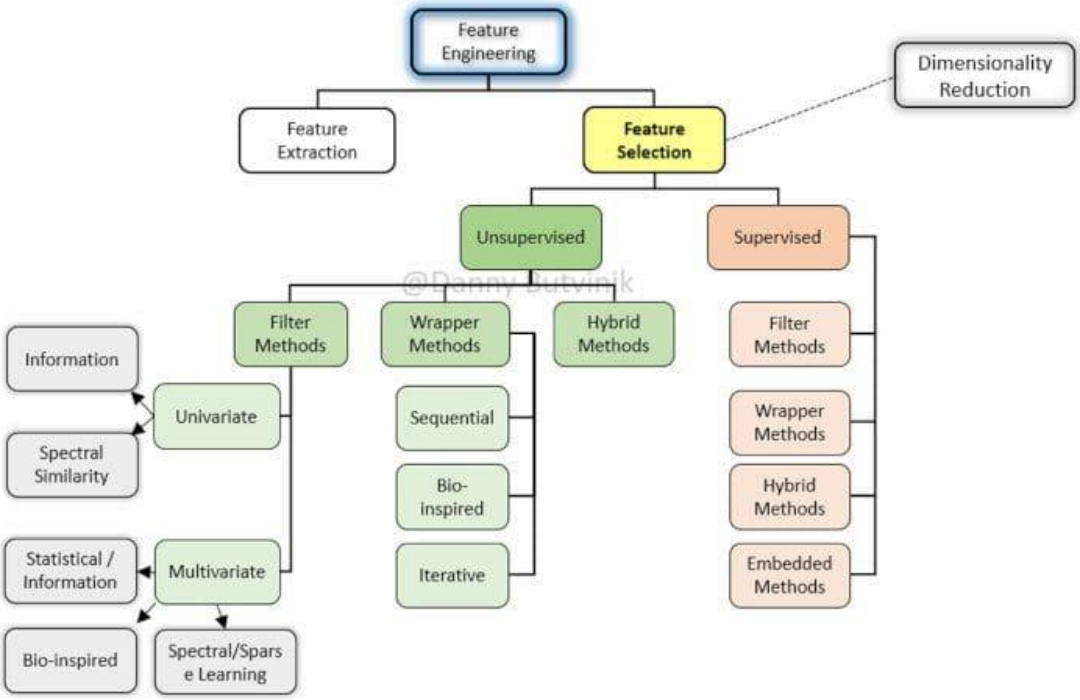


Figure 2.10:High-level Taxonomy for feature selection

2.3 Feature selection techniques

2.3.1 Supervised Techniques

These techniques can be used for labelled data and to identify the relevant features for increasing the efficiency of supervised models like classification and regression. For Example- linear regression, decision tree, SVM, etc.

2.3.2 Unsupervised Techniques

These techniques can be used for unlabeled data. For Example- K-Means Clustering, Principal Component Analysis, Hierarchical Clustering, etc.

From a taxonomic point of view, these techniques are classified into filter, wrapper, embedded, and hybrid methods.

Now, let's discuss some of these popular machine learning feature selection methods in detail according to [14].

2.4 When Features Are Born

Feature selection research dates back to the 1960s. Hughes used a general parametric model to study the accuracy of a Bayesian classifier as a function of the number of features. He concludes: "measurement selection, reduction and combination are not proposed as developed techniques. Rather, they are illustrative of a framework for further investigation." Since then, the research in feature selection has been a challenging field despite the skepticism of some researchers, for example, in a discussion of Miller, R.L. Plackett stated: "If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems." Nowadays, it is well known that feature selection plays a crucial role in some real problems to reduce their dimensionality, and there are numerous books and papers regarding this issue and their number is continuously growing [15].

2.5 Reasons why we need feature selection

A popular claim is that modern machine learning techniques do well without feature selection. After all, a model should be able to learn that particular features are useless, and it should focus on the others according to [16].

Well, this reasoning makes sense to some extent. Linear models could, in theory, assign a weight of zero to useless features, and tree-based models should learn quickly not to make splits on them. In practice, however, many things can go wrong with training when the inputs are irrelevant or redundant – more on these two terms later. On top of this, there are many other reasons why simply dumping all the available features into the model might not be a good idea. Let's look at the seven most prominent ones.

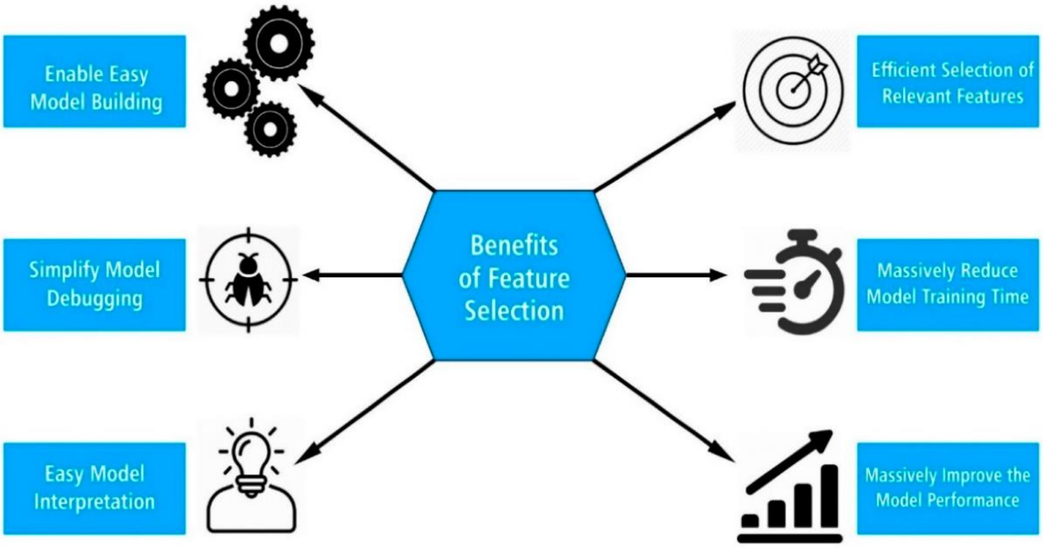

Enable Easy Model Building
Simplify Model Debugging
Easy Model Interpretation
Benefits of Feature Selection
Efficient Selection of Relevant Features
Massively Reduce Model Training Time
Massively Improve the Model Performance

Figure 2.11:The benefits of feature selection [17]

2.5.1 Irrelevant and redundant features

Some features might be irrelevant to the problem at hand. This means they have no relation with the target variable and are completely unrelated to the task the model is designed to solve. Discarding irrelevant features will prevent the model from picking up on spurious correlations it might carry, thus fending off overfitting.

Redundant features are a different animal, though. Redundancy implies that two or more features share the same information, and all but one can be safely discarded without information loss. Note that an important feature can also be redundant in the presence of another relevant feature. Redundant features should be dropped, as they might pose many problems during training, such as multicollinearity in linear models.

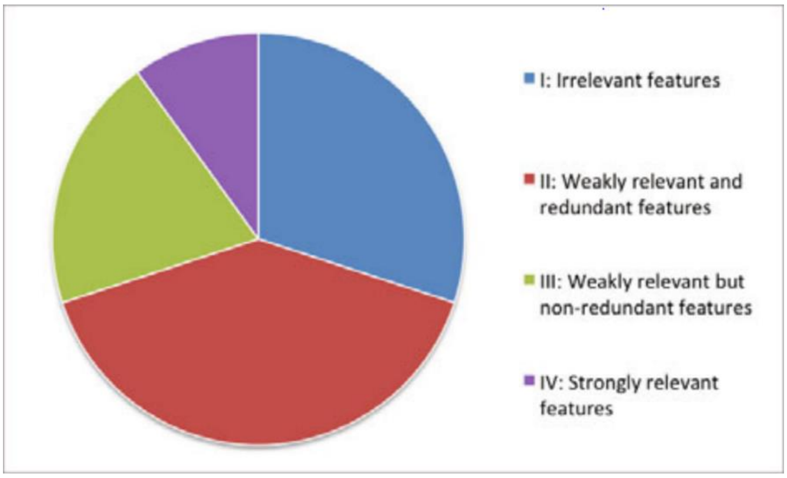


Figure 2.4: Overview of feature relevance and redundancy [15]

2.5.2 Curse of dimensionality

Feature selection techniques are especially indispensable in scenarios with many features but few training examples. Such cases suffer from what is known as the curse of dimensionality: in a very high-dimensional space, each training example is so far from all the other examples that the model cannot learn any useful patterns. The solution is to decrease the dimensionality of the features space, for instance, via feature selection.

2.5.3 Training time

The more features, the more training time. The specifics of this trade-off depend on the particular learning algorithm being used, but in situations where retraining needs to happen in real-time, one might need to limit oneself to a couple of best features.

2.5.4 Deployment effort

The more features, the more complex the machine learning system becomes in production. This poses multiple risks, including but not limited to high maintenance effort, entanglement, undeclared consumers, or correction cascades.

2.5.5 Interpretability

With too many features, we lose the explainability of the model. While not always the primary modelling goal, interpreting and explaining the model's results are often important and, in some regulated domains, might even constitute a legal requirement.

2.5.6 Occam's Razor

According to this so-called law of parsimony, simpler models should be preferred over the more complex ones as long as their performance is the same. This also has to do with the machine learning engineer’s nemesis, overfitting. Less complex models are less likely to overfit the data.

2.5.7 Data-model compatibility

Finally, there is the issue of data-model compatibility. While, in principle, the approach should be data-first, which means collecting and preparing high-quality data and then choosing a model which works well on this data, real life may have it the other way around.

You might be trying to reproduce a particular research paper, or your boss might have suggested using a particular model. In this model-first approach, you might be forced to select features that are compatible with the model you set out to train. For instance, many models don’t work with missing values in the data. Unless you know your imputation methods well, you might need to drop the incomplete features [16].

2.6 Challenges in Feature Selection

2.6.1 High dimensionality

With increasing amounts of data, the number of features can quickly become unwieldy, leading to the curse of dimensionality. Feature selection helps to reduce the number of features, making it easier to model the data.

2.6.2 Correlated features

In many datasets, features can be highly correlated, leading to multicollinearity issues. This can make it difficult to determine which features are most important, as they may all have a similar effect on the outcome.

2.6.3 Overfitting

Feature selection can sometimes lead to overfitting, where a model is fit too closely to the training data, leading to poor generalization to new data. This can be mitigated by using techniques such as cross-validation or regularization.

2.6.4 Loss of information

Removing features can sometimes result in the loss of important information or signal. Careful consideration must be given to the features selected to ensure that relevant information is not discarded.

2.6.5 Time complexity

Feature selection can be computationally expensive, particularly when dealing with large datasets and complex algorithms. This can limit the ability to scale and make feature selection a bottleneck in the modelling process [18].

2.7 How to Choose a Feature Selection Method for Machine Learning

Statistics for Filter-Based Feature Selection Methods

According to [19] in Data Preparation,

It is common to use correlation type statistical measures between input and output variables as the basis for filter feature selection.

As such, the choice of statistical measures is highly dependent upon the variable data types.

Common data types include numerical (such as height) and categorical (such as a label), although each may be further subdivided such as integer and floating point for numerical variables, and Boolean, ordinal, or nominal for categorical variables.

Common input variable data types:

- Numerical Variables:
    - Integer Variables.
    - Floating Point Variables.
- Categorical Variables:
    - Boolean Variables (dichotomous).
    - Ordinal Variables.
    - Nominal Variables.

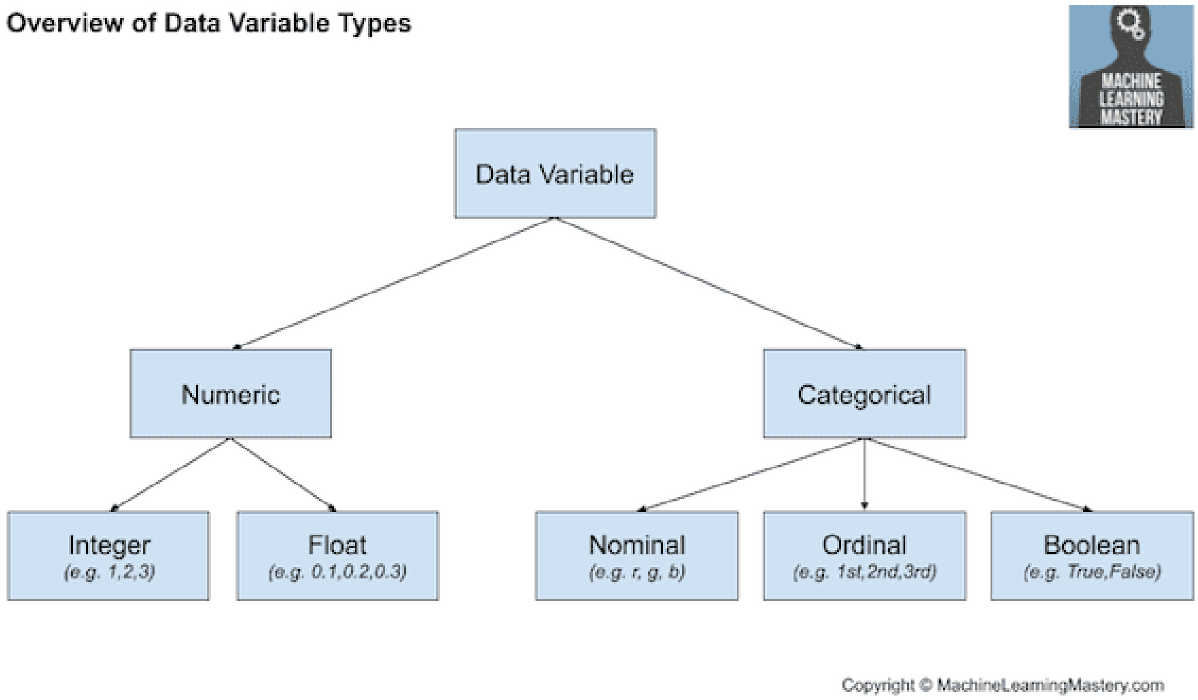


Figure 2.5: Overview of data variable types

The more that is known about the data type of a variable, the easier it is to choose an appropriate statistical measure for a filter-based feature selection method.

In this section, we will consider two broad categories of variable types: numerical and categorical; also, the two main groups of variables to consider: input and output.

Input variables are those that are provided as input to a model. In feature selection, it is this group of variables that we wish to reduce in size. Output variables are those for which a model is intended to predict, often called the response variable.

The type of response variable typically indicates the type of predictive modelling problem being performed. For example, a numerical output variable indicates a regression predictive modelling problem, and a categorical output variable indicates a classification predictive modelling problem.

- **Numerical Output:** Regression predictive modelling problem.
- **Categorical Output:** Classification predictive modelling problem.

The statistical measures used in filter-based feature selection are generally calculated one input variable at a time with the target variable. As such, they are referred to as univariate statistical measures. This may mean that any interaction between input variables is not considered in the filtering process.

Most of these techniques are univariate, meaning that they evaluate each predictor in isolation. In this case, the existence of correlated predictors makes it possible to select important, but redundant, predictors. The obvious consequences of this issue are that too many predictors are chosen and, as a result, collinearity problems arise.

With this framework, let's review some univariate statistical measures that can be used for filter-based feature selection.

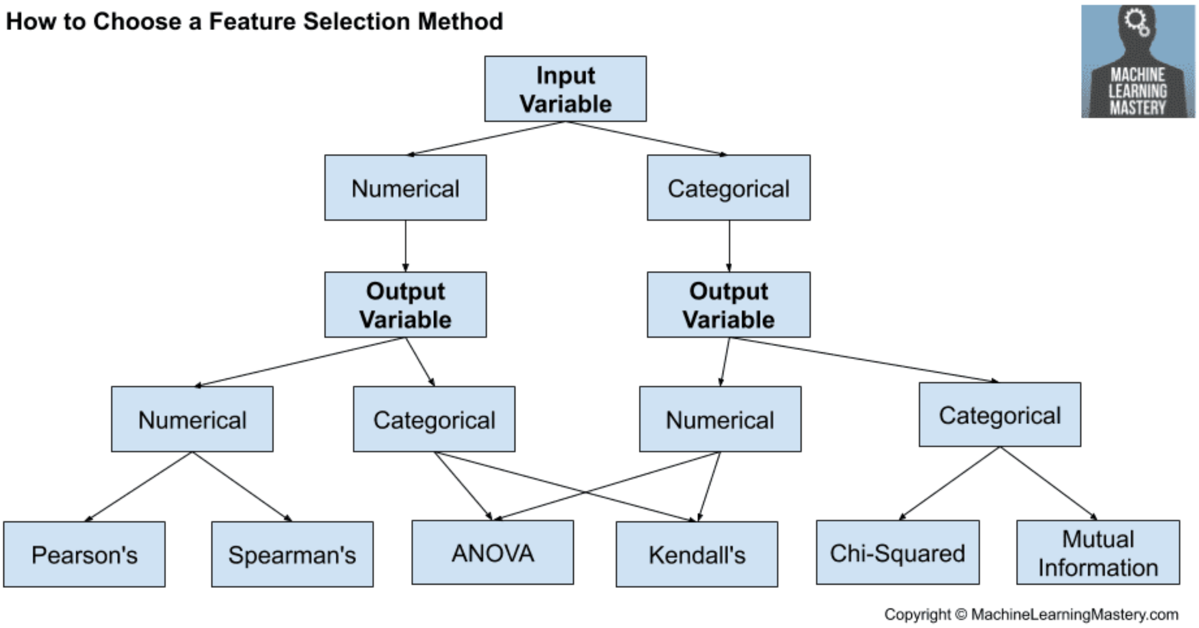


Figure 2.6: How to choose feature selection methods for Machine Learning

Numerical Input, Numerical Output

This is a regression predictive modelling problem with numerical input variables.

The most common techniques are to use a correlation coefficient, such as Pearson's for a linear correlation, or rank-based methods for a nonlinear correlation.

- Pearson's correlation coefficient (liner).
- Spearman's Rank coefficient (nonlinear)

Numerical Input, Categorical Output

This is a classification predictive modelling problem with numerical input variables.

This might be the most common example of a classification problem,

Again, the most common techniques are correlation based, although in this case, they must take the categorical target into account.

- ANOVA correlation coefficient (linear).
- Kendall’s rank coefficient (nonlinear).

Kendall does assume that the categorical variable is ordinal.

Categorical Input, Numerical Output

This is a regression predictive modelling problem with categorical input variables.

This is a strange example of a regression problem (e.g. you would not encounter it often).

Nevertheless, you can use the same “*Numerical Input, Categorical Output*” methods (described above), but in reverse.

Categorical Input, Categorical Output

This is a classification predictive modelling problem with categorical input variables.

The most common correlation measure for categorical data is the chi-squared test. You can also use mutual information (information gain) from the field of information theory.

- Chi-Squared test (contingency tables).
- Mutual Information.

In fact, mutual information is a powerful method that may prove useful for both categorical and numerical data, e.g. it is agnostic to the data types.

2.8 Feature Selection by Label Information:

In terms of availability of label information, the feature selection technique can be roughly classified into three families: supervised methods, semi-supervised methods, and unsupervised methods.

The availability of label information allows supervised feature selection algorithms to effectively select discriminative and relevant features to distinguish samples from different classes. There are supervised feature selection algorithms that identify the relevant features for best achieving the goal of the supervised model (e.g. classification or a regression problem) and they rely on the availability of labelled data.

Unsupervised feature selection techniques ignore the target variable, such as methods that remove redundant variables using correlation. Supervised feature selection techniques use the target variable, such as methods that remove irrelevant variables.

When a small portion of data is labelled, we can utilize semi-supervised feature selection which can take advantage of both labelled data and unlabeled data. Most of the existing semi-supervised feature selection algorithms rely on the construction of the similarity matrix and select those features that best fit the similarity matrix. Due to the absence of labels that are used for guiding the search for discriminative features, unsupervised feature selection is considered a much harder problem [20].

3 CHAPTER III: Feature selection methods

3.1 Feature selection methods:

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.

Some predictive modelling problems have a large number of variables that can slow the development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

Many models, especially those based on regression slopes and intercepts, will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model.

One way to think about feature selection methods are in terms of supervised and unsupervised methods.

An important distinction to be made in feature selection is that of supervised and unsupervised methods. When the outcome is ignored during the elimination of predictors, the technique is unsupervised [19].

3.2 Supervised Feature Selection Methods:

Supervised feature selection methods are classified into four types, based on the interaction with the learning model as Filter, Wrapper, Hybrid, and Embedded Methods [20].

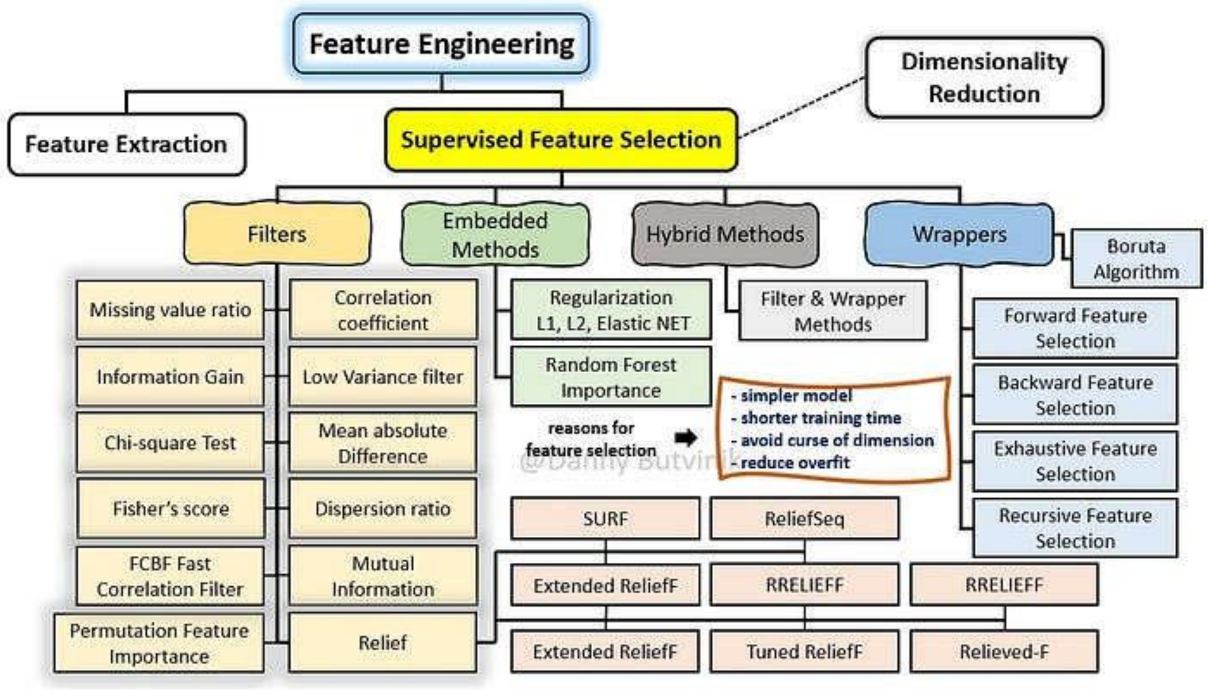


Figure 3.1: Extended taxonomy of supervised feature selection methods and techniques

3.2.1 Filter method:

Feature selection using filter method is made by using some information, distance, or correlation measures. Here, the features' sub-setting is generally done using one of the statistical measures like the Chi-square test, information gain, ANOVA test, or correlation coefficient. These help in selecting the attributes that are highly correlated with the target variable. Here, we work on the same model by changing the features [21]

According to [20] in the Filter method, features are selected based on statistical measures. It is independent of the learning algorithm and requires less computational time. Information gain, chi-square test, Fisher score, correlation coefficient, and variance threshold are some of the statistical measures used to understand the importance of the features.

Filter methodology uses the selected metric to identify irrelevant attributes and also filter out redundant columns from the models. It gives the option of isolating selected measures that enrich a model. The columns are ranked following the calculation of the feature scores.

By choosing and implementing the right features, the accuracy and efficiency of classification models potentially can be improved.

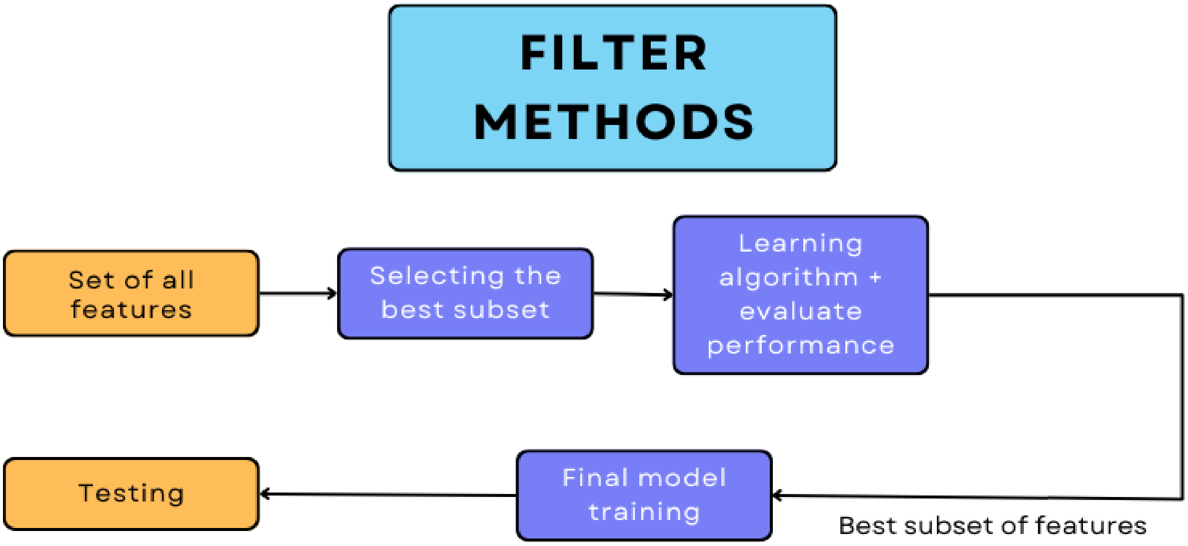


Figure 3.2: Filter method

3.2.1.1. Information gain:

Calculation of reduction in entropy from the transform. It evaluates the information gain of each variable in the context of the target variable.

3.2.1.2. Chi-square Test:

Applied on categorical features; it is calculated between each feature and selects features with the best Chi-squared score. There are three conditions: (1) feature must be categorical (2) sampled independently (3) expected frequency of the feature above 5.

3.2.1.3. Fisher's Score:

The features selected such that in the data space spanned by the selected features, distances between samples in different classes are as large as possible, while the distances between samples in the same class are as small as possible; Features with high quality should assign similar values to instances in the same class and different values to instances from different classes [20].

3.2.1.4. Correlation Coefficient:

Features should be correlated with the target but should be uncorrelated among themselves.

3.2.1.5. Low Variance Filter:

Low variance filter removes all zero-variance (that have the same value in all samples) features by default. The assumption is that features with higher variance may contain more useful information. Relationships between feature or feature and target variables are not taken into account, which is one of the drawbacks of this filter method.

3.2.1.6. Mean Absolute Difference (MAD):

The higher the MAD, the higher the discriminatory power.

3.2.1.7. Dispersion Ratio:

Dispersion ratio is the ratio between the arithmetic mean and geometric mean AM/GM. Higher dispersion implies a higher value of AM/GM, thus a more relevant feature.

3.2.1.8. Permutation Feature Importance:

Permutation feature importance is a model inspection technique that can be used for any fitted estimator when the data is tabular. This is especially useful for non-linear or opaque estimators. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. This technique benefits from being model agnostic and can be calculated many times with different permutations of the feature.

3.2.1.9. Missing value ratio:

Missing value ratio equals the number of missing values divided by a total number of observations and multiplied by 100 per column; decisions need to be taken on the threshold and drop all features that have a missing value ratio more than this threshold.

$$Ratio\ of\ missing\ values = \frac{Number\ of\ missing\ values}{Total\ number\ of\ observations} \times 100$$

3.2.1.10. Fast Correlation Based Filter (FCBF):

Fast Correlation Based Filter (FCBF) algorithm is a Filter method that uses the idea of ‘predominant correlation’. It selects features with high correlation with the target variable, but little correlation to other features.

Notably, the correlation used here is not the typical Pearson, Kendall, or Spearman you may be used commonly. It's known as Symmetrical Uncertainty (SU), which is based on information theory, drawing from the concepts of Shannon Entropy and Information Gain.

The algorithm initially selects features correlated above a given threshold by SU with the class variable. After this initial filtering, it detects predominant correlations of features with the class. The definition is that for a predominant feature, for example, feature X, no other feature is more correlated to X than X is correlated to Y.

3.2.1.10.1. Methodology:

- Creates groups of the features as per different criteria.
- Creates a benchmark for each group.
- Tests the correlation of features inside the group, compared to the predetermined group benchmark.
- Keeps only the features that are less correlated to each other than to the group benchmark.

3.2.1.11. Relief:

Relief's filter method approach is notably sensitive to feature interactions. It was originally designed for application to binary classification problems with discrete or numerical features. Relief calculates a feature score for each feature which can then be applied to rank and select top-scoring features for feature selection.

Alternatively, these scores may be applied as feature weights to guide downstream modelling. Relief feature scoring is based on the identification of feature value differences between nearest-neighbor instance pairs. If a feature value difference is observed in a neighboring instance pair with the same class (a 'hit'), the feature score decreases.

Alternatively, if a feature value difference is observed in a neighboring instance pair with different class values (a 'miss'), the feature score increases. Original Relief algorithm has since inspired a family of Relief-based feature selection algorithms (RBAs).

RBAs have been adapted to:

- Perform more reliably in noisy problems.
- Generalize to multi-class problems.
- Generalize to numerical outcome (i.e. regression) problems.
- Make them robust to incomplete (i.e. missing) data [20].

- **Reasons for choosing the filter method:**
  - It does not rely on the model's bias and instead depends only on the characteristics of the data. Hence, the same feature subset can be used to train different algorithms.
  - The time taken by information or distance-related measures is very; hence, a filter method can produce subsets faster.
  - They can handle large amounts of data [21]

3.2.2 Wrapper method:

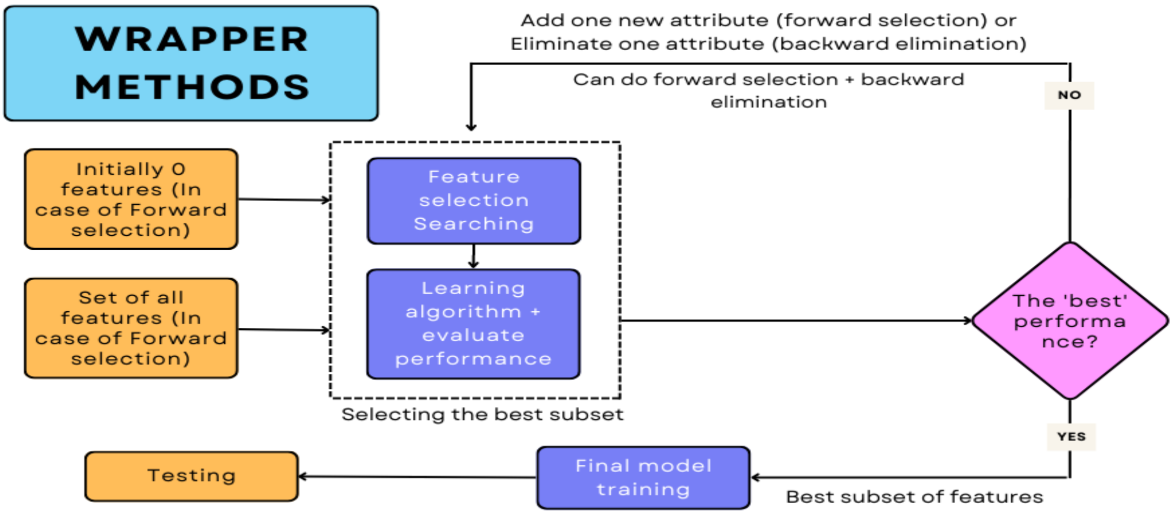

WRAPPER METHODS
Add one new attribute (forward selection) or Eliminate one attribute (backward elimination)
Can do forward selection + backward elimination
NO
Initially 0 features (In case of Forward selection)
Set of all features (In case of Forward selection)
Feature selection Searching
Learning algorithm + evaluate performance
Selecting the best subset
The 'best' performa nce?
YES
Testing
Final model training
Best subset of features

Figure 3.3: Wrapper method

According to [20] the Wrapper methodology considers the selection of feature sets as a search problem, where different combinations are prepared, evaluated, and compared to other combinations. A predictive model is used to evaluate a combination of features and assign model performance scores.

The performance of the Wrapper method depends on the classifier. The best subset of features is selected based on the results of the classifier.

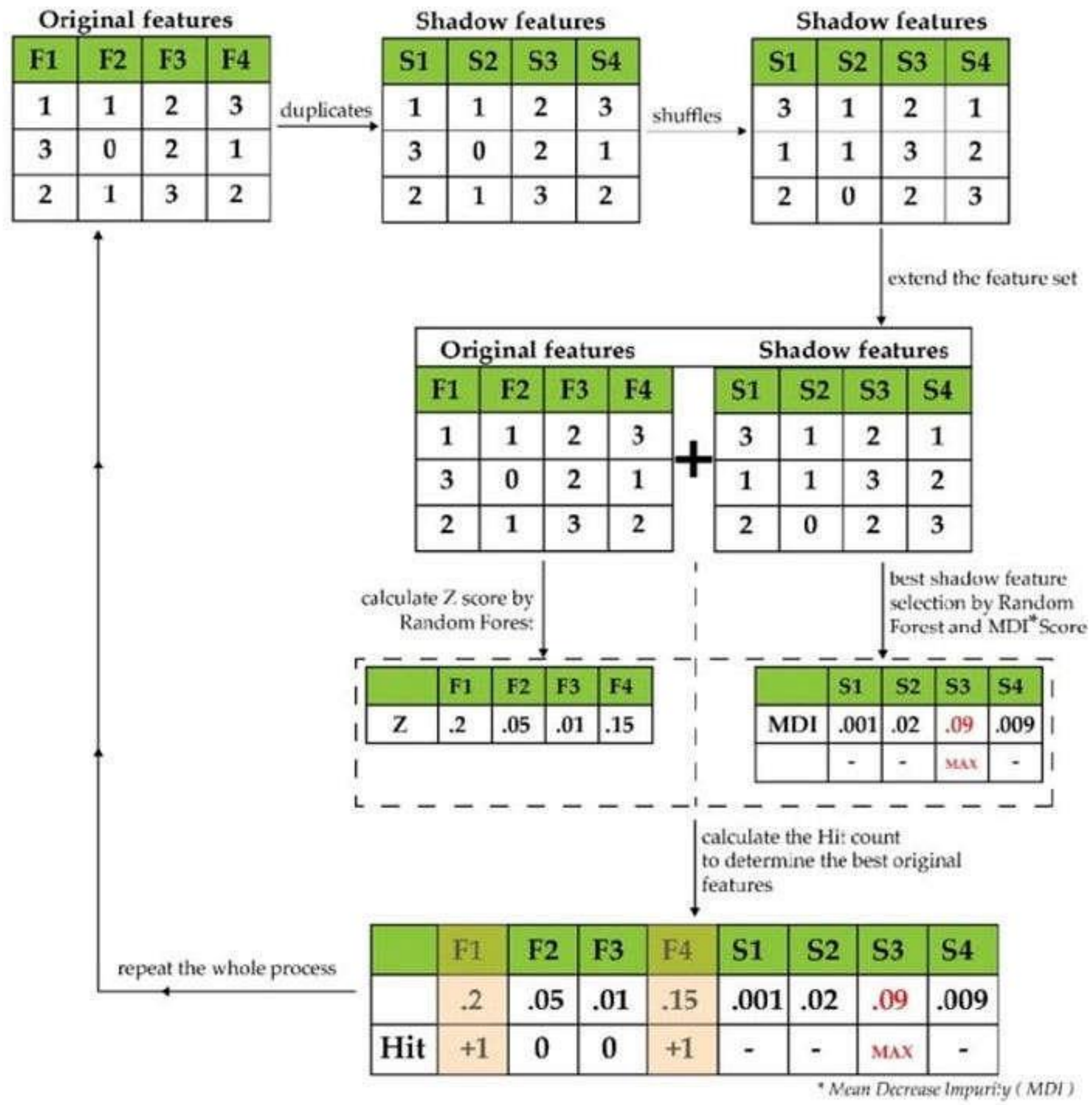
Wrapper methods are computationally more expensive than filter methods, due to the repeated learning steps and cross-validation. However, these methods are more accurate than the filter method. Some of the examples are Recursive feature elimination, Sequential feature selection algorithms, and Genetic algorithms.

3.2.2.1. Boruta Algorithm:

For this demonstration, we’ve chosen to implement the Boruta algorithm, with XGBoost as our wrapper classifier. By doing so, we found it to be better on the performance and efficiency fronts. The classifier tries to capture all important and interesting features that may be hiding in your data set concerning an outcome variable.

3.2.2.1.1. Methodology (Boruta):

- Creating duplicate features and shuffle their values in each column. These features are called shadow features.
- Trains a classifier (XGBoost) several times, on the Dataset and calculates all feature importance at all iterations.
- For all shadow features, we create a benchmark based on the mean importance and algorithm configuration parameter.
- Then, the algorithm checks for each of your real features if they have higher importance. That is, whether the feature has an importance greater than the benchmark. and keeps only the ones greater.


duplicates
shuffles
extend the feature set
calculate Z score by Random Forest
best shadow feature selection by Random Forest and MDI* Score
calculate the Hit count to determine the best original features
repeat the whole process
* Mean Decrease Impurity (MDI)

Original features

F1	F2	F3	F4
1	1	2	3
3	0	2	1
2	1	3	2

Shadow features

S1	S2	S3	S4
1	1	2	3
3	0	2	1
2	1	3	2

Shadow features

S1	S2	S3	S4
3	1	2	1
1	1	3	2
2	0	2	3

Original features				Shadow features			
F1	F2	F3	F4	S1	S2	S3	S4
1	1	2	3	3	1	2	1
3	0	2	1	1	1	3	2
2	1	3	2	2	0	2	3

	F1	F2	F3	F4
Z	.2	.05	.01	.15

	S1	S2	S3	S4
MDI	.001	.02	.09	.009
	-	-	MAX	-

	F1	F2	F3	F4	S1	S2	S3	S4
	.2	.05	.01	.15	.001	.02	.09	.009
Hit	+1	0	0	+1	-	-	MAX	-

Figure 3.4: Schematic flow of Boruta algorithm

According to [21] in wrapper methods, we generate a new model for each feature subset that is generated. The performance of each of these is recorded and the features which produce the best performance model are used for training and testing the final algorithm. Unlike filter methods that use distance or information-based measures for feature selection, wrapper methods use many simple techniques for choosing the most significant attributes.

3.2.2.2. Forward selection:

It is an iterative greedy process where you start with absolutely no features and in each iteration, you keep adding one most significant feature. Here, the variables are added in the decreasing order of their correlation with the target variable.

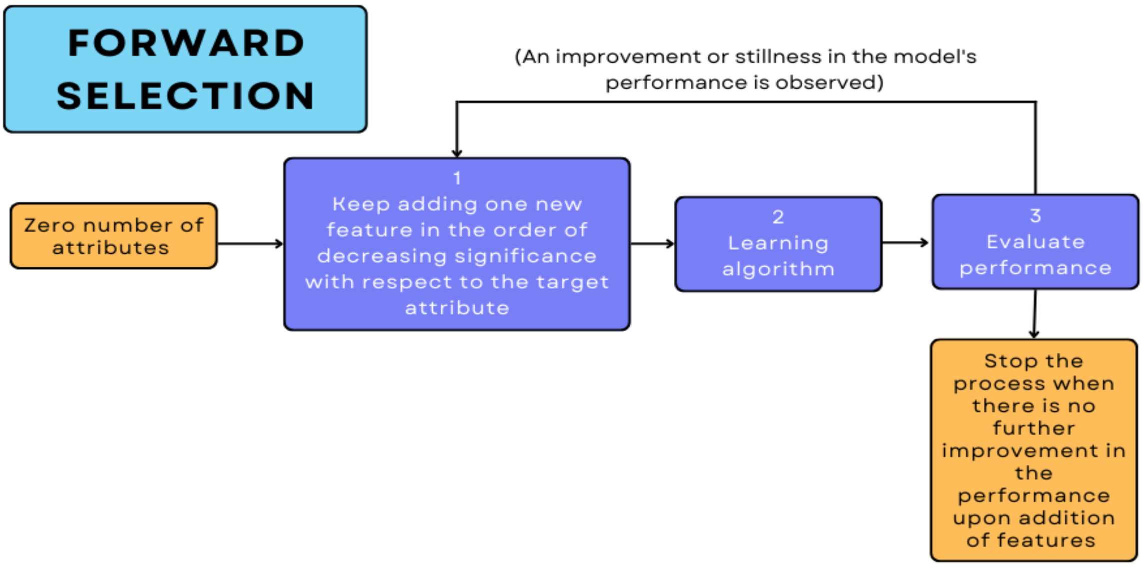

FORWARD SELECTION
(An improvement or stillness in the model's performance is observed)
Zero number of attributes
1 Keep adding one new feature in the order of decreasing significance with respect to the target attribute
2 Learning algorithm
3 Evaluate performance
Stop the process when there is no further improvement in the performance upon addition of features

Figure 3.5: Forward selection

This addition of a new attribute is done until the model's performance does not increase on further adding other features that are when you reach the point where you get the best possible performance.

3.2.2.3. Backward elimination:

As the name suggests, here we start with all the features present in the dataset, and with each iteration, we remove one least significant variable.

We remove the attributes until there is no improvement in the model's performance on eliminating features. The least correlated feature with the target variable is chosen based on certain statistical measures. In contrast to the filter methods, the features are removed in the increasing order of correlation with the target variable.

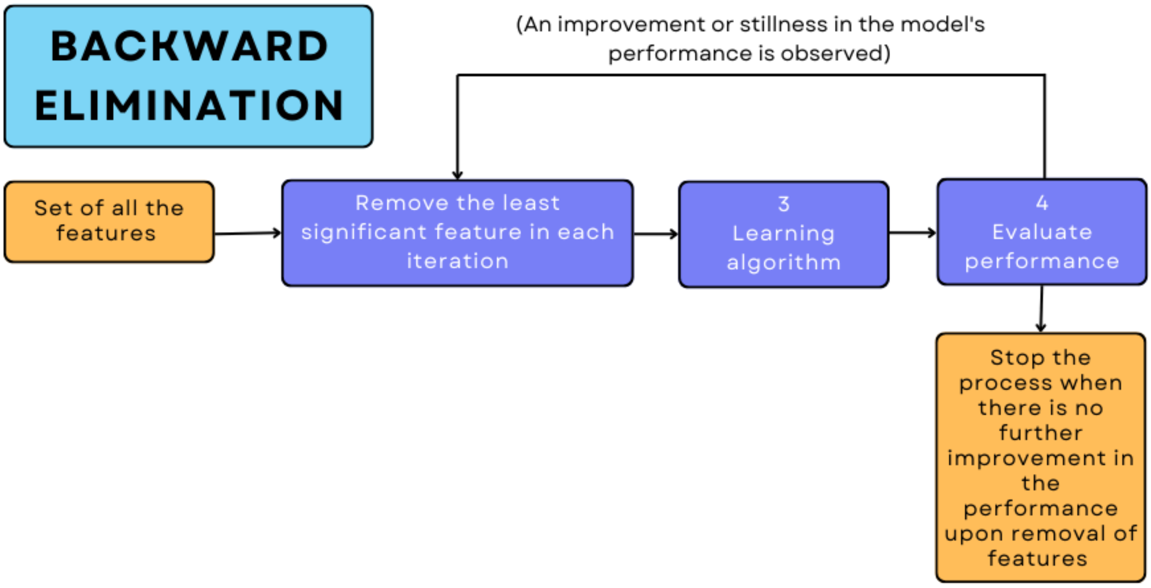

BACKWARD ELIMINATION
(An improvement or stillness in the model's performance is observed)
Set of all the features
Remove the least significant feature in each iteration
3 Learning algorithm
4 Evaluate performance
Stop the process when there is no further improvement in the performance upon removal of features

Figure 3.6: Backward elimination

It is also possible to combine both these methods. This is often called Bidirectional Elimination. This is similar to forward selection, but the only difference is that if it finds any already added feature to be insignificant at a later stage when a new feature is added, it removes the former through backward elimination.

3.2.2.4. Exhaustive Feature Selection:

Brute-force evaluation of each feature subset. This means that it tries every possible combination of features and returns the best performing subset [20]

3.2.2.5. Recursive Feature Elimination:

Given weights estimator for features (coefficient of the linear model), the goal is to select features by recursively considering smaller sets of features; Estimator trained, and importance of each feature obtained, and then least important features pruned from the feature set. This process repeats recursively on pruned set until the desired num of features reached [20]

It is worth noting that wrapper methods may work very effectively for certain learning algorithms. However, the computational costs are extremely high when these wrapper methods as compared to filter methods [21]

3.2.3 Hybrid method:

The process of creating hybrid feature selection methods depends on what you choose to combine. The main priority is to select the methods you're going to use, then follow their processes. The idea here is to use these ranking methods to generate a feature ranking list in the first step, then use the top k features from this list to perform wrapper methods. With that, we can reduce the feature space of our dataset using these filter-based rangers to improve the time complexity of the wrapper methods [20].

3.2.4 Embedded method:

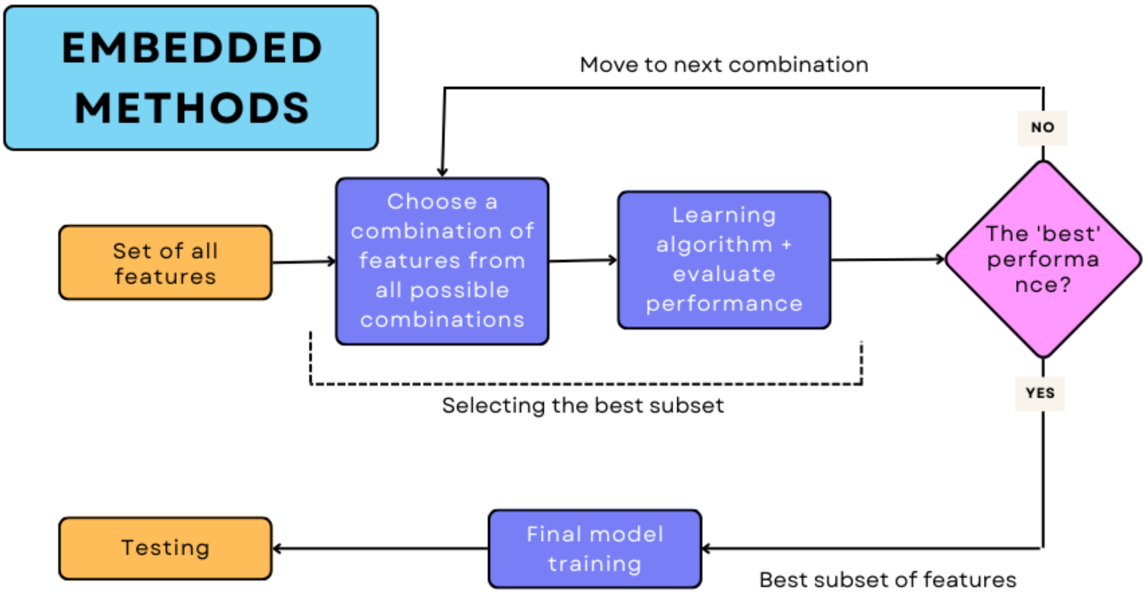


Figure 3.7: Embedded method

All possible feature combinations are created using embedded algorithms. After the model has been trained using each of these attribute combinations, its performance is monitored as usual. For the final training, the combo that performs the best is selected.

For feature selection, the Embedded approach offers hybrid learning and ensemble learning techniques. Because it makes decisions collectively, it performs better than the other two models.

Among them is Random Forest. In comparison to wrapper approaches, it requires less computing power. Nevertheless, there is a learning model-specific disadvantage to this approach.

In embedded techniques, the feature selection algorithm is integrated as part of the learning algorithm. The most typical embedded technique is the decision tree algorithm. Decision tree algorithms select a feature in each recursive step of the tree growth process and divide the sample set into smaller subsets [20]

The choice of technique used for feature selection depends on the application and the dataset's size and requires an in-depth understanding of the dataset. As mentioned before [21]

3.2.4.1. LASSO Regularization L1:

Shrinks some of the coefficients to zero. Therefore, that feature can be removed from the model [20]

3.2.4.2. Random Forest Importance:

Random Forests inherently rank by how well they improve the purity of node, decreasing impurity (Gini impurity) over all trees. Nodes with the greatest decrease in impurity happen at the start of trees, while nodes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features [20]

3.3 Unsupervised Feature Selection Methods:

According to [20] Due to the scarcity of readily available labels, unsupervised feature selection methods are widely adopted in the analysis of high-dimensional data. However, most of the existing UFS methods primarily focus on the significance of features in maintaining the data structure while ignoring the redundancy among features. Moreover, the determination of the proper number of features is another challenge.

There are two issues involved in developing an automated Feature Subset Selection Algorithm for unlabeled data:

- The need for finding the number of clusters in conjunction with feature selection.
- The need for normalizing the bias of feature selection criteria concerning the dimension.

Unsupervised feature selection methods are classified into four types, based on the interaction with the learning model as Filter, Wrapper, and Hybrid methods.

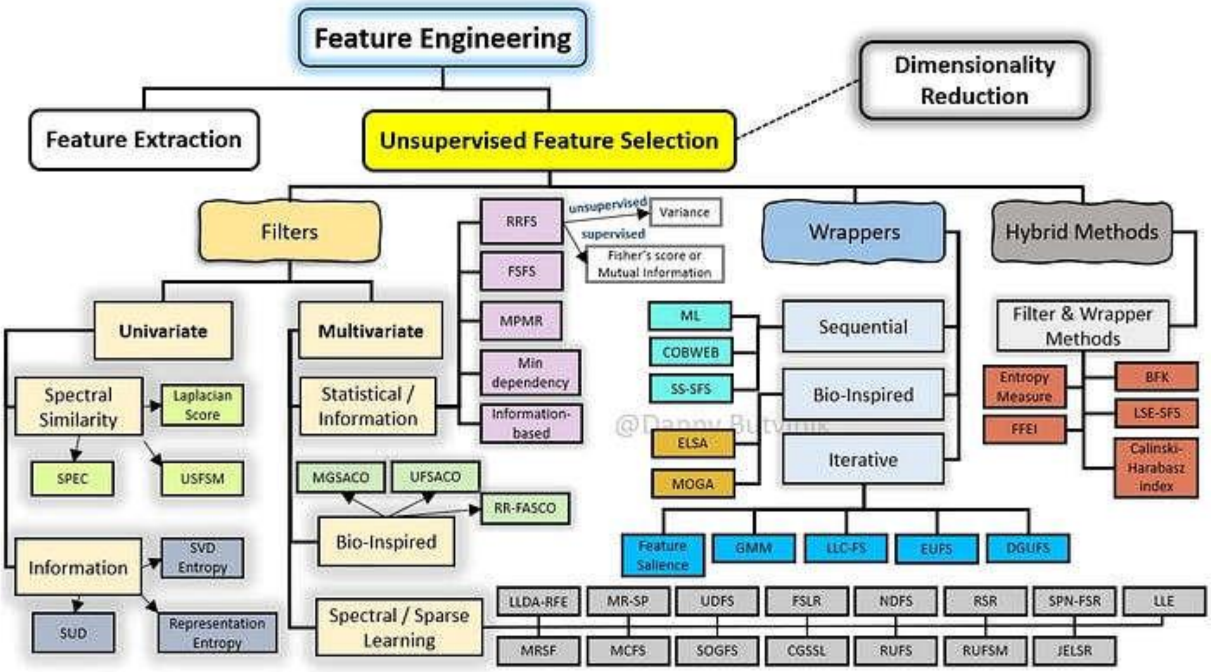


Figure 3.8: Extended taxonomy of unsupervised feature selection methods and techniques

3.3.1 Filter Methodology:

Unsupervised feature selection methods based on the filter approach can be categorized as univariate and multivariate. Univariate methods, aka ranking-based unsupervised feature selection methods use certain criteria to evaluate each feature to get an ordered ranking list of features, where the final feature subset is selected according to this order. Such methods can effectively identify and remove irrelevant features, but they are unable to remove redundant ones since they do not take into account possible dependencies among features. On the other hand, multivariate filter methods evaluate the relevance of the features jointly rather than individually.

Multivariate methods can handle redundant and irrelevant features; thus, in many cases, the accuracy reached by learning algorithms using the subset of features selected by multivariate methods is better than the one achieved by using univariate methods [20].

3.3.1.1. Univariate Filter Methods:

Within the univariate filter methods, two main groups can be highlighted:

- Methods that assess the relevancy of each feature based on Information.
- Methods that evaluate features based on Spectral Analysis using the similarities among objects.

In methods based on information, the idea is to assess the degree of dispersion of the data through measures such as entropy, divergence, mutual information, among others, to identify cluster structures in the data.

On the other hand, methods based on Spectral Analysis — Similarity, aka Spectral Feature Selection methods, follow the idea of modelling or identifying the local or global data structure using the eigensystem of Laplacian or normalized Laplacian matrices derived from an object similarity matrix.

3.3.1.1.1. Information-based Methods:

3.3.1.1.1.1. Sequential Backward Selection Method for Unsupervised Data:

SUD is a filter method that weighs features using a measure of the entropy of similarities based on distance, which is defined as the total entropy induced from a similarity matrix, where the elements of this matrix contain the similarity between pairs of objects in the dataset. The idea is to measure the entropy of the data based on the fact that when every pair of objects is very close or very far, the entropy is low, and it is high if most of the distances between pairs of objects are close to the average distance. Therefore, if the data has low entropy, there are well-defined cluster structures, while there are not when the entropy is high. The relevance of each feature is quantified using a leave-one-out sequential backward strategy jointly with the entropy measure above mentioned. The final result is a feature ranking ordered from the most to the least relevant feature.

3.3.1.1.1.2. SVD-Entropy:

The central idea is to select those features that best represent the data, measuring the entropy of the original data matrix through its singular values. This entropy varies between 0 and 1, in such a way that when the entropy is low (close to zero), well-formed clusters are generated, since the spectrum of the data matrix is not uniformly distributed.

When the entropy is high, the spectrum is uniformly distributed, and the cluster structure is not well-defined. Through a leave-one-out comparison, the contribution of each feature to the entropy (CE) is evaluated, and the features are sorted according to their respective CE values.

3.3.1.1.1.3. Representation Entropy:

This method ranks features using information theory. The aim is to weigh each feature using the concept of Representation Entropy. Representation Entropy is a measure of information compression in a dataset, and it is computed from the entropy of eigenvalues of the covariance matrix of the data. Representation Entropy ranges from 0 to 1, where 1 represents the maximum compression, and 0 is the minimum one.

3.3.1.1.2. Spectral Similarity-based Methods:

3.3.1.1.2.1. Laplacian Score:

One of the most referenced and relevant univariate filters in unsupervised feature selection methods based on Spectral Feature Selection is Laplacian Score. In Laplacian Score, the importance of a feature is evaluated by its variance and its power of locality preserving. This method assigns high weights to those features that most preserve the predefined graph structure represented by the Laplacian matrix. This idea comes from the observation that two objects are probably related to the same cluster if they are close to each other; in such a way that those features that take similar values for close objects, and dissimilar values for the far away ones are the most relevant.

3.3.1.1.2.2. Spectrum Decomposition:

SPEC evaluates the relevance of a feature by its consistency with the structure of the graph induced by the similarities among objects.

This method consists of three steps: (1) building the object similarity matrix as well as its graph representation; (2) evaluating features using the eigensystem of the graph by measuring the consistency between each feature and those nontrivial eigenvectors of the Laplacian matrix. And (3), ranking features in descending order in terms of their feature relevance and consistency. In fact, SPEC is a generalization of the Laplacian Score.

3.3.1.1.2.3. Unsupervised Spectral Feature Selection:

Unsupervised Spectral Feature Selection Method applied on mixed data. It assesses features by analyzing the changes in the spectrum distribution (spectral gaps) of the first non-trivial eigenvalues of the Normalized Laplacian matrix when each feature is excluded from the whole set of features separately. Features are sorted in descending order according to their respective spectral gap values.

3.3.1.2. Multivariate Filter Methods:

Multivariate filter methods can be divided into three main following groups: Statistical / Information, Bio-inspired, and Spectral / Sparse Learning-based methods.

According to [20] statistical/Information group includes unsupervised feature selection methods that perform the selection using statistical and/or information theory measures such as variance-covariance, linear correlation, entropy, mutual information, among others.

The bio-Inspired group includes unsupervised feature selection methods that use stochastic search strategies based on the swarm intelligence paradigm for finding a good subset of features, which satisfies some criterion of quality.

Spectral/Sparse learning group includes unsupervised feature selection methods that are based on spectral analysis or a combination of spectral analysis and sparse learning. This group of methods is sometimes considered embedded methods because feature selection is achieved as part of the learning process, commonly through the optimization of a constrained regression model. But embedded methods also could be considered as a subcategory of filter, wrapper, and hybrid approaches.

3.3.1.2.1. Statistical / Information Group:

3.3.1.2.1.1. Feature Selection using Feature Similarity:

FSFS method introduces a statistical measure of dependency/similarity to reduce feature redundancy; this measure called Maximal Information Compression Index (MICI) is based on the variance-covariance between features. The idea of this method is partitioning the original set of features into clusters, such that those features in the same cluster are highly similar, while those in different clusters are dissimilar. Feature clustering is done iteratively based on the KNN principle as follows: in each iteration, FSFS computes the k-nearest features of each feature (using MICI). Then, the feature with the most compact subset of k-nearest features (determined by the distance to its farthest feature among the k-nearest) is selected, and its k nearest features are discarded. This procedure is repeated for the remaining features until all of them are either picked or discarded.

3.3.1.2.1.2. Relevance Redundancy Feature Selection:

RRFS is a supervised/unsupervised feature selection method, which selects features in two steps. First, the features are sorted according to a relevance measure (variance for the unsupervised version and the Fisher's Ratio or mutual information for the supervised one). Then, in the second step, following the order generated in the previous step, the features are evaluated using a feature similarity measure to quantify the redundancy between them. Afterward, the first p features with the lowest redundancy are selected.

3.3.1.2.1.3. Maximum Projection and Minimum Redundancy:

MPMR is an unsupervised feature selection method based on a criterion of maximum projection and minimum redundancy. The core idea is to select a feature subset such that all original features are projected into a feature subspace (applying a linear transformation) with minimum reconstruction error while aiming to maintain low redundancy, a term for quantifying the redundancy among features, estimating redundancy rate with added Pearson correlation coefficient.

3.3.1.2.1.4. Information based:

In information-based, the basic idea is to select features using a measure of the entropy of similarities based on distance.

3.3.1.2.1.5. Minimum Dependency:

Min Dependency is a multivariate statistical-based filter method that holds the objective to remove redundant features using the concept of minimization of the feature dependency. The idea is to find independent features (relevant) by choosing a set of coefficients such that the linear dependency of features (expressed by the error vector E) could be close to zero. At each iteration, the feature with the largest absolute coefficients is removed, and the effect of its removal is updated. This process is iterated until all the remaining error vectors E are smaller than a threshold fixed by the user.

3.3.1.2.2. Bio-Inspired Methods:

3.3.1.2.2.1. Unsupervised Feature Selection based on Ant Colony Optimization:

UFSACO's main objective is to select feature subsets with low similarity among features (low redundancy). In this case, the search space is represented as a complete undirected graph, where the nodes represent the features, and the weights of the edges represent the similarities between features. This similarity is computed using the cosine similarity function. The idea is that if two features are similar, then these features are redundant. Each node in the graph has a desirability value called pheromone, which is updated by agents (ants) in the function of its current value, a pre-specified decay rate, and the number of times that a given feature has been selected by an agent. The agents traverse the graph iteratively preferring high pheromone values and low similarities until a pre-specified stop criterion (number of iterations) is reached. Finally, those features with the highest pheromone value are selected. Consequently, it is expected to pick feature subsets with low redundancy.

3.3.1.2.2.2. Microarray Gene Selection based on Ant Colony Optimization and Relevance-Redundancy Feature Selection based on ACO (ant colony optimization):

In both **MGSACO** and **RR-FSACO**, in addition to quantifying the feature redundancy as in the UFSACO method, they also measure the relevance of each feature through its variance. Therefore, the main objective of all these methods is to select features that minimize redundancy and at the same time maximize relevance.

3.3.1.2.3. Spectral / Sparse Learning Methods:

Some multivariate methods based on spectral analysis derived from the SPEC and the Laplacian Score.

3.3.1.2.3.1. Minimum-Redundancy Spectral Feature Selection:

In MR-SP that combines the SPEC ranking and the minimum redundancy optimality criterion. The basic idea of this method is to add a way for controlling the feature redundancy in SPEC, by introducing an evaluation measure for quantifying the similarity of each pair of features through a modified cosine similarity function.

3.3.1.2.3.2. Laplacian Linear Discriminant Analysis-based Recursive Feature Elimination:

LLDA-RFE method extends the Linear Discriminant Analysis (LDA) to the unsupervised case using the similarities among objects; this extension is called LLDA. The idea is to recursively remove features with the smallest absolute values of the discriminant vectors of the LLDA to identify features that potentially reveal clusters in the samples. LLDA-RFE is closely related to Laplacian Score; the main difference is that LLDA-RFE is a multivariate method, which allows selecting features that in combination contribute to discrimination.

3.3.1.2.3.3. Multi-Cluster Feature Selection:

MCFS consists of three steps: (1) spectral analysis, (2) sparse coefficient learning, and (3) feature selection. In the first step, spectral analysis is applied to the dataset to detect the cluster structure of the data. Then, in the second step, since the embedding clustering structure of the data is known, through the first k eigenvectors of the Laplacian matrix, MCFS measures the importance of the features by a regression model with an L1-norm regularization. In the third step, after solving the regression problem, MCFS selects d features based on the highest absolute values of the coefficients obtained through the regression problem.

3.3.1.2.3.4. Minimize the feature Redundancy for Spectral Feature Selection:

MRSF evaluates the features all together to eliminate redundant features. The idea is to formulate the feature selection problem as a multi-output regression problem, and the selection is performed by enforcing the sparsity applying the norm L2, instead of the L1 norm.

3.3.1.2.3.5. Unsupervised Discriminative Feature Selection Algorithm:

UDFS performs feature selection by simultaneously exploiting discriminative information contained in the scatter matrices and feature correlations. This method proposes to address the feature selection problem taking into account the trace criterion into the regression problem. Furthermore, UDFS adds some additional constraints to the regression problem and proposes an efficient algorithm to optimize it. UDFS ranks each feature according to the corresponding weight value in descending order, and the top-ranked features are selected.

3.3.1.2.3.6. Joint Embedding Learning and Sparse Regression:

JELSR operates on the same objective function as MRSF, and it only differs in the construction of the Laplacian graph, since in this work, locally linear approximation weight is used to measure local similarity for building the Laplacian graph.

3.3.1.2.3.7. Nonnegative Discriminative Feature Selection:

NDFS like UDFS and MRFS, performs feature selection exploiting the discriminative information and feature correlations in a unified framework. First, NDFS uses spectral analysis to learn pseudo-class labels (defined as non-negative real values). Then, a regression model regularization is formulated and optimized through a special solver also proposed in this work. The main difference between NDFS and UDFS is that NDFS adds a non-negativity constraint to the regression problem, since removing this constraint NDFS becomes UDFS.

3.3.1.2.3.8. Feature subset with Sparsity and Low Redundancy:

FSLR employs spectral analysis to represent the data in a lower dimension and introduces a novel regularization term into the objective function with a non-negative constraint.

3.3.1.2.3.9. Structured Optimal Graph Feature Selection:

SOGFS simultaneously performs feature selection and local structure learning, which was proposed. SOGFS adaptively learns local manifold structure by introducing a similarity matrix in a sparse optimization model based on minimization on both loss function and regularization. Features are selected according to the corresponding weights once the proposed model has been optimized.

3.3.1.2.3.10. Clustering-Guided Sparse Structural Learning:

CGSSL is a general method for feature selection that jointly exploits nonnegative spectral analysis and structural learning with sparsity. The idea is to use the cluster indicators (learned with nonnegative spectral clustering) in a linear model to provide label information for structural learning.

3.3.1.2.3.11. Robust Unsupervised Feature Selection:

RUFS's objective is to achieve both robust clustering and robust feature selection. Unlike the unsupervised feature selection methods above mentioned such as MCFS, UDFS, and NDFS, RUFS learns the pseudo cluster labels via local learning regularized robust nonnegative matrix factorization.

3.3.1.2.3.12. Robust Unsupervised Feature Selection via Matrix Factorization:

RUFSM selects features by performing discriminative feature selection and robust clustering simultaneously. The main difference between RUFS and RUFSM is that the latter uses the cluster centers as an objective concept rather than the pseudo labels of the data.

3.3.1.2.3.13. Regularized Self-Representation Model for Unsupervised Feature Selection:

RSR reflects the idea that if a feature is important, then it will participate in the representation of most of the other features. The feature selection is done by the minimization of the self-representation error for the characterization of residuals, and the most representative features (those with high feature weights) are selected.

3.3.1.2.3.14. Structure-Preserving Non-negative Feature Self-Representation:

SPNFSR method takes into account both the self-representation and the structure-preserving ability of features by optimizing a model. The general idea of this method methods is to optimize a model (objective function) take into account three aspects: (1) the self-representation of features; (2) the local manifold geometrical structure of the original data using a graph-based norm regularization term; 3) a regularization term W to reflect the importance of each feature. The optimization problem is solved through an efficient iterative algorithm. At the final stage, each feature is sorted according to the corresponding W values in descending order and the top p ranked features are selected.

3.3.1.2.3.15. Locally Linear Embedding:

LLE is a set of non-convex sparse regularization functions in sparse learning models. The idea is to characterize the intrinsic local geometric through an LLE graph-based instead of the typical pairwise similarity matrix jointly with a structure regularization term. For each feature, a feature-level reconstruction score based on the LLE graph is defined, and the final feature subset is selected according to this score.

3.3.2 Wrapper Methodology:

According to [20] unsupervised feature selection methods based on the wrapper approach can be divided into three broad categories.

According to the feature search strategy: sequential, bio-inspired, and iterative. In sequential methodology, features are added or removed sequentially. Methods based on sequential search are easy to implement and fast.

On the other hand, bio-inspired methodology tries to incorporate randomness into the search process, aiming to escape from local optima.

Iterative methods address the unsupervised feature selection problem by casting it as an estimation problem and thus avoiding a combinatorial search.

Wrapper methods evaluate feature subsets using the results of a specific clustering algorithm. Methods developed under this approach are characterized by finding feature subsets that contribute to improving the quality of the results of the clustering algorithm used for the selection. However, the main disadvantage of wrapper methods is that they usually have a high computational cost, and they are limited to be used in conjunction with a particular clustering algorithm.

3.3.2.1. Sequential Methods:

3.3.2.1.1. Maximum Likelihood:

Two feature selection criteria are evaluated: the criterion of ML and the scatter separability criterion. This method searches through the space of feature subsets, evaluating each candidate subset as follows: First, Expectation-Maximization (EM) or KMeans clustering algorithms are applied on the data described by each candidate subset. Then, the obtained clusters are evaluated with the ML or T separability criteria. The method uses a forward selection search for generating subsets of features that will be evaluated as described above. The method ends when the change in the value of the used criterion is smaller than a given threshold.

3.3.2.1.2. Category Utility by COBWEB:

This method is based on a measure called category utility, which is used to measure the quality of the clusters found by the COBWEB hierarchical clustering algorithm. This method generates subsets of features with two search strategies: forward selection and backward elimination. Feature selection is performed by running the COBWEB algorithm using the subset of features generated by the search strategy and evaluating the category utility for this feature subset. The process ends when no higher category utility score can be obtained in the backward or forward selection.

3.3.2.1.3. Simplified Silhouette Sequential Forward Selection:

SS-SFS method selects a feature subset that provides the best quality according to the simplified silhouette criterion.

In this method, a forward selection search is used for generating subsets of features. Each feature subset is used to cluster the data using the k-means clustering algorithm, and the quality of the feature subset is evaluated through the quality of the clusters measured with the simplified silhouette criterion. The feature subset that produces the best value of this criterion in the forward selection is selected.

3.3.2.2. Bio-Inspired Methods:

3.3.2.2.1. Evolutionary Local Selection Algorithm:

ELSA method searches feature subsets as well as the number of clusters based on the KMeans and Gaussian Mixture clustering algorithms. Each solution provided by the clustering algorithms is associated with a vector whose elements represent the quality of the evaluation criteria, which are based on the cohesion of the clusters, inter-class separation, and maximum likelihood. Those features that optimize the objective functions in the evaluation stage are selected.

3.3.2.2.2. Multi-Objective Genetic Algorithm:

MOGA method proposes a multi-objective fitness function that minimizes the intra-cluster distance (uniformity) and maximizes the inter-cluster distance (separation). Each chromosome represents a solution, which is composed of a set of k cluster centroids (cluster center for continuous features and cluster mode for categorical features) described by a subset of features. The number of features used for each centroid in each chromosome is randomly generated, and the cluster centers and cluster modes of chromosomes in the initial population are created by generating random numbers, and feature values from the same feature domain, respectively. Then, for re-assigning cluster centroids, MOGA uses the k-prototypes clustering algorithm which obtains its inputs from the initial population generated in the previous step. Afterward, the crossover, mutation, and substitution operators are applied, and the process is repeated until a prespecified stop criterion is met. In the final stage, this method returns the feature subset that optimizes the fitness function jointly with the clusters that they produced.

3.3.2.3. Iterative Methods:

3.3.2.3.1. Feature Salience:

The idea is to estimate a set of weights (real values in $[0 - 1]$) called feature saliences (one for each feature) to quantify the relevance of each feature. This estimation is carried out by a modified EM algorithm derived for the task. The method returns the parameters of the density functions that model the components (clusters), as well as the set of feature salience values. Then, the user can consider those feature saliencies that best discriminate between different components (those with the highest values).

3.3.2.3.2. Local Learning-Based Clustering:

LLC framework formulates the final ridge regression model. Feature selection is done by introducing a binary feature selection vector τ to the local discriminant function of the model. In the end, after the convergence, the output is the vector τ along with a discretized cluster indicator matrix.

3.3.2.3.3. Embedded Unsupervised Feature Selection:

EUFS method directly embeds the feature selection in the clustering algorithm via Sparse Learning. In this method non-convex sparse regression model uses a loss function based on L2 norm and optimizes through an Alternating Direction Method of Multipliers (ADMM). EUFS uses the KMeans clustering algorithm to initialize a pseudo cluster indicator matrix U and a latent feature matrix V (used for indicating feature weights) in the final model. Once the model has converged, the output is a feature ranking sorted according to the final values of the latent feature matrix along with the pseudo clusters indicators.

3.3.2.3.4. Dependence Guided Unsupervised Feature Selection:

DGUFS method simultaneously performs feature selection and clustering using a constraint model. The model is optimized using a modified algorithm based on the iterative Alternating Direction Method of Multipliers (ADMM).

3.3.2.3.5. Gaussian Mixture Models:

In GMM the idea is to apply feature selection and clustering simultaneously, using a Gaussian mixture model. The objective is to optimize the Gaussian mixture model via the Expectation-Maximization clustering algorithm, where the maximization step of this algorithm was reformulated as an L1-constraint LASSO problem. The method returns the clusters as well as the coefficients of the model; the coefficients indicate the relevance of each feature.

3.3.3 Hybrid Methodology:

Hybrid methods try to exploit the qualities of both approaches, filter, and wrapper, trying to have a good compromise between efficiency (computational effort) and effectiveness (quality in the associated objective task when using the selected features).

To take advantage of the filter and wrapper approaches, hybrid methods, in a filter stage, the features are ranked or selected applying a measure based on intrinsic properties of the data. While, in a wrapper stage, certain feature subsets are evaluated for finding the best one, through a specific clustering algorithm. We can distinguish two types of hybrid methods: methods based on ranking and methods non-based on the ranking of features.

It is worth noting that, in the literature, some hybrid unsupervised feature selection methods like (Jashki et al. 2009; Hu et al. 2009; Yang et al. 2011a; Yu 2011) designed specifically for handling data in specific domains also have been proposed [20].

3.3.3.1. Fuzzy Evaluation Index:

In FFEI method exponential entropy measure is combined with the fuzzy evaluation index for feature ranking and feature subset selection, respectively. The method employs sequential search considering subsets of features based on the generated ranking and using the fuzzy evaluation index as a quality measure. In the wrapper stage, to select even a smaller feature subset, the fuzzy-c-means algorithm and the scatter separability criterion are used to select a compact subset of features.

3.3.3.2. Calinski-Harabasz Index:

In this method, spectral feature selection is combined with the Calinski-Harabasz index for selecting a relevant feature subset. The feature selection is divided into two stages: (1) Feature ranking and, (2) feature subset selection. In the first stage, the idea is to identify those features that preserve the data structure computing for each feature the Laplacian Score. This produces a feature ranking. After, in the second stage, taking advantage of the ranking generated in the previous stage and using forward or backward selection search, feature subsets are evaluated through a modified internal evaluation index called Weighted Normalized Calinski-Harabasz index (WNCH). The feature subset with the highest WNCH value is selected.

3.3.3.3. Bayesian Filter KMeans:

BFK is a non-biased ranking method that combines KMeans and a Bayesian filter. This method, unlike all the above-mentioned hybrid methods, begins with the wrapper stage, by running the KMeans clustering algorithm on the dataset with a range of clusters specified by the user. The clusters are evaluated with the simplified silhouette criterion and the one with the highest value is selected. Subsequently, in the filter stage, using the concept of Markov blanket, a feature subset is selected through a Bayesian network, where each cluster represents a class, the nodes represent features, and the edges represent relationships between features.

3.3.3.4. Least Square Estimation using Sequential Forward Selection:

LSE-SFS is a non-based on ranking method that removes both irrelevant and redundant features. This method performs feature selection in two steps: in the first step, a subset of features is founded by applying the least-square estimation LSE-based evaluation. The second step works only with those features identified in the first step, and by using a Sequential Forward Selection search the best feature subset that maximizes the clustering performance (using a modified version of the EM clustering algorithm).

3.3.3.5. Entropy Measure:

This method one of the first ranking unsupervised hybrid feature selection methods that are based on the entropy measure where filter stage joints with the internal scatter separability criterion (wrapper stage). In the filter stage, each feature, one by one, is removed from the whole set of features, and the entropy generated in the dataset after the elimination of the feature is computed. This produces a sorted list of features according to the degree of disorder that each feature generates when it was removed from the whole set of features. Once all features have been sorted, in the wrapper stage, a forward selection search is applied jointly with the k-means clustering algorithm to build clusters that are evaluated using the scatter separability criterion. This method selects the feature subset that reaches the highest value for the separability criterion.

3.4 Conclusion:

In the vast landscape of machine learning, both approaches: supervised and unsupervised methods serve distinct purposes and play a crucial role in solving real-world problems.

Selecting the right approach depends on the problem context, data availability, and desired outcomes. Supervised methods excel when labeled data is abundant, while unsupervised methods explore hidden structures in unlabeled data.

4 CHAPTER IV: CASE STUDY ─ Industrial Safety and Health Analytics Database

4.1 Overview

The database comes from one of the biggest industries in Brazil and in the world. There is an urgent need for companies to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such environment.

4.2 Objectives

- Analyze real labor accident data aiming to help manufacturing plants to find solutions that improve the level of safety by reducing the rate of accidents and save people lives!
- Showcase the use of feature selection techniques to assess the risks and enhance safety

4.3 Details

- Coding language for database analysis: **Python**
- The Python IDE software (environment used to run the code): **PyCharm**
- The Python libraries used (data handling tools): **Pandas, Scikit-learn**

4.4 Data description

The database is basically records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident. (**database source from *www.kaggle.com***) [22]

4.4.1 Columns description

- Data: timestamp or time/date information
- Countries: which country the accident occurred (anonymized)
- Local: the city where the manufacturing plant is located (anonymized)
- Industry sector: which sector the plant belongs to
- Accident level: from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)

- Potential Accident Level: Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident)
- Gender: if the person is male of female
- Employee or Third Party: if the injured person is an employee or a third party
- Critical Risk: some description of the risk involved in the accident
- Description: Detailed description of how the accident happened.

4.5 Methodology of application of Feature Selection

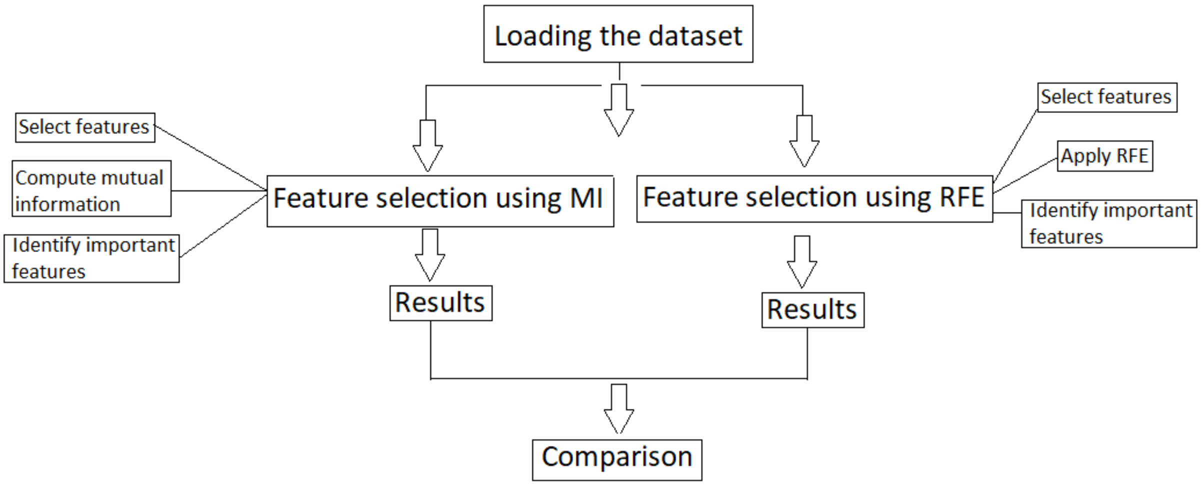

Loading the dataset
Select features
Compute mutual information
Identify important features
Feature selection using MI
Feature selection using RFE
Select features
Apply RFE
Identify important features
Results
Results
Comparison

Figure 4.12: Methodology used for case studying

4.5.1 Feature Selection using RFE method (Wrapper)

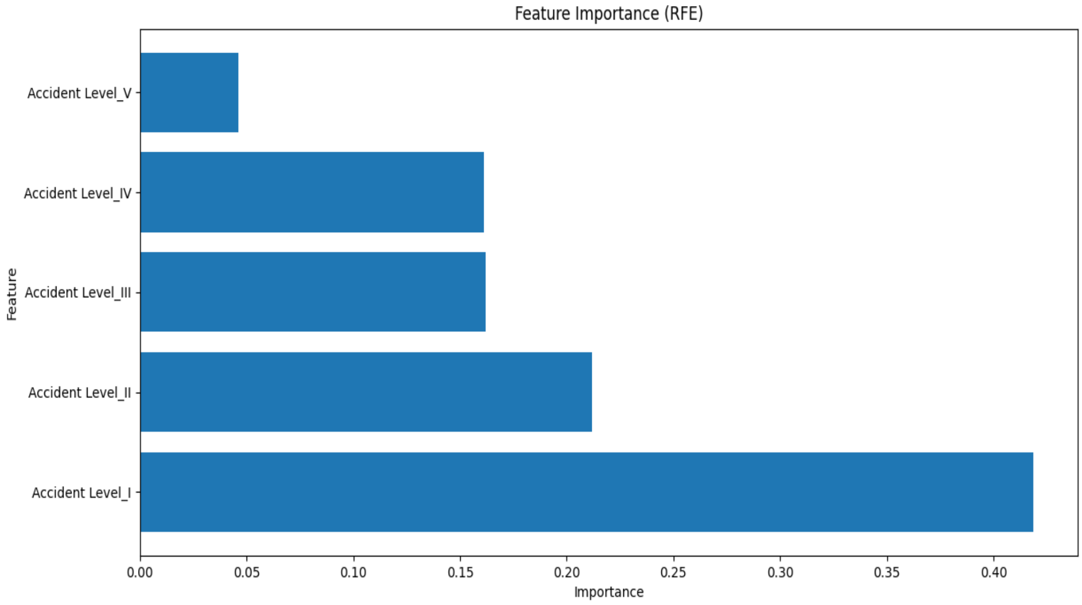


Figure 4.13: The importance score of selected features using Recursive Feature Elimination (RFE)

4.5.1.1. Results and discussion

The output from the script will indicate which features have been selected as the most predictive according to the RandomForest model within the RFE framework. The selected features' names provide insights into which factors are most influential in predicting the accident level according to the model's internal logic. The ranking provides a straightforward interpretation of feature importance, where a rank of 1 indicates a selected feature, and higher numbers indicate lesser importance.

Selected Features: The printed selected feature names will reveal the features chosen by the RFE algorithm. In this case, only the features related to the 'Accident Level' are selected, this suggests that, according to the feature selection criteria, the severity level of accidents is deemed as the most crucial factor in predicting the outcome.

Here are the selected features:

1. Accident Level_I
2. Accident Level_II
3. Accident Level_III
4. Accident Level_IV
5. Accident Level_V

Each accident level has an associated importance score. Accident Level_I has the highest importance score, followed by Level II, III, IV, and V.

Comparison with Original Dataset:

- The original dataset likely contains various accident-related features such as the type of industry, location, critical risk factors, etc.
- The selected features (Accident Levels) are likely derived from the target variable itself. This suggests that the accident levels are crucial in understanding the severity and frequency of incidents.

Risks and Major Risks:

- Accident levels can be interpreted as a classification of risks, where higher levels indicate more severe risks.
- The importance scores in the RFE output indicate which levels contribute the most to the model's ability to predict or classify accidents.
- If Accident Level I has the highest importance, it means that the most severe accidents (assuming Level I is the most severe) are the most critical to consider in safety and risk assessments.
- Conversely, if Level V is the least important, it might indicate that these are less severe and perhaps less critical for immediate focus.

Discussion:

- The output from RFE shows that higher severity levels (such as Level V) are more important, which aligns with the goal of identifying major risks.

- This suggests that in terms of risk prioritization, focusing on preventing and mitigating higher-level accidents (Levels IV and V) could be more beneficial.
- If the original dataset includes a variety of other risk factors, it’s important to ensure that these factors are also considered alongside the accident levels to get a comprehensive view of safety and health risks.

Final assessment:

The RFE results indicate that higher accident levels are major risks, which is expected as more severe accidents typically have more significant consequences, thus this approach ensures that the major risks are accurately identified and prioritized.

Model Performance:

While the selected features provide valuable insights into the dataset, it's crucial to assess the performance of the model using these features. Evaluation metrics such as accuracy, precision, recall, or F1-score can help determine how well the model performs in predicting the 'Accident Level' based on the selected features.

4.5.2 Feature Selection using MI method (Filter)

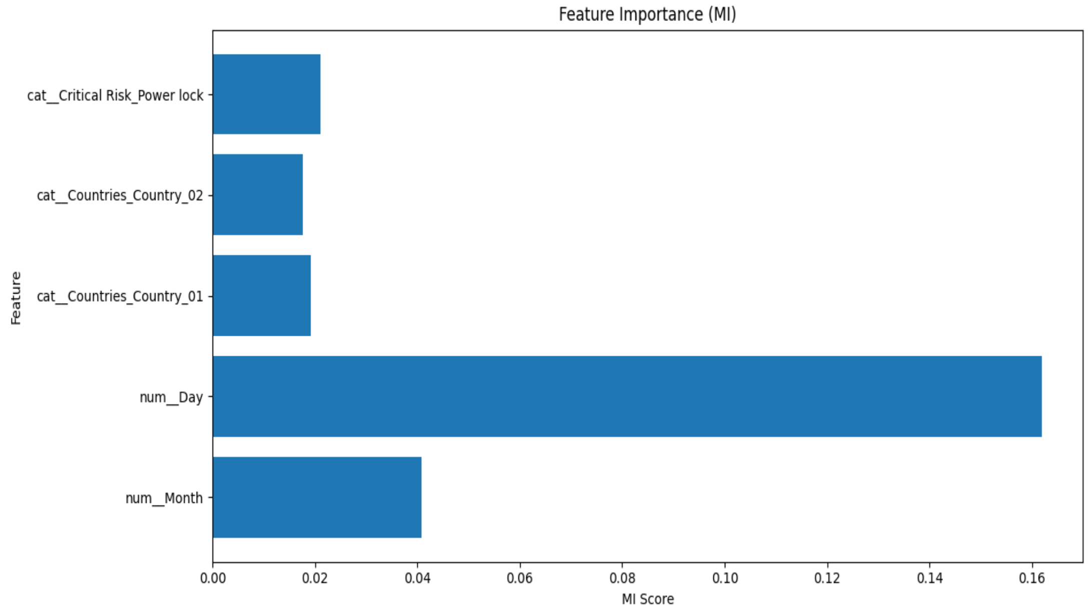


Figure 4.3: Mutual information scores of the top 5 selected features

4.5.2.1. Results and discussion

The graph illustrates the importance of various features in predicting the target variable, with each feature's importance measured by its Mutual Information (MI) score. The dataset likely includes multiple features, out of which the graph highlights six key features that contribute significantly to the predictive model. Here's a breakdown of these features and their potential implications:

- **num__Day:** This feature has the highest MI score, making it the most influential factor in predicting the target variable, "Accident Level." Its significance suggests that the specific day of an event plays a crucial role in the occurrence or severity of accidents, possibly due to recurring patterns or specific dates that are more accident-prone.
- **num__Month:** The second most important feature, "num__Month," also has a substantial MI score, although lower than "num__Day." This implies that the month or time of the year is a critical determinant in understanding accident trends, potentially due to seasonal variations or periodic factors.
- **cat__Countries_Country_01:** This categorical feature, representing a specific country, has a moderate MI score. It indicates that geographical location is a

significant factor, possibly reflecting differences in safety standards, regulations, or environmental conditions across different countries.

- **cat__Countries_Country_02:** Another geographical feature, this time representing a different country, shows a slightly lower MI score compared to "cat__Countries_Country_01." This further underscores the importance of location in accident analysis, suggesting that different countries may have unique risk profiles.
- **cat__Local_Local_05:** This feature, representing a specific local factor, has a moderate MI score. It suggests that local conditions or regulations in specific areas are important in predicting the accident level, possibly due to localized risk factors or environmental conditions.
- **cat__Critical Risk_Power lock:** Although it has the lowest MI score among the selected features, this factor is still significant. It represents a critical risk related to "Power lock," indicating that this specific risk factor has a noteworthy influence on accident severity or frequency. It may highlight the importance of safety measures related to power management and machinery operations.

These features were chosen based on their high relevance and importance in predicting the target variable within the risks database.

Interpretation of selected features:

The dataset likely contains various features related to accidents, such as "Countries," "Local Factors," "Critical Risks," and "Temporal Features." The graph emphasizes both numerical and categorical features, with a few key insights:

- **Temporal Features (num__Day, num__Month):** The prominence of "Day" and "Month" as significant features suggests that temporal factors play a vital role in accident occurrence. This could be related to specific days or months when accidents are more frequent, perhaps due to weather patterns, seasonal activities, or operational schedules.
- **Geographical Features (cat__Countries_Country_01, cat__Countries_Country_02):** The importance of country-specific features indicates that the location where an accident occurs is crucial in predicting its severity. This could reflect differences in safety regulations, industry practices, or regional environmental conditions that influence accident rates.
- **Local and Risk Factors (cat__Local_Local_05, cat__Critical Risk_Power lock):** The inclusion of these features highlights the importance of considering localized factors and specific risks when analyzing accidents. "Power lock" as a critical risk emphasizes the need for stringent safety protocols around machinery and power management to mitigate accident risks.

Discussion on Major Risks:

The feature importance graph provides valuable insights into the key factors influencing accident levels:

- **Temporal Trends:** The high MI scores for "Day" and "Month" suggest that certain times of the year or specific days are more prone to accidents. This could help organizations plan and implement targeted safety measures during these periods.
- **Geographical Differences:** The significance of country-specific features points to the need for localized safety improvements. Understanding why certain countries or regions have higher accident levels can guide targeted regulatory and policy changes.
- **Specific Risk Factors:** The "Power lock" risk being highlighted indicates that it is a critical factor in accident prevention. Organizations might need to focus on improving safety measures related to power management and machinery operations to reduce accident rates.

Final assessment: The features selected based on their MI scores provide a clear understanding of the influential factors in the dataset. The results emphasize the importance of temporal and geographical factors, along with specific risks like "Power lock," in predicting accident levels. These insights can guide targeted interventions to improve safety and reduce accident rates in the most affected areas and times. However, it is recommended to conduct further domain-specific analysis and consult with subject matter experts to validate these findings and ensure their practical relevance.

4.5.3 Comparison

Different Approach: Mutual Information focuses on individual feature dependencies with the target variable, while RFE uses a model (Random Forest) to evaluate feature importance based on how well they contribute to the model's prediction accuracy.

Potential for Different Results: It's possible for the two methods to select different features, especially if there are complex interactions between features that might not be captured by mutual information alone.

The optimal method:

- The optimal method depends on the dataset and the specific problem. RFE can be particularly useful when we want to select features that work well together within a specific model (in this case, the Random Forest).
- Model Interpretation: If we need features that have clear, interpretable relationships with the target variable, MI might be more suitable.
- Computational Efficiency: MI is generally faster to compute than RFE, especially for large datasets.

RFE with RandomForest: This method assesses feature importance based on a specific model's performance, making it model-dependent. It integrates the interaction between features inherently as the model evaluates feature subsets.

Mutual Information: This method is model-independent and evaluates the mutual dependence between variables. It provides a measure of how much information the presence/absence of a feature contributes to making the correct prediction on the target variable.

Results Interpretation:

- Features selected through RFE are likely to be more tailored to producing the best results with the RandomForest model, potentially capturing complex patterns and interactions not evident through mutual information.
- MI-based selection offers a more theoretical approach, identifying features that have the strongest statistical relationships with the target variable but without consideration of any specific model dynamics.

Practical Considerations:

- RFE can be computationally more intensive than MI, especially with complex models like RandomForest and large datasets, because it involves retraining the model multiple times.
- MI can be quicker to compute and provides a good starting point for feature selection, especially when computational resources are limited or when a model-agnostic overview is desired.

4.5.4 Conclusion

Both feature selection methods have their place in data science workflows, however the optimal method in this case is Mutual Information (MI) since it is quicker and provides features that have clear relationships with the target variable.

General conclusion

Feature selection is a pivotal yet underutilized technique in risk assessment. Our study illustrates its crucial role in enhancing model performance and stability, thereby facilitating accurate risk prediction. Through the application of RFE and MI, we identified key features that significantly contribute to the predictive power of the used model. This not only underscores the importance of feature selection but also provides a clear methodology for practitioners to improve their risk assessment models.

By integrating feature selection into the risk assessment process, we can achieve more reliable and meaningful predictions, ultimately leading to better-informed decisions and improved safety outcomes. The practical implementation using Python further demonstrates the accessibility and effectiveness of these techniques, encouraging wider adoption in the field of risk assessment.

Obstacles:

- Lack of databases for case study
- Difficulties in Machine Learning

Recommendations:

- Increase awareness and training
- Regularly update feature sets
- Leverage automated feature selection tools
- Evaluate and validate models

By implementing these recommendations, organizations can enhance their risk assessment capabilities, leading to more accurate predictions, better decision-making, and improved safety outcomes. The integration of feature selection as a critical component of risk assessment will contribute to the development of robust and reliable predictive models.

References

[1] - HSA Health and Safety Authority (2023) https://www.hsa.ie/eng/your_industry/fishing/management_of_health_and_safety/risk_assessment/

[2] - Reid Middleton (2017) Civil & Structural Engineering | Surveying https://www.reidmiddleton.com/reidourblog/hazards-vs-risks-whats-the-difference/

[3] - Alexander S. Gillis technical writer and editor (2012) TechTarget https://www.techtarget.com/searchsecurity/definition/risk-assessment

[4] - Jairus Andales SafetyCulture https://safetyculture.com/topics/risk-assessment/

[5] - Adam Wyatt (October 18 ,2019) Employsure Protect https://employsure.com.au/blog/what-is-a-risk-assessment

[6] - Patricia Guevara in SafetyCulture https://safetyculture.com/topics/risk-assessment/risk-assessment-examples/#what-are-risk-assessment-examples

[7] - British Safety council (2023) because experience counts https://www.britsafe.org/training-and-learning/informational-resources/risk-assessments-what-they-are-why-they-re-important-and-how-to-complete-them

[8] - Nirvana Training Academy (2022) study to overcome poverty https://nirvanatraining.com/course/risk-assessment-course/

[9] - OiRA European agency for safety and health at work https://oira.osha.europa.eu/en/roles-and-responsibilities

[10] - ENISA European union agency for cybersecurity https://www.enisa.europa.eu/topics/risk-management/current-risk/risk-management-inventory/rm-process/risk-assessment

[11] - Safran Software Solutions (09 December 2019) https://www.google.com/amp/s/www.safran.com/blog/whats-the-difference-between-qualitative-and-quantitative-risk-analysis%3fhs_amp=true

[12] - Canadian Centre for Occupational Health and Safety CCOHS (2017) https://www.ccohs.ca/oshanswers/hsprograms/hazard/risk_assessment.html

[13] - Danny Butvinik, Chief Data Scientist at NICE Actimize KDnuggets https://www.kdnuggets.com/2021/06/feature-selection-overview.html

[14] - Aman Gupta (21 dec ,2023) Analytics Vidhya https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

[15] - Verónica Bolón-Canedo, Noelia Sánchez-Marõno, Amparo Alonso-Betanzos (March 2015) Artificial intelligence: Foundations, Theory, and algorithms "Feature Selection for High-Dimensional Data" book

[16] - Michał Oleszak Machine Learning Engineer with a statistics background (1st august 2023) https://neptune.ai/blog/feature-selection-methods

[17] - Nilimesh Halder, PhD (15 July 2023) Medium https://pub.aimind.so/feature-selection-for-machine-learning-techniques-benefits-and-challenges-eb29c598f172

[18] - Dr. Nagaraj S. Data Scientist (5 Feb, 2023) Linkedin https://www.linkedin.com/pulse/challenges-feature-selection-data-science-dr-nagaraj-s-#:~:text=Time%20complexity%3A%20Feature%20selection%20can,bottleneck%20in%20the%20modeling%20process

[19] - Jason Brownlee (August 20, 2020) in Data Preparation Machine Learning Mastery https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

[20] - Danny Butvinik (21 May 2021) Analytics Vidhya Medium https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

[21] -Yamini Ane (14 Jun 2023) Analytics Vidhya https://www.analyticsvidhya.com/blog/2022/11/feature-selection-101-the-manual-for-beginners/

[22] - Database link: https://www.kaggle.com/datasets/ihmstefanini/industrial-safety-and-health-analytics-database (2018)