UNIVERSITY OF WEST ATTICA IN COOPERATION WITH UNIVERSITY OF LIMOGES

DEPARTMENT OF ENGINEERING

MASTER IN «ARTIFICIAL INTELLIGENCE AND VISUAL COMPUTING»

Master Diploma Thesis

« Predictive Business Process Monitoring with Automated Machine Learning »

Supervisor: **Dr. Alexandros Bousdekis**

Savvas Kaftantzis

Master of Science, University of West Attica in cooperation with University of Limoges

A T H E N S , 2 0 2 4

Members of the Examination Committee including the rapporteur

The thesis was successfully examined by the following Examination Committee:

	Name	Position	Digital Signature
1	Alexandros Bousdekis	Adjunct Lecturer and Senior Researcher	ΑΛΕΞΑΝΔΡΟΣ ΜΠΟΥΣΔΕΚΗΣ Digitally signed by ΑΛΕΞΑΝΔΡΟΣ ΜΠΟΥΣΔΕΚΗΣ Date: 2024.03.15 09:59:30 +02'00'
2	Paris Mastorocostas	Professor	Paris Mastorokostas Digitally signed by Paris Mastorokostas Date: 2024.02.28 09:06:27 +02'00'
3	Georgios Miaoulis	Professor Emeritus	Georgios MIAOULIS Digitally signed by Georgios MIAOULIS Date: 2024.03.14 13:32:14 +02'00'

Acknowledgments

I express my sincere thanks to Professor A. Bousdekis for his excellent guidance, unwavering support and invaluable knowledge throughout my thesis. His guidance played a decisive role in shaping my research and academic path.

I sincerely thank my parents, Manos and Joanna, for their unconditional support, encouragement and belief in my abilities. Their unwavering commitment was a constant source of motivation.

To my brother Nikos for his support and valuable advice, to my girlfriend Giota for the unlimited support, patience, calmness, understanding and love she gave me throughout this difficult period.

And finally to my circle of friends, I owe their understanding, encouragement and endless support through the challenges and triumphs of this academic endeavor.

Christos, Jim, George, Theodore, Michael, thank you for being by my side.

Your presence made this trip more meaningful.

Contents

1. INTRODUCTION 7

1.1 BACKROUND 7

1.2 CONTRIBUTIONS TO THE FIELD 8

1.3 OBJECTIVES OF THE STUDY 9

2. LITERATURE REVIEW 10

2.1 OVERVIEW OF BUSINESS PROCESS MANAGEMENT 10

2.1.1 TYPES OF BUSINESS PROCESS MANAGEMENT 10

2.1.2 PROS AND CONS OF BUSINESS PROCESS MANAGEMENT 11

2.1.3 FIELDS OF APPLICATION FOR BUSINESS PROCESS MANAGEMENT 12

2.2 PROCESS MINING 12

2.2.1 DISTINGUISHING PROCESS MINING FROM BPM 13

2.2.2. POSITIVE AND NEGATIVE ASPECTS OF PROCESS MINING 13

2.3 MACHINE LEARNING IN PROCESS MINING 14

2.4 AUTOMATED MACHINE LEARNING 15

2.4.1 INTRODUCTION TO AUTOMATED MACHINE LEARNING (AUTO ML) 15

2.4.2 HOW AUTOMATED MACHINE LEARNING WORKS 15

2.4.3 EVOLUTION OF AUTOMATED MACHINE LEARNING 17

2.4.4 POSITIVES AND NEGATIVES OF AUTOMATED MACHINE LEARNING 18

2.4.5 AUTOMATED MACHINE LEARNING IN BUSINESS PROCESS MANAGEMENT 19

2.5 PREDICTIVE BUSINESS PROCESS MONITORING 20

2.5.1 GOALS AND ACHIEVEMENTS 21

3. METHODOLOGY 22

3.1 DATASET 22

3.2 EVENT LOG DATA STRUCTURE 22

3.3 PREPROCESSING OF EVENT LOG DATA 23

3.3.1 PROCESS DISCOVERY USING PROCESS MINING ALGORITHMS 25

3.4 MODEL SELECTION AND CONFIGURATON 27

3.4.1 TPOT LIBRARY 28

3.4.2 TPOT REGRESSION 31

3.4.3 TPOT CLASSIFICATION 33

3.5 EVALUATION METRICS 34

3.5.1 CLASSIFICATION METRICS 34

3.5.2 REGRESSION METRICS 36

4. RESULTS AND ANALYSIS 38
4.1 LIBRARIES AND FRAMEWORKS 38
4.2 DATASET 39
4.3 PRESENTATION OF PREDICTIVE MODELS 43
4.3.1 NEXT EVENT PREDICTION (ANALYSIS AND RESULTS) 43
4.3.2 TIME PREDICTION (ANALYSIS AND RESULTS) 51
4.3.3 REMAINING CYCLE TIME (ANALYSIS AND RESULTS) 54
4.4 COMPARISON OF PREDICTIVE MODELS AND IMPACT ON BUSINESS PROCESSES 58
4.4.1 COMPARATIVE ANALYSIS OF ONE VS TWO NEXT FUTURE PREDICTIONS 58
4.4.2 COMPARATIVE ANALYSIS OF TIME AND REMAINING CYCLE TIME 59
4.4.3 COMPARISON OF RESULTS WITH OTHER RESEARCH ARTICLES 60
5. CONCLUSION AND FURURE WORK 64
5.1 SUMMARY OF FINDINGS 64
5.2 FURURE WORK AND RECOMMENDATIONS 65
REFERENCES 67

Table of Figures

Figure 1 - Auto ML Architecture Search 17
Figure 2 - Metamodel of XES format 23
Figure 3 - Heuristic Network Example 26
Figure 4 - Automated Model by TPOT Library 28
Figure 5 - Confusion Matrix 35
Figure 6 - 10 First Lines of the Dataset (BPI2012) 40
Figure 7 - Heuristic Net for BPI2012 42
Figure 8 - Model Efficiency for Next Event Prediction 45
Figure 9 - Model Efficiency for next two events prediction 50
Figure 10 - Time Prediction: Actual vs Predicted 53
Figure 11 - Remaining Cycle Time: Regression Plot 57
Figure 12 - Comparative Analysis 60

1. INTRODUCTION

In the contemporary business landscape, the intersection of Predictive Business Process Monitoring (PBPM) with Automated Machine Learning (AutoML) represents a cutting-edge paradigm poised to revolutionize how organizations approach operational intelligence. As businesses grapple with the ever-increasing complexity of processes, the amalgamation of predictive analytics and automated machine learning emerges as a transformative strategy. This study delves into the background, contributions to the field, and the overarching objectives, shedding light on the intricate synergy between Predictive Business Process Monitoring and Automated Machine Learning.

1.1 BACKROUND

Traditional business process monitoring primarily involves the retrospective analysis of historical data to identify deviations and inefficiencies after they have occurred. While this approach offers valuable insights for post-mortem analysis, it falls short in addressing the demands of contemporary business environments where agility and proactive decision-making are paramount. PBPM represents an evolution in this domain, introducing a forward-looking dimension that goes beyond mere observation, allowing organizations to predict future process behavior. Traditional approaches to process monitoring often fall short in the face of dynamic and unpredictable workflows. Predictive Business Process Monitoring, as a discipline, addresses this gap by introducing forward-looking analytics, allowing businesses to anticipate events, predict cycle times, and strategically align operations with temporal dynamics. Automated Machine Learning adds a layer of efficiency to this framework, automating the modeling and optimization processes, thereby enhancing the scalability and accessibility of predictive analytics. The synergy between PBPM and AutoML holds immense promise for organizations seeking to navigate the complexities of modern business processes. By combining predictive capabilities with automated model generation, this intersection allows businesses to not only predict future events and process durations but also to do so in a scalable and efficient manner. The fusion of PBPM and AutoML represents a paradigm shift, offering a holistic approach to operational intelligence and decision-making.

In summary, the background of this study is rooted in the evolution of business process monitoring, the challenges faced by organizations in managing complex processes, and the transformative potential of Predictive Business Process Monitoring enhanced by the automation capabilities of Automated Machine Learning.

1.2 CONTRIBUTIONS TO THE FIELD

The integration of Predictive Business Process Monitoring (PBPM) with Automated Machine Learning (AutoML) presents a novel and multifaceted contribution to the field of business process optimization. At its core, this synergy represents a significant advancement in operational intelligence, addressing longstanding challenges and offering transformative capabilities for organizations navigating the complexities of modern business processes.

One notable contribution lies in the establishment of a holistic approach to predictive analytics within the domain of business process management. PBPM, with its predictive capabilities, transcends the limitations of traditional monitoring by offering proactive insights into potential issues and deviations. The integration of AutoML complements this by automating the intricate processes of model generation and optimization. The result is a comprehensive framework that empowers organizations to harness the full potential of predictive analytics without being constrained by the technical intricacies traditionally associated with model development. The synergy between PBPM and AutoML contributes significantly to the democratization of predictive capabilities. Historically, the deployment of advanced predictive models often required specialized expertise in data science. However, the integration of AutoML automates many of the technical aspects, making predictive analytics more accessible to a broader range of professionals. This democratization facilitates the widespread adoption of predictive insights within organizations, empowering decision-makers, business analysts, and process experts to actively contribute to operational excellence without extensive training in data science.

Furthermore, the contributions extend to the real-world applicability and scalability of predictive analytics. The combined framework allows organizations to implement predictive models on event log datasets, a common representation in business processes. This practical application ensures that the insights derived from predictive analytics are not confined to theoretical discussions but are directly applicable to real-world scenarios. Additionally, the automation capabilities of AutoML enhance the scalability of predictive analytics, enabling organizations to deploy models efficiently across diverse business processes, thereby maximizing the impact of operational intelligence.

In essence, the contributions of integrating PBPM with AutoML transcend theoretical advancements, actively shaping the landscape of operational intelligence. This synergy not only addresses existing challenges within business process management but also propels the field forward by making predictive analytics more accessible, applicable, and scalable, thereby empowering organizations to navigate the intricacies of their processes with foresight and efficiency.

1.3 OBJECTIVES OF THE STUDY

To achieve a comprehensive understanding of the interplay between Predictive Business Process Monitoring (PBPM) and Automated Machine Learning (AutoML), this study is guided by four overarching goals:

1. **In-Depth Exploration of Predictive Business Process Monitoring:**

The first goal entails a meticulous examination of the scientific domain of Predictive Business Process Monitoring. This involves an extensive literature review to comprehend the theoretical foundations, methodologies, and existing advancements within PBPM. By conducting a thorough study of the scientific landscape, we aim to establish a robust foundation for subsequent analyses and evaluations.

2. **Assessing the Potential of Automated Machine Learning in PBPM:**

Building upon the insights gained from the exploration of PBPM, the second goal focuses on identifying the potential of Automated Machine Learning (AutoML) in addressing challenges within the PBPM scientific area. This involves a critical analysis of how AutoML can augment and enhance predictive analytics in business processes, with a particular emphasis on its capabilities to streamline modeling processes and optimize predictive outcomes.

3. **Implementation of AutoML Algorithms on Event Log Data Sets:**

The third goal is a hands-on exploration of AutoML's capabilities. We aim to implement AutoML algorithms on datasets formatted as event logs, a common representation in business process contexts. This step involves the practical application of AutoML methodologies to produce predictions in the realm of business processes. Through this implementation, we seek to understand the adaptability and efficacy of AutoML in real-world scenarios.

4. **Evaluation of AutoML Algorithm Results:**

The final goal centers on the critical evaluation of the results derived from the application of AutoML algorithms. This involves a systematic assessment of the predictive accuracy, efficiency, and overall performance of AutoML in the context of business process monitoring. By rigorously evaluating the outcomes, we aim to provide empirical evidence of the effectiveness of AutoML as a tool for enhancing predictive capabilities within business processes.

In essence, these goals collectively form a comprehensive research framework, guiding the study towards a nuanced understanding of how the marriage of Predictive Business Process Monitoring and Automated Machine Learning can contribute to operational intelligence and decision-making within the complex dynamics of modern business processes.

2. LITERATURE REVIEW

2.1 OVERVIEW OF BUSINESS PROCESS MANAGEMENT

Business Process Management (BPM) serves as a structured methodology for organizations seeking to optimize their operational efficiency, effectiveness, and adaptability. It encompasses a systematic approach to the design, execution, monitoring, and improvement of business processes. [1] The ultimate goal is to align these processes with organizational objectives, fostering continuous improvement and overall enhanced performance. A key piece to achieving this business process improvement –understands the BPM lifecycle.

The life cycle of BPM involves a series of interconnected stages, each contributing to the holistic management and optimization of business processes. The initial Design phase involves identifying and documenting existing processes, followed by the creation of detailed process models during the Modeling stage. [2] The Execution phase sees the implementation of designed processes in real-world business environments, often facilitated by BPM software and automation tools. The Monitoring stage entails continuous tracking of key performance indicators (KPIs) and process metrics. Finally, the Optimization phase utilizes monitoring data to identify areas for improvement, implementing changes to enhance efficiency and effectiveness. [2]

2.1.1 TYPES OF BUSINESS PROCESS MANAGEMENT

Business Process Management (BPM) encompasses diverse approaches tailored to address specific organizational needs, fostering efficiency and effectiveness. One prominent category is Process-Centric BPM, which centers on the refinement and optimization of individual business processes. [1] Organizations employing this type of BPM delve deeply into the intricacies of specific operational functions, aiming to enhance efficiency, reduce bottlenecks, and improve overall performance within distinct functional areas.

In contrast, Human-Centric BPM places a spotlight on processes that involve human interactions. [2] This approach acknowledges the vital role played by individuals in the execution of processes, emphasizing collaboration, communication, and decision-making. Human-Centric BPM seeks to optimize not only the procedural aspects but also the human elements within workflows, recognizing that organizational success often hinges on effective human collaboration. [2]

Another facet of BPM is Integration-Centric BPM, which concentrates on the seamless integration of diverse systems and technologies. In a rapidly evolving technological landscape, organizations rely on numerous applications and tools. [2] Integration-Centric BPM ensures the harmonious flow of data and communication between these systems, promoting interoperability and efficiency. It addresses the challenges associated with disparate technologies, enabling a more unified and streamlined operational environment.

These distinct types of BPM are not mutually exclusive; organizations often adopt a hybrid approach based on their unique requirements. For instance, a comprehensive BPM strategy may incorporate elements of Process-Centric, Human-Centric, and Integration-Centric BPM, offering a nuanced and tailored solution that aligns with the organization's overarching goals and objectives. [2] The flexibility inherent in these BPM types allows organizations to adapt their approach to the specific characteristics and demands of different processes within the business ecosystem. As a result, BPM becomes a dynamic and evolving framework, capable of addressing the multifaceted nature of modern business operations.

2.1.2 PROS AND CONS OF BUSINESS PROCESS MANAGEMENT

Business Process Management (BPM) offers a range of advantages, making it a widely adopted approach for organizational optimization. One of the key benefits is the potential for significant efficiency improvement. [1] BPM enables organizations to streamline their operations, eliminate redundancies, and enhance overall workflow efficiency. By identifying and addressing bottlenecks or inefficiencies within processes, businesses can experience cost savings and improved resource utilization. Furthermore, BPM provides a framework for adaptability, allowing organizations to respond promptly to changing market conditions. In a dynamic business environment, the ability to modify and optimize processes quickly is a valuable asset. [1] BPM facilitates this agility by providing a structured methodology for continuous improvement, ensuring that organizations can align their operations with evolving strategic objectives. Improved customer satisfaction is another notable advantage of BPM. As processes become more efficient and customer-centric, the overall experience for clients and stakeholders is enhanced. Whether it's in terms of faster service delivery, more accurate information, or smoother interactions, BPM contributes to heightened customer satisfaction, fostering loyalty and positive relationships. [2]

Despite these advantages, the adoption of BPM is not without challenges. One significant drawback is the potential for high implementation costs. [1] Integrating BPM systems and software into existing organizational structures may require substantial financial investment. This cost factor can be a barrier for smaller organizations or those operating on tighter budgets, limiting the accessibility of BPM solutions. Resistance to change from employees is another notable challenge. Employees accustomed to established processes may resist alterations introduced by BPM initiatives. [1] This resistance can lead to implementation challenges, necessitating effective change management strategies to ensure a smooth transition and acceptance of new methodologies. Lastly, the inherent complexity of managing and optimizing business processes can pose difficulties. [2] Organizations may find it challenging to navigate the intricacies of BPM, requiring specialized skills and resources for successful implementation. The complexity factor reinforces the importance of thorough planning, employee training, and ongoing support to maximize the benefits of BPM.

In summary, the pros and cons of BPM reflect its multifaceted nature. While it offers substantial advantages in terms of efficiency, adaptability, and customer satisfaction, organizations must carefully consider potential challenges such as implementation costs, employee resistance, and the inherent complexity associated with managing and optimizing intricate business processes. [2]

2.1.3 FIELDS OF APPLICATION FOR BUSINESS PROCESS MANAGEMENT

The application of BPM extends across diverse industries, each benefiting from the optimization of specific processes. [1] In manufacturing, BPM aids in streamlining production processes, managing inventory, and optimizing supply chain operations. In finance and banking, BPM finds utility in streamlining loan approval processes, enhancing transaction handling, and improving risk management. The healthcare sector leverages BPM to optimize patient care processes, streamline appointment scheduling, and manage medical records efficiently. Retail organizations benefit from BPM in areas such as inventory management, order processing, and customer service enhancement. [1] Information technology sectors utilize BPM to manage software development processes, enhance IT service delivery, and facilitate seamless system integrations.

In conclusion, Business Process Management stands as a multifaceted approach, encompassing various types, a comprehensive life cycle, and presenting both advantages and challenges. Its applications span across industries, driving improvements in diverse business processes.

2.2 PROCESS MINING

Process Mining is a cutting-edge analytical discipline that leverages data logs generated by information systems to gain insights into business processes. [4] Unlike traditional Business Process Management (BPM), which relies heavily on predefined models, Process Mining extracts process information directly from event logs, offering a more dynamic and data-driven approach to process analysis. Process Mining plays a pivotal role in enhancing process transparency, offering organizations an in-depth understanding of how their processes truly operate. By analyzing event data, it provides a factual representation of workflow execution, facilitating the identification of inefficiencies, compliance issues, and potential areas for improvement. [4] This data-driven insight is crucial for organizations striving to enhance operational efficiency, compliance, and overall performance.

2.2.1 DISTINGUISHING PROCESS MINING FROM BPM

Process Mining and Business Process Management (BPM) represent distinct approaches to understanding and optimizing business processes. While BPM focuses on the design, modeling, and improvement of processes based on predefined structures, Process Mining takes a different, more data-centric approach. [3] In the realm of BPM, organizations traditionally rely on predefined process models, emphasizing the proactive design and optimization of processes before their execution. In contrast, Process Mining operates retrospectively. Instead of relying on predefined models, it extracts insights directly from event logs generated by information systems during actual process executions. [3] By analyzing these logs, Process Mining reconstructs the real-life sequence of activities, providing an accurate and factual representation of how processes are executed within an organization.

One key distinction lies in the source of information. BPM starts with a predefined model and seeks to align processes with this model. Process Mining, on the other hand, derives its understanding directly from the data generated during the execution of processes. [3] This allows Process Mining to uncover variations, deviations, and nuances that may exist in real-world process execution but are not accounted for in BPM models. While BPM offers a structured and planned approach to process optimization, it may struggle to capture the dynamic and evolving nature of actual process execution. Process Mining, by directly leveraging data logs, offers a more adaptive and reactive strategy. [3] It excels in revealing the true intricacies of processes, allowing organizations to identify inefficiencies, bottlenecks, and deviations that might escape the purview of traditional BPM methodologies.

In essence, the key difference lies in their orientation – BPM is proactive and model-centric, while Process Mining is retrospective and data-driven. [3] This distinction empowers organizations to complement their strategic process design with a more granular and adaptive understanding derived from real-world execution, providing a comprehensive view essential for continuous improvement.

2.2.2. POSITIVE AND NEGATIVE ASPECTS OF PROCESS MINING

One of the significant advantages of Process Mining is its ability to provide an objective and comprehensive view of processes. [3] It offers a clear visualization of how tasks are executed in reality, allowing organizations to make informed decisions based on factual evidence. The transparency afforded by Process Mining enables organizations to identify and rectify inefficiencies, optimize resource allocation, and ensure adherence to compliance standards. Moreover, Process Mining is instrumental in facilitating continuous improvement. By continuously monitoring and analyzing processes, organizations can identify opportunities for optimization, ensuring that their operations stay agile and responsive to changing

business dynamics. [3] This adaptability is crucial in the modern business landscape, where organizations must be able to adjust quickly to stay competitive.

However, the implementation of Process Mining is not without challenges. Integration with existing systems and data sources can be complex, and organizations may face hurdles in ensuring data accuracy and completeness. [3] Additionally, there may be concerns related to data privacy and security, as Process Mining relies on analyzing detailed event logs, potentially containing sensitive information. Furthermore, interpreting Process Mining results requires a nuanced understanding of both the technical and business aspects. Misinterpretation of the mined data can lead to misguided decisions. [3] Additionally, organizations may encounter resistance from employees who might perceive Process Mining as intrusive or fear potential repercussions from increased process transparency.

In conclusion, Process Mining offers a transformative approach to understanding and optimizing business processes. Its importance lies in its ability to provide real-time, data-driven insights, enhancing transparency, efficiency, and adaptability. While it presents substantial positive aspects, organizations must navigate potential challenges related to data integration, privacy concerns, and the need for a nuanced interpretation of results to fully harness its benefits.

2.3 MACHINE LEARNING IN PROCESS MINING

The intersection of Machine Learning (ML) and Process Mining represents a powerful synergy, where advanced analytical techniques enhance the capabilities of traditional process analysis. [5] Unlike conventional process analysis methods, which may rely on predefined rules and models, ML in Process Mining introduces a data-driven and adaptive dimension to uncover hidden patterns and insights within vast datasets. One significant application of ML in Process Mining is the automated discovery of process models. Traditional Process Mining techniques extract process models from event logs [4], but ML algorithms take this a step further. Machine Learning algorithms can autonomously identify patterns, relationships, and variations within the data, contributing to the automatic generation and refinement of process models. This not only accelerates the analysis process but also ensures a more dynamic and responsive adaptation to evolving business processes.

Predictive analytics is another realm where ML contributes significantly to Process Mining. By leveraging historical process data, ML algorithms can forecast future process behaviors, identifying potential bottlenecks or deviations before they occur. [5] This predictive capability empowers organizations to proactively address issues, optimize resource allocation, and enhance overall process efficiency. Furthermore, ML enhances the precision of anomaly detection in Process Mining. Anomalies, deviations, or outliers within process execution can be indicative of inefficiencies or potential risks. ML algorithms excel in recognizing subtle patterns within large datasets, facilitating the early identification of irregularities that may go unnoticed with traditional analysis methods. [5] Despite these advantages, the integration of ML into Process Mining is not without challenges. The

complexity of ML algorithms and the need for large, high-quality datasets can pose implementation hurdles. [5] Moreover, interpreting the output of ML models requires a nuanced understanding of both the domain-specific processes and the intricacies of the employed algorithms. Organizations need to strike a balance between the predictive power of ML and the interpretability required for actionable insights.

In summary, the incorporation of Machine Learning into Process Mining introduces a data-driven paradigm shift. From automated process model discovery to predictive analytics and anomaly detection, ML enriches the capabilities of Process Mining, providing organizations with advanced tools to glean deeper insights and optimize their business processes. The challenges in implementation notwithstanding, the fusion of ML and Process Mining promises a more dynamic, adaptive, and effective approach to process analysis and optimization.

2.4 AUTOMATED MACHINE LEARNING

2.4.1 INTRODUCTION TO AUTOMATED MACHINE LEARNING (AUTO ML)

Automated Machine Learning (AutoML) represents a significant advancement in the field of machine learning, streamlining and democratizing the process of developing predictive models. The primary goal of AutoML is to automate various aspects of the machine learning workflow, making it accessible to a broader audience and significantly reducing the time and expertise required for model development.

The primary difference between AutoML and traditional machine learning lies in the level of automation. Classic machine learning often requires extensive manual intervention in tasks such as feature engineering, algorithm selection, and hyperparameter tuning. AutoML, on the other hand, automates these processes, aiming to optimize the entire workflow seamlessly. This shift towards automation democratizes access to machine learning, enabling a wider range of professionals to leverage its benefits. [8]

2.4.2 HOW AUTOMATED MACHINE LEARNING WORKS

Auto ML operates on the principle of automating the end-to-end process of building machine learning models, traditionally a labor-intensive and specialized task. [6] The workflow of Auto ML typically involves several key steps:

- **Data Preprocessing**: Auto ML systems handle data preprocessing tasks, such as handling missing values, encoding categorical variables, and scaling features. [8] This automation ensures that the data is appropriately prepared for modeling.

- **Feature Engineering**: Automated techniques are employed to generate and select relevant features from the dataset. [8] This involves transforming raw data into a format that enhances the performance of machine learning algorithms.

- **Algorithm Selection**: Auto ML explores a range of machine learning algorithms, from simple to complex, to identify the most suitable models for a given prediction task. This automated selection process considers the characteristics of the data and the nature of the prediction problem. [8]

- **Hyperparameter Tuning**: Hyperparameters are the configuration settings that govern the behavior of machine learning algorithms. Auto ML systematically explores different combinations of hyperparameter values to optimize the model's performance. [8]

- **Model Evaluation and Selection**: Auto ML evaluates the performance of multiple models using predefined metrics, such as accuracy or precision-recall. [8] The best-performing model is then selected for further use.

- **Model Deployment**: Once a satisfactory model is identified, Auto ML facilitates the deployment of the model into production environments. This often includes generating code for deployment or providing interfaces for seamless integration with other systems. [8]

The key innovation lies in the automation of these steps, allowing users to interact with an intuitive interface, upload their data, and receive a fully optimized machine learning model [7]. Auto ML platforms abstract away the complexities of model development, making the power of predictive analytics accessible to a broader audience. However, we must note that the human factor is important no matter what. [7] While Automated Machine Learning (Auto ML) automates many aspects of the machine learning workflow, including algorithm selection, hyperparameter tuning, and model evaluation, it is common for users to perform their own data preprocessing before using Auto ML. Data preprocessing is a critical step in the machine learning pipeline, including tasks such as handling missing values, coding categorical variables, scaling features, and addressing any other data quality issues. [8] Auto ML platforms automate some aspects of data preprocessing, but users often need to ensure that their data is in an appropriate format and meets certain standards before feeding it into an Auto ML system.

In essence, Auto ML is a catalyst for innovation, eliminating entry barriers to machine learning by automating the intricate tasks associated with model development. It empowers organizations to leverage the full potential of their data without necessitating extensive expertise in data science or machine learning. [7] This paradigm shift holds promise for a future where predictive analytics becomes an integral part of decision-making across diverse sectors.

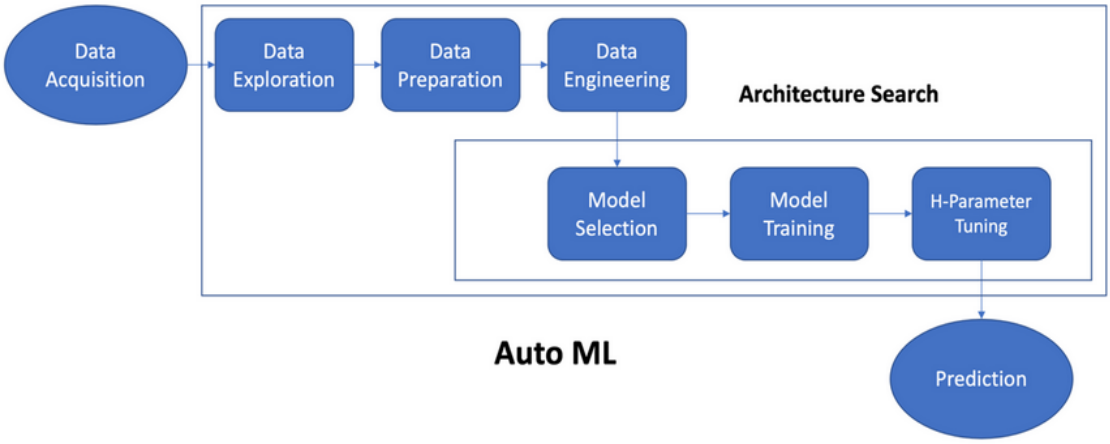

Figure 1 - Auto ML Architecture Search

2.4.3 EVOLUTION OF AUTOMATED MACHINE LEARNING

The evolution of Automated Machine Learning represents a dynamic journey that has significantly transformed the landscape of machine learning over the years. From its early beginnings to its current state, the evolution of AutoML can be traced through several key phases, marked by advancements in algorithms, methodologies, and the democratization of machine learning capabilities. The nascent phase of AutoML saw a focus on simplifying the machine learning process for users with limited expertise. [7] Early tools aimed to automate basic tasks such as algorithm selection and hyperparameter tuning. These tools provided a user-friendly interface to make machine learning more accessible, allowing users to experiment with predictive modeling without delving into the intricacies of algorithms or coding. [8] As the field progressed, the integration of hyperparameter optimization became a pivotal advancement. Hyperparameter tuning, a critical step in achieving optimal model performance, was automated to enhance efficiency. AutoML platforms started employing advanced optimization algorithms to systematically explore hyperparameter spaces, fine-tuning models for better predictive accuracy. [8]

The mid-phase of AutoML evolution witnessed the integration of ensemble methods and model stacking. AutoML systems began incorporating ensemble learning techniques, leveraging the strengths of multiple models to improve predictive performance. [8] Model stacking, which involves combining the outputs of diverse models, became a prevalent strategy, further enhancing the robustness and generalization capabilities of AutoML models. Recent years have seen a paradigm shift towards end-to-end automation and democratization of machine learning. [7] Comprehensive AutoML platforms emerged, offering users a one-stop solution for the entire machine learning workflow. These platforms automated not only algorithm selection and hyperparameter tuning but also data

preprocessing, feature engineering, and model deployment. [7] Democratization became a focal point, empowering a broader audience, including business analysts and domain experts, to leverage the full potential of machine learning without extensive technical expertise.

The integration of Neural Architecture Search (NAS) represents a cutting-edge development in AutoML. NAS automates the design of neural network architectures, optimizing them for specific tasks. [6] This advancement has particularly impacted deep learning applications, allowing AutoML systems to automatically discover and fine-tune neural network structures, thereby pushing the boundaries of model complexity and performance. In response to the growing importance of model interpretability, recent advancements in AutoML include a focus on explainability. [8] AutoML platforms now incorporate techniques to generate interpretable models and provide insights into the decision-making process. This addresses concerns related to the "black-box" nature of complex machine learning models.

In conclusion, the evolution of AutoML reflects a journey from simplifying the machine learning process to achieving comprehensive end-to-end automation and democratization. The field continues to advance, incorporating state-of-the-art techniques such as Neural Architecture Search and emphasizing the importance of model explainability, making AutoML an integral part of the contemporary machine learning landscape.

2.4.4 POSITIVES AND NEGATIVES OF AUTOMATED MACHINE LEARNING

Automated Machine Learning (AutoML) offers several positive attributes that have contributed to its widespread adoption. One of its primary advantages lies in its efficiency and time-saving capabilities. By automating tasks such as algorithm selection, hyperparameter tuning, and feature engineering, AutoML significantly reduces the time and effort required for model development, enabling swift deployment. [8] Another key benefit is the accessibility and democratization it brings to machine learning. User-friendly interfaces empower individuals across various domains, including business analysts and domain experts, to leverage advanced predictive modeling without extensive technical expertise. [7] Furthermore, AutoML often leads to optimized model performance, as its systematic exploration of the model space results in superior hyperparameter configurations and algorithm selections. Additionally, its adaptability to changing data ensures that models remain accurate and relevant over time. The incorporation of ensemble learning techniques enhances overall model robustness, providing more reliable predictions. [7]

Despite its numerous advantages, AutoML is not without challenges. One significant concern is the interpretability of automated models. The complexity introduced by the automated processes can result in models that are challenging to interpret, posing issues in scenarios where understanding the decision-making process is crucial, such as in healthcare or finance. [8] Another drawback is the potential limitation in customization for experts in machine learning and data science. Experienced practitioners may find certain AutoML

platforms restrictive in terms of flexibility and control, limiting their ability to fine-tune models manually. [7] The automated nature of some advanced models, particularly those involving deep learning or complex ensembles, can render them as "black-box" models. This lack of transparency raises concerns about accountability and the ability to understand model decisions. Additionally, AutoML's effectiveness is heavily reliant on the quality of input data; if the data is noisy or biased, the automated processes may not yield optimal results. [7] Moreover, the resource intensiveness of training complex models, especially those generated through neural architecture search, can be a computational challenge. Lastly, there is a risk of overfitting if the automated processes are not carefully controlled, emphasizing the importance of proper validation and testing procedures in AutoML applications. Balancing these positives and negatives requires a thoughtful consideration of specific use cases, data quality, and the expertise of users involved in the machine learning process.

2.4.5 AUTOMATED MACHINE LEARNING IN BUSINESS PROCESS MANAGEMENT

The integration of Automated Machine Learning (AutoML) into business processes has emerged as a strategic imperative, revolutionizing how organizations approach decision-making, efficiency, and innovation. AutoML's analytical prowess in this context lies in its capacity to streamline and optimize predictive modeling without the necessity of extensive data science expertise.

One of the primary analytical advantages of AutoML is its ability to enhance efficiency in business processes. By automating complex tasks such as algorithm selection, hyperparameter tuning, and feature engineering, AutoML accelerates the model development lifecycle. [8] This efficiency translates into quicker insights, enabling organizations to make timely, data-driven decisions that directly impact operational performance. Also AutoML empowers organizations to embrace a data-driven culture in their decision-making processes. By leveraging machine learning models trained on historical data, businesses gain the analytical capability to forecast trends, identify patterns, and make informed decisions based on quantitative insights. [8] This analytical shift facilitates a proactive approach to problem-solving within the context of business processes. Business processes are inherently dynamic, influenced by evolving market conditions, consumer behavior, and internal factors. AutoML's analytical adaptability shines in this context. The ability to continuously monitor and retrain models ensures that predictive analytics remain robust and reflective of the current business landscape. [9] This adaptability is crucial for organizations seeking to navigate and thrive in a rapidly changing environment.

From an analytical standpoint, AutoML aids in the optimal allocation of resources within business processes. By automating resource-intensive tasks and refining models for efficiency, organizations can strategically deploy their resources, both human and computational. [7] This analytical optimization contributes to cost-effectiveness and ensures that resources are directed towards initiatives that yield maximum impact. Analytically, AutoML acts as a catalyst for continuous improvement in business processes. The iterative

nature of AutoML processes allows organizations to learn from model performance, identify areas for enhancement, and refine strategies over time. [8] This analytical feedback loop supports a culture of continuous improvement, fostering resilience and adaptability in the face of changing business landscapes.

In conclusion, the analytical impact of AutoML in business processes is multifaceted, encompassing efficiency gains, data-driven decision-making, adaptability, resource optimization, enhanced predictive power, and a commitment to continuous improvement. As organizations increasingly recognize the analytical potential of AutoML, it becomes a cornerstone in their quest for operational excellence, providing a data-centric lens through which to analyze, optimize, and innovate in the intricate landscape of business processes.

2.5 PREDICTIVE BUSINESS PROCESS MONITORING

In the dynamic landscape of modern business, where agility and efficiency are paramount, Predictive Business Process Monitoring (PBPM) has emerged as a pivotal strategy for organizations seeking to elevate their operational intelligence. This innovative approach transcends traditional process monitoring by harnessing the power of predictive analytics to anticipate, identify, and address potential issues before they impact operational efficiency. [4] As we embark on understanding the significance of PBPM, its operational mechanics, and the overarching goals it aims to achieve, we unravel a transformative tool that holds the key to unlocking unparalleled insights into business processes.

The importance of PBPM for businesses cannot be overstated. In an era where every competitive edge matters, the ability to proactively manage and optimize business processes is a strategic imperative. PBPM provides organizations with the foresight needed to navigate complex operational landscapes, mitigate risks, and seize opportunities. By shifting from reactive to proactive process management, businesses can not only enhance efficiency but also bolster their capacity for strategic decision-making. [9] At its core, PBPM leverages advanced predictive modeling techniques to analyze historical data, real-time process metrics, and patterns of behavior within business processes. Machine learning algorithms, time series analysis, and anomaly detection are intricately woven into the fabric of PBPM. [9] This analytical ensemble enables organizations to predict future process behavior, identify deviations from expected norms, and continuously monitor the health of their operational workflows.

2.5.1 GOALS AND ACHIEVEMENTS

The primary goal of PBPM is to usher in a new era of operational excellence by providing actionable insights that go beyond mere monitoring. [9] Organizations aspire to achieve several key objectives through PBPM, including:

- **Proactive Issue Identification**: PBPM aims to shift from reactive problem-solving to proactive issue identification. [9] By anticipating potential challenges, organizations can take preemptive actions to prevent disruptions.

- **Enhanced Efficiency**: Through continuous monitoring and optimization, PBPM contributes to enhanced process efficiency. [9] It identifies areas for improvement, streamlines workflows, and supports resource allocation for maximum impact.

- **Data-Driven Decision-Making**: PBPM promotes a culture of data-driven decision-making by providing decision-makers with predictive insights. [9] This analytical approach empowers leaders to make informed choices, aligning strategic decisions with the anticipated trajectory of business processes.

- **Adaptability to Change**: The adaptability of PBPM to changing conditions ensures that predictive insights remain accurate and relevant. [9]This is crucial in a business environment where flexibility and adaptability is the key to staying competitive.

It is important to point out that Predictive Business Process Monitoring goes beyond traditional process management by introducing a long-term dimension to operational analytics. Within this innovative framework, critical aspects such as: Next Event Prediction, Remaining Cycle Time and Time Prediction stand out. These elements not only redefine how businesses approach process optimization, but also pave the way for a more proactive and strategically aligned business landscape. By harnessing the power of predictive analytics, organizations can navigate the intricate temporal landscape of their business processes with precision, maximizing efficiency and making informed decisions that shape the future of their operations. [10]

In essence, Predictive Business Process Monitoring is not merely a technological innovation but a strategic imperative for businesses aiming to thrive in the face of complexity. By embracing the power of prediction, organizations can embark on a journey towards operational excellence, where insights gleaned from data become a compass guiding them through the intricacies of the modern business landscape.

3. METHODOLOGY

3.1 DATASET

The chosen dataset for this research is part of the Business Process Intelligence (BPI) collection, comprising a series of event logs capturing various business processes. These event logs provide a comprehensive record of activities within different organizational workflows. The dataset encompasses a diverse range of events, allowing for a broad exploration of process dynamics and temporal sequences. The events recorded in these logs span from the initiation to the completion of various business processes, offering valuable insights into the patterns and the trends associated with these workflows.

The decision to utilize BPI datasets stems from their significance in the field of business process analysis and monitoring. These datasets serve as valuable resources for studying the behaviors and trends within diverse processes, aiding in the development of predictive models and analytical frameworks. The overarching goal is to extract meaningful information from these event logs, contributing to a deeper understanding of process efficiency, bottlenecks, and potential areas for optimization within organizational workflows.

3.2 EVENT LOG DATA STRUCTURE

Most BPMSs and also other enterprise systems record events corresponding to the execution of work items and other relevant events such as the receipt of a message related to a given case of a process. These event records can be extracted from the database of the BPMS or enterprise system and represented as an event log. An event log is a collection of timestamped event records. Each event record tells us something about the execution of a work item (and hence a task) of the process (e.g., that a task has started or has been completed), or it tells us that a given message event, escalation event, or other relevant event has occurred in the context of a given case in the process. For example, an event record in an event log may capture the fact that Chuck has confirmed a given purchase order at a given point in time. So a single event has a unique event ID. Furthermore, it refers to one individual case, it has a timestamp, and it shows which resources executed which task.

Simple event logs are commonly represented as tables and stored in a Comma-Separated-Values (CSV) format. However, in more complex event logs, where the events have data attributes (e.g., the amount of a loan application, the shipping address of a purchase order), a flat CSV file is not a suitable representation. A more versatile file format for storing and exchanging event logs is the eXtensible Event Stream (XES) format standardized by the IEEE Task Force on Process Mining. The majority of process mining tools can handle event logs in XES. The structure of an XES file is based on a data model, partially depicted in Figure 2. An XES file represents an event log. It contains multiple traces, and each trace can contain multiple events. All of them can contain different attributes. [1]

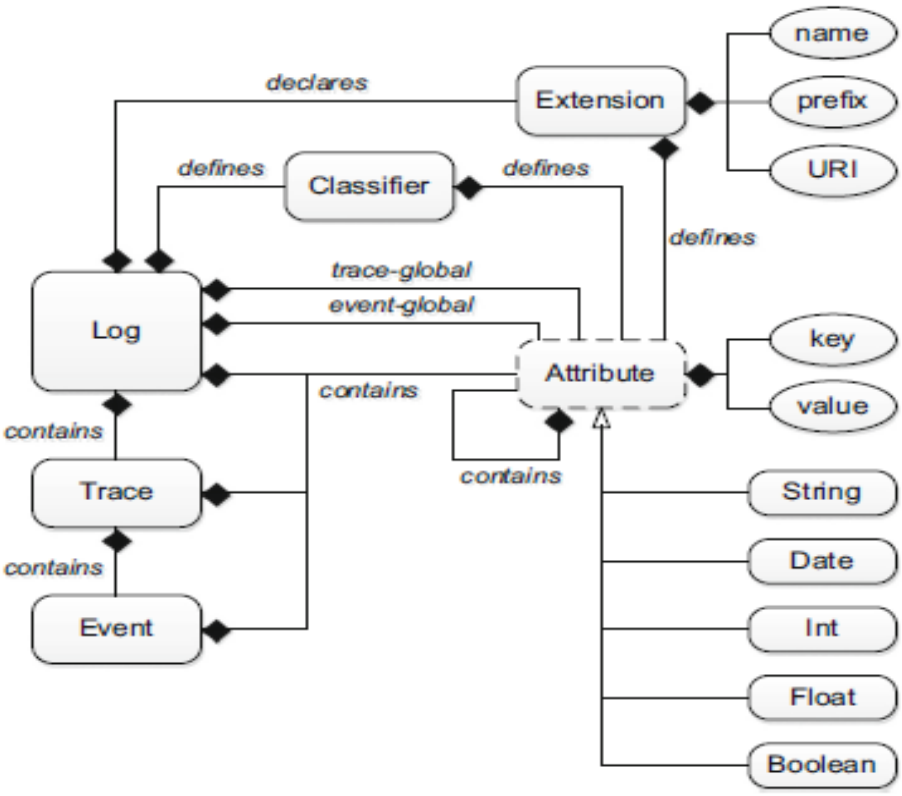

Figure 2 - Metamodel of XES format

An Attribute has to be a string, date, int, float, or boolean element as a key-value pair. Attributes have to refer to a global definition. There are two global elements in the XES file: one for defining trace attributes, the other for defining event attributes. Several classifiers can be defined in XES. A classifier maps one of more attributes of an event to a label that is used in the output of a process mining tool. In this way, for instance, events can be associated with tasks. [1]

3.3 PREPROCESSING OF EVENT LOG DATA

First, an analysis of each data to give the categorical values, the continuous values, the blank values and also the type of each variable, is important. This will help to convert the variable categories when we need them, to convert them into numerical values and scale them parallel to the continuous values. Thus, the categorical variables in the event log data were subjected to an exhaustive coding process. This transformative step was and is decisive for the corresponding incorporation of categorical features into machine learning algorithms. The goal was to enable meaningful interpretation and use of these variables in subsequent prediction models. At the same time, the numerical features of the data set were subjected to scaling procedures. Standardizing the sizes of these numerical variables

emerged as a critical measure to mitigate the potential impact of different scales on the performance of machine learning models. This harmonization ensured a balanced contribution of different features to the overall prediction framework.

Regarding the treatment of missing values, the methodology we followed was either to delete a column that has too many empty values but at the same time in our predictions it will not be used at all and will be left out of the model, or we filled the empty values with the exact previous value. This happened because some time points corresponding to an event had an empty value after converting them to datetime due to a problem with the conversion function itself and so we had to find a way to fill these values. This method was to fill them with the exact previous value from each empty value. It is important to note that certain columns or values may be in a different language from the language we either know or want to work with. This happens because the datasets we may work with are from different countries, universities, companies, organizations, groups, etc. Thus, an exploration of the prices and in terms of the language would be important so that we can convert it to the language we want.

A very important step is to see the distribution of time with everything formatted in the correct form, column of time but also to break this time into smaller pieces. In other words, let's break time into hours of the day, day of the week, month. In this way we will be able to do a very specific exploration and get 'into' the time variable and if we extract important information, where in a different case we would never learn, by seeing the distribution of these values. By extracting the day, month, and time, we can capture temporal patterns in our data. Many business processes vary based on the time of day, day of the week, or month. Note that these exported components can serve as additional features for our model. For example, we may discover that certain events or patterns are more prevalent during certain hours or days, providing valuable information for your predictive model. It will also obviously help if our dataset exhibits seasonality, understanding the month and day can help our model for variations related to specific seasons or recurring events.

There are many processes that can exist in a dataset, especially in the form of event logs. In these processes it is important to see the frequency of transitions of each activity separately because determining the frequency of transitions helps in process discovery, which involves revealing the actual sequence of activities and events within a business process. Knowing which transitions occur frequently provides insight into the most common paths taken by instances. Indeed unusual or infrequent transitions may indicate anomalies or deviations from the expected process flow. Monitoring transition frequencies helps identify irregular patterns that may require further investigation. Similarly, high-frequency transitions may indicate critical paths or stages in a process. Identifying frequent transitions allows you to identify potential bottlenecks or areas where processes may slow down, providing opportunities for optimization.

Controlling the event log format for start and end activities is crucial for process mining and business process analysis for several reasons. At first glance, identifying start and end activities help define the boundaries of a process. It clarifies where a process begins and ends, allowing a clear understanding of the entire workflow. Since initiation activities

represent the beginning of a process, while termination activities indicate its completion, then analyzing these events will help to highlight sequences of activities. Finally we ensure the verification of the occurrence of the start and end activities by ensuring the completeness and validity of the event log.

A final very important step is variance analysis in the context of business process analysis and process mining. A process variant is a unique path from the beginning of a process to the end. Briefly, a process variant is a unique sequence of activities, like a process 'map'. Thus variance analysis will allow us to measure performance fluctuations within a business process. Understanding how activities deviate from expected norms provides insights into efficiency and effectiveness. In detecting and investigating unexpected behavior or deviations from standard operating procedures. Consistent and low variance indicates a stable process, while high variance may indicate instability or unpredictability.

3.3.1 PROCESS DISCOVERY USING PROCESS MINING ALGORITHMS

The Process Discovery phase plays a central role in uncovering the intricacies of business processes from event log data. Leveraging process mining algorithms is a key step toward extracting meaningful information and patterns in any event log dataset. A powerful technique used in this phase is the implementation of a heuristic network through the heuristic miner. Understanding the dynamics of business processes is fundamental to organizations and businesses seeking to improve efficiency, optimize workflows and identify areas for improvement. Process Discovery enables us to uncover the true flow of activities as recorded in event logs, shedding light on both expected and unexpected process behaviors.

So we use a heuristic network for our analysis. A heuristic network is a graphical representation of observed behavior in an event log, providing a visual abstraction of process structure. Unlike traditional process models, heuristic networks embrace flexibility and capture the inherent uncertainty present in real-world processes. They are very important because they handle incomplete or noisy data, making them suitable for the diverse and dynamic nature of event logs. A heuristic miner will give a better result than applying the alpha algorithm because of the noise. The Alpha Miner is an algorithm designed for discovering a process model without assuming any a priori knowledge of the underlying process structure. It leverages the direct succession relation present in event logs to construct a Petri net, a formalism that captures both concurrency and synchronization in process execution. So, the heuristic net gives more information on the reliability of the used paths and therefore is more suitable for determining the main process. [1]

A very simple example of heuristic network is below:

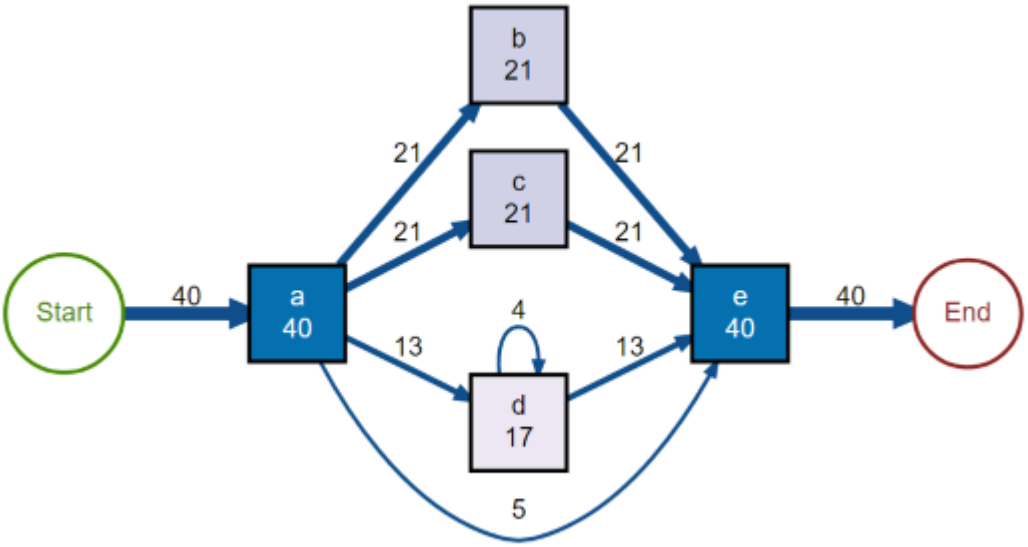

Figure 3 - Heuristic Network Example

Heuristic Miner is a widely adopted algorithm for constructing heuristic nets from event logs. Its significance lies in its ability to discover process models without relying on a predefined process model structure. Instead, it intelligently infers the most likely process flow by analyzing the temporal relationships and dependencies present in the event log data.

- Key Advantages of Heuristic Miner:

1. *Adaptability to Real-world Complexity:*

Heuristic Miner excels in scenarios where processes exhibit variations, exceptions, or ad-hoc deviations from a rigid structure, making it an ideal choice for modeling real-world complexity.

2. *Handling Noisy Data:*

Incomplete or noisy event logs are common in practical scenarios. The Heuristic Miner's resilience to such imperfections ensures that the discovered heuristic net accurately reflects the observed behavior.

3. *Efficient Handling of Large Event Logs:*

Scaling to large datasets is crucial for practical applications. The efficiency of Heuristic Miner makes it suitable for processing extensive event logs, facilitating analysis in real-world, large-scale business environments.

4. *Incremental Discovery:*

As processes evolve, the Heuristic Miner allows for incremental discovery, adapting to changes over time and supporting continuous process improvement efforts.

In summary, the implementation of a heuristic net through the application of the Heuristic Miner is a valuable approach in Process Discovery. It not only accommodates the complexities inherent in real-world processes but also provides a foundation for subsequent analysis and improvement initiatives.

3.4 MODEL SELECTION AND CONFIGURATON

In the methodology of predictive business process monitoring, careful attention is paid to the selection and configuration of models, a process facilitated by the application of automated machine learning (AutoML). The TPOT library, a notable component of this methodology, stands out as an invaluable resource for automating the complex tasks associated with model selection and hyperparameter tuning. This library works on the premise of optimizing the machine learning pipeline, exploring a variety of algorithms and configurations to identify the most effective ones. This automated approach not only speeds up the model development process, but also ensures a thorough exploration of potential architectural models. [6]

Over the course of the predictions in the code, we implemented four different prediction models. In two of them the TROT library was used for classification and in the other two models for regression. In classification prediction works, attempts were made to predict the next process of a business process, among a set of processes where each case, that is, each loan request, supposedly has some historical patterns. The choice of a classifier model aligns with the nature of event prediction and our goal. In parallel, the predictions do not stop only at the prediction of one subsequent process, but also at the prediction and classification of the two subsequent processes. These are done in many different experiments, for both models, where we give a different number of historical patterns for the algorithms to learn the different patterns according to the processes where they have been given. TPOT Classifier, through its automated optimization, adapts to the dynamic sequences inherent in the business process, offering a flexible solution for event prediction. On the other hand, to predict time-related metrics, such as the time duration of each process from one to the next and the remaining cycle time, TPOT Regressor was used. In these predictions, we create models where we initially try to predict the time it takes to move from each process within a request to the next. By this we mean the length of time for each procedure separately, but for all procedures in all cases-loan applications, which can certainly be very important for the management of resources in a bank. Predicting the remaining cycle time in a business process offers many practical advantages and is of significant value in optimizing workflow performance. This is how we try to predict the remaining cycle time for each loan application, where we will see in detail below which procedures we follow to make these predictions. [11]

Thus, this regression model, adapted to predict continuous numerical values, proves invaluable in capturing the temporal nuances embedded in the business process timeline in each and similar dataset.

3.4.1 TPOT LIBRARY

There are a lot of components we have to consider before solving a machine learning problem some of which includes data preparation, feature selection, feature engineering, model selection and validation, hyperparameter tuning, etc. In theory, you can find and apply a plethora of techniques for each of these components, but they all might perform differently for different datasets. The challenge is to find the best performing combination of techniques so that you can minimize the error in your predictions. This is the main reason that nowadays people are working to develop Auto-ML algorithms and platforms so that anyone, without any machine learning expertise, can build models without spending much time or effort. Such a platform is available as a python library: *TPOT*, and we use it during the implementation of the specific thesis. You can consider TPOT as your Data Science Assistant.

TPOT, which stands for "Tree-based Pipeline Optimization Tool," is an open-source Python library designed for automated machine learning (AutoML). Developed by Randy Olson, TPOT is particularly known for its use of genetic programming to optimize machine learning pipelines automatically. The goal of TPOT is to automate the process of selecting the best machine learning model and its hyperparameters for a given dataset. Leveraging genetic programming, TPOT intelligently explores diverse machine learning models, optimizing not only for model selection but also fine-tuning hyperparameters to enhance predictive performance. Furthermore, TPOT incorporates feature engineering techniques, such as polynomial features and interactions, during its evolutionary algorithm-driven search for the most effective machine learning pipeline. This comprehensive approach streamlines the often complex process of selecting, configuring, and refining machine learning models, empowering users with an automated tool capable of efficiently producing high-performing pipelines tailored to specific datasets. [11]

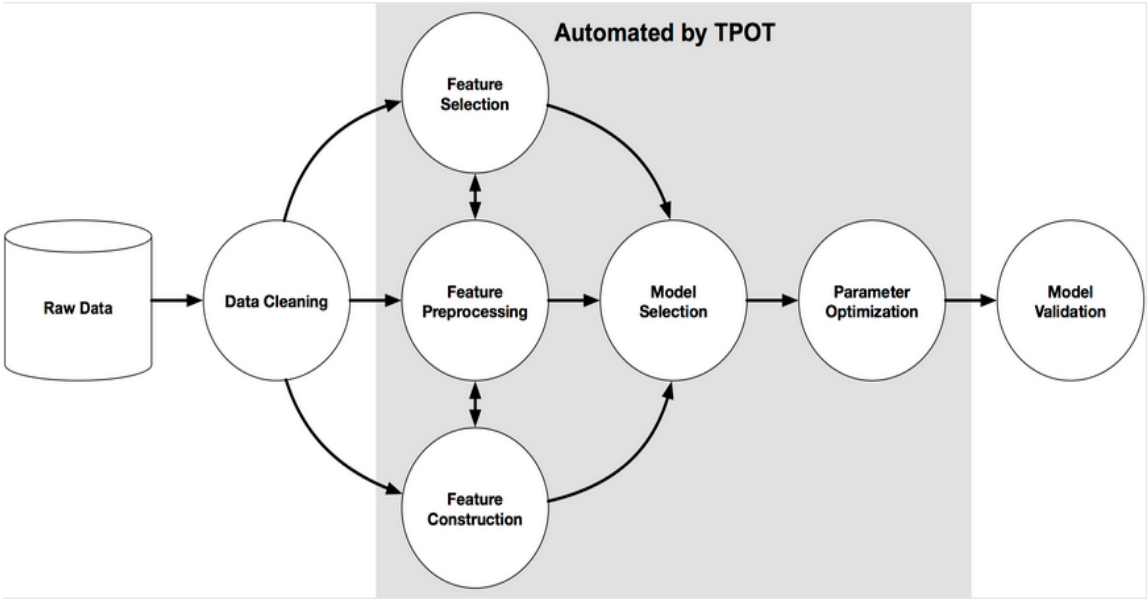

Figure 4 - Automated Model by TPOT Library

With the right data, computing power and machine learning model you can discover a solution to any problem, but knowing which model to use can be challenging for you as there are so many of them like Decision Trees, SVM, KNN, etc. That's where genetic programming can be of great use and provide help. Genetic algorithms are inspired by the Darwinian process of Natural Selection, and they are used to generate solutions to optimization and search problems in computer science. Broadly speaking, Genetic Algorithms have these properties:

- *Initialization*: TPOT starts with a population of randomly generated pipelines, each representing a combination of machine learning algorithms and their hyperparameters.

- *Evaluation*: The fitness of each pipeline is assessed by its performance on the specified task (regression or classification). The evaluation is typically based on metrics like accuracy or mean squared error.

- *Selection*: High-performing pipelines are selected to form the basis for the next generation. Selection is influenced by the fitness scores achieved during evaluation.

- *Crossover*: Selected pipelines undergo crossover, where elements of two parent pipelines are combined to create new child pipelines. This mimics the process of genetic recombination.

- *Mutation*: Random mutations are introduced to the child pipelines, exploring new possibilities in the search space. This adds diversity to the population and helps discover potentially better solutions.

- Repeat: Steps 2-5 are repeated over multiple generations, allowing TPOT to iteratively refine its search space and converge towards optimal machine learning pipelines.

Over several generations, TPOT refines its search space, adapting to the characteristics of the dataset to identify the most effective machine learning models. By combining the power of genetic algorithms with automated machine learning, TPOT provides a comprehensive and effective solution for optimizing predictive modeling tasks. Its adaptability and automated nature make it a valuable asset in the development of accurate and efficient models for business process monitoring.

Consequently, it turns out that choosing the right machine learning model and all the best hyperparameters for that model is itself an optimization problem for which genetic programming can be used. The Python library TPOT built on top of Scikit-Learn uses genetic programming to optimize your machine learning pipeline. For instance, in machine learning, after preparing your data you need to know what features to input to your model and how you should construct those features. Once you have those features, you input them into your model to train on, and then you tune your hyperparameters to get the optimal results.

Instead of doing this all by yourselves through trial and error, TPOT automates these steps for you with genetic programming and outputs the optimal code for you when it's done.

As we will see in detail below, the library is divided into two main parts. One is classification and the other is regression where we choose each problem accordingly and adapting the algorithms where it will search to find the appropriate one for the corresponding problem and dataset we have. [6] [11]

- **HOW WE USED THE TPOT LIBRARY**

In this section we will briefly look at the intricacies of leveraging the Tree-based Pipeline Optimization Tool (TPOT) library to build automated and optimized machine learning pipelines tailored to the BPI 2012 event log dataset we chose for our experiments, which we will see more detailed in the next section.

Raw Data: In this case, to import the data we have an XES file where we analyzed previously how these files work, and we import them using the pm4py library. We then convert it to a dataframe via the Pandas library so that we can use the data for input into the prediction models later. Noteworthy is the selective use of specific columns—timestamp, case ID, and concept name—streamlining the dataset to the essential components for our predictive tasks.

Data Cleaning: TPOT addresses the complexities of raw data by incorporating powerful data cleansing techniques. Handling missing values becomes a critical aspect, ensuring the integrity of the data set and laying the groundwork for accurate predictive modeling. Although it is done automatically, we handle some of the prices we have manually. That is, we convert the timestamp column to datetime and fill in the values that are missing after this conversion due to problems with the library function to fill in values where it is the first value for each case. That is, we fill them with 0. We can also, where there are too many values missing and we have no sure way to fill them in and that they are correct, delete the entire column if it is not very useful in our predictions, as we do in the 'org:resource' column, which represents the department that implements the every procedure every time and it is not sure how to fill these values.

Feature Engineering: The predictive power of our models is enhanced by extracting temporal features from the event log data. TPOT's feature engineering capabilities delve into temporal dynamics, capturing the sequence of events and temporal patterns that are critical to the predictions we will be experimenting with. So it is very important to extract characteristics from the timestamp column, in days, weeks, hours, etc. We helped in this way to have more features for our model to learn different patterns between them and to see different distributions

Model Selection: An automated approach to model selection is orchestrated by TPOT. Through genetic programming, the library navigates a diverse landscape of machine learning algorithms and configurations, selecting models that exhibit optimal performance for our specific predictive tasks. Thus there are two models which we will see in detail in the next subsection. The two models are divided into classification and regression where we use them respectively for the predictions of the next event and time duration or remaining time cycle.

Parameter Optimization: Genetic programming extends its ability to optimize hyperparameters. TPOT evolves and refines the hyperparameter configurations for the selected machine learning models, ensuring that the models are finely tuned to the intricacies of the event log data. In our models we mainly follow parameters with 8 generations and 25 populations. In parallel, we set the verbosity to 2 so that we can see the process in detail and not just the final results.

Model Validation: The robustness of our predictive models is rigorously assessed through cross-validation techniques. TPOT partitions the dataset, trains models on subsets, and validates on remaining data, ensuring reliable performance evaluation and guarding against overfitting.

In summary, the use of TPOT in this integrated machine learning pipeline reflects a commitment to accuracy, efficiency, and adaptability in the field of predictive analytics, reinforced by the intentional inclusion of timestamp, case identifier, and concept name columns to focus on key information. The following sections provide a detailed exploration of each of the two models in the TPOT library as well as the evaluation metrics for the performance of the models where they output, as we will see in more detail in the next section, either predicting the next event, or the two next events, or duration from each event to the next, or forecast of remaining cycle time. In other words, according to the predictive model we are considering.

3.4.2 TPOT REGRESSION

The TPOT Regressor performs an intelligent search over machine learning pipelines that can contain supervised regression models, preprocessors, feature selection techniques, and any other estimator or transformer that follows the scikit-learn API. The TPOTRegressor will also search over the hyperparameters of all objects in the pipeline. By default, TPOT Regressor will search over a broad range of supervised regression models, transformers, and their hyperparameters. However, the models, transformers, and parameters that the TPOT Regressor searches over can be fully customized. Users have the flexibility to customize the search space by specifying the allowed algorithms and their hyperparameters. This level of customization enables practitioners to tailor TPOT Regressor to their specific predictive modeling requirements. For instance, if a user prefers to focus solely on tree-based models

or linear regression, they can restrict the search space accordingly. This fine-grained control ensures that the automated selection aligns with the user's preferences and the characteristics of the business process dataset.

The algorithms listed for TPOT Regressor encompass a versatile set of regression models, each offering unique advantages in capturing patterns and making predictions about numerical metrics. Here's a summary of the regression algorithms:

- *Linear Regression*: This algorithm models the relationship between the dependent variable and one or more independent variables by fitting a linear equation to the observed data. It is widely used for its simplicity and interpretability, assuming a linear relationship between the input features and the target variable.

- *Decision Trees*: Decision trees are non-linear models that recursively split the dataset based on features, forming a tree-like structure. They are capable of capturing complex relationships within the data, making them adept at handling intricate patterns in the business process timeline.

- *Random Forest*: An ensemble method, Random Forest combines multiple decision trees to improve predictive accuracy and control overfitting. It works by constructing a multitude of trees and averaging their predictions, providing robustness and mitigating the impact of outliers.

- *Gradient Boosting*: Gradient Boosting builds a series of weak learners (typically decision trees) sequentially, with each subsequent tree correcting the errors of the previous ones. It excels at capturing subtle patterns and dependencies within the data, making it well-suited for nuanced time-related predictions.

- *Support Vector Machines (SVM):* SVM is a versatile algorithm that can be applied to both regression and classification tasks. It aims to find the hyperplane that best separates the data into different classes or predicts a numerical outcome. SVM is particularly effective in high-dimensional spaces.
- *k-Nearest Neighbors (k-NN):* k-NN is a simple and intuitive algorithm that predicts the target variable based on the majority vote or average of its k-nearest neighbors in the feature space. It is well-suited for tasks where local patterns are important.

These algorithms collectively offer a rich toolkit for TPOT Regressor, enabling automated selection based on the characteristics of the business process data and the specific requirements of the regression task at hand. Users can further customize the algorithmic search space based on their domain knowledge and preferences, ensuring a tailored approach to model selection. [11]

3.4.3 TPOT CLASSIFICATION

TPOT Classifier focuses on classification tasks, making it suitable for predicting discrete outcomes. The TPOT Classifier performs an intelligent search over machine learning pipelines that can contain supervised classification models, preprocessors, feature selection techniques, and any other estimator or transformer that follows the Scikit-Learn API. The TPOT Classifier will also search over the hyperparameters of all objects in the pipeline. By default, TPOT Classifier will search over a broad range of supervised classification algorithms, transformers, and their parameters. However as in TPOT Regression so here too apparently, the algorithms, transformers, and hyperparameters that the TPOT Classifier searches over can be fully customized using the 'config_dict' parameter.

The algorithms listed for TPOT Classifier provide a diverse set of tools for automated model selection in classification tasks. Here's a concise overview of each algorithm:

- Logistic Regression: Logistic Regression models the probability of a binary outcome, making it suitable for classification tasks. It works by applying a logistic function to a linear combination of input features, providing probabilities that can be transformed into class predictions.

- *Decision Trees*: Decision trees in the context of classification create branches based on features to classify instances into different classes. They are effective at capturing complex decision boundaries and interactions within the feature space.

- *Random Forest*: Similar to TPOT Regressor, Random Forest in TPOT Classifier combines multiple decision trees to form a robust ensemble, improving classification accuracy and generalization.

- *Gradient Boosting*: Gradient Boosting for classification builds a series of weak learners to iteratively correct errors. It excels at capturing subtle dependencies within the data and is particularly effective when dealing with imbalanced classes.
- *Support Vector Machines (SVM):* SVM in classification aims to find the hyperplane that best separates different classes in the feature space. It is powerful in scenarios where the decision boundary is non-linear or complex.

- *k-Nearest Neighbors (k-NN):* k-NN for classification predicts the class based on the majority vote of its k-nearest neighbors. It is a simple yet effective algorithm, particularly useful when local patterns are essential in the classification task.

These algorithms collectively form a comprehensive set within TPOT Classifier, facilitating automated selection based on the characteristics of the business process data and the specific requirements of the classification task at hand. Users can further customize the algorithmic search space based on their domain knowledge and preferences, ensuring a tailored approach to model selection in classification scenarios. [11]

3.5 EVALUATION METRICS

Evaluation metrics are crucial in assessing the performance of machine learning models. They provide quantitative measures that guide the selection of models and the tuning of hyperparameters. Different tasks require different metrics, and understanding which metric to use is the key to interpreting model results effectively. These metrics provide insights into how well a model is achieving its objectives, whether it's classification, regression, clustering, or another type of task. When evaluating a machine learning model, it is crucial to assess its predictive ability, generalization capability, and overall quality. The choice of evaluation metrics depends on the nature of the task and the specific goals of the analysis.

As we knew before about predictive models, the performance metrics are separate for classification problems and regression problems. In classification tasks, where the output is a discrete label, common evaluation metrics include: Accuracy, Precision and Recall, F1 Score, ROC Curve and AUC. And for regression tasks where the model predicts continuous values, common metrics include: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Squared Error (MSE). Also for multi-class classification problems, evaluation metrics are extended or adapted from binary classification we have: Confusion Matrix and Classification Report.

As we will see in the next chapter more specifically about the results of the specific dataset and the models we used, we calculate all the above model performance metrics except for the ROC Curve and AUC. So now we will see what each measurement calculates and why it is useful for the performance of the model. [12]

3.5.1 CLASSIFICATION METRICS

- **Accuracy**

Accuracy is the simplest evaluation metric for classification. It is the ratio of correctly predicted observations to the total observations and provides a quick measure of how often the model is correct. It is a straightforward metric that provides a high-level assessment of the model's performance. One of the main reasons why model accuracy is an important metric, is that it is an extremely simple indicator of model performance. However, it may not be suitable for imbalanced datasets.

- **Classification Report**

A classification report is used to measure the quality of predictions from a classification algorithm. The report presents the main classification metrics—precision, recall, and F1-score—for each class. The classification report is crucial for understanding the model's performance across different classes, especially in scenarios with imbalanced datasets.

- **Confusion Matrix**

A confusion matrix is a table that presents a summary of the model's predictions. It includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It provides a detailed breakdown of the types and quantities of classification errors. This information is crucial for diagnosing model performance, understanding where the model excels, and identifying areas for improvement. The printing of a confusion matrix is of the type:

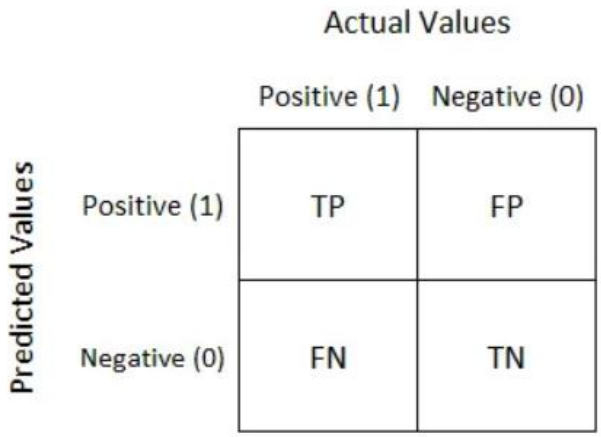

Figure 5 - Confusion Matrix

TN / True Negative: The model correctly predicts a negative class for a negative case.

TP / True Positive: The model correctly predicts a positive class for a positive case.

FN / False Negative: The model incorrectly predicts a negative class for a positive case.

FP / False Positive: The model incorrectly predicts a positive class for a negative case.

- **Precision**

Precision can be seen as a measure of a classifier's exactness. For each class, it is defined as the ratio of true positives to the sum of true and false positives. Said another way, "for all instances classified positive, what percent was correct?". The Precision formula is the following:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall**

Recall is a measure of the classifier's completeness; the ability of a classifier to correctly find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives. Said in another way, "for all instances that were actually positive, what percent was classified correctly?". The Recall formula is the following:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F1 Score**

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. The F1 Score formula is the following:

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

3.5.2 REGRESSION METRICS

- **Mean Squared Error (MSE)**

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

The formula for MSE is the following.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

- **Root Mean Squared Error (RMSE)**

The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points. RMSE quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values.

The RSME formula for a sample is the following:

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

- **Mean Absolute Error (MAE)**

Mean Absolute Error (MAE) is a measure of the average size of the mistakes in a collection of predictions, without taking their direction into account. It is measured as the average absolute difference between the predicted values and the actual values and is used to assess the effectiveness of a regression model.

The **Mean Absolute Error**(MAE) is the average of all absolute errors. The formula is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

4. RESULTS AND ANALYSIS

4.1 LIBRARIES AND FRAMEWORKS

In the pursuit of unraveling the intricate dynamics of Predictive Business Process Monitoring (PBPM) with Automated Machine Learning (AutoML), the choice and implementation of libraries and frameworks play a pivotal role. This section provides a succinct overview of the key libraries employed in the analysis, shedding light on their individual contributions. Additionally, we touch upon the concept of frameworks, outlining their role in guiding the structure and methodology of the research. The application of these tools is further contextualized within the dataset used for the study, focusing on the Bank Loan Application event logs from the Business Process Intelligence (BPI) Challenge 2012 dataset.

Libraries:

1. **Pandas and NumPy:** Pandas and NumPy were instrumental for data manipulation and numerical operations, providing a robust foundation for handling datasets and performing essential computations.

2. **PM4Py:** PM4Py, a process mining library, played a pivotal role in extracting insights from event logs, enabling the application of process discovery and conformance checking techniques.

3. **Seaborn and Matplotlib:** Seaborn and Matplotlib served as powerful data visualization tools, facilitating the creation of insightful graphs and plots for a comprehensive analysis of predictive model outputs.

4. **Plotly:** Plotly enriched the visual representation of data with interactive plots, enhancing the communicative aspects of the results and providing an immersive experience in exploring patterns and trends.

5. **Scikit-learn (sklearn):** Scikit-learn, a versatile machine learning library, offered a broad spectrum of tools for model selection, evaluation, and preprocessing, streamlining the implementation of predictive models.

6. **TPOT:** TPOT, an automated machine learning framework, played a crucial role in optimizing model selection and hyperparameter tuning, automating the tedious aspects of the machine learning pipeline.

The research methodology adhered to a systematic framework designed to unlock the predictive potential of business process monitoring within the context of bank loan applications. Commencing with a meticulous exploration during the Business Understanding phase, the emphasis was placed on defining and comprehending the intricacies of predicting next event, next two events, time duration, and remaining cycle time in the loan application process. The subsequent Data Preparation phase involved a judicious selection of relevant

features, managing missing values, and optimizing the dataset structure by retaining essential columns such as ‘case:concept:id’, ‘time:timestamp’, and ‘concept:name’. But also the 'time_duration' column where it is used accordingly in the remaining cycle time problem.

Proceeding to the Modeling phase, the automated machine learning TPOT assumed a central role. Tailored functions were employed to prepare the dataset for each specific prediction problem, ensuring compatibility with TPOT's automated model selection and hyperparameter tuning capabilities. This phase was characterized by a meticulous integration of TPOT's automated pipeline optimization, facilitating efficient and effective model development for each predictive task.

The Evaluation phase witnessed the deployment of comprehensive metrics, including accuracy, classification report, confusion matrix, precision, recall, MSE, RMSE, and MAE, to assess the performance of each TPOT-optimized model. This iterative process allowed for a nuanced understanding of the efficacy of predictive models in addressing the diverse challenges presented by next event prediction, multiple next events, time duration prediction, and remaining cycle time prediction. The structured deployment of this framework ensured a systematic and insightful exploration of predictive analytics within the dynamic domain of business process monitoring for bank loan applications.

4.2 DATASET

The Business Process Intelligence (BPI) Challenge is an annual competition that invites participants from the data science and business process management communities to address real-world challenges in process mining and analytics. The BPI Challenge serves as a benchmark for evaluating and advancing techniques in the field of business process management.

For this research endeavor, the chosen dataset emanates from the BPI Challenge 2012, a notable installment in the series. The dataset centers around the domain of bank loan applications, presenting a rich and multifaceted landscape for exploration. It encapsulates a myriad of events, timestamps, and process-related information that captures the intricacies of the loan application process. The dataset, originating from real-world scenarios, provides a representative and practical foundation for studying predictive business process monitoring within the context of financial processes.

Researchers and practitioners interested in accessing the BPI 2012 dataset for their own analyses can find it on the official BPI Challenge website or related repositories. The dataset's accessibility and relevance make it a valuable resource for studies seeking to unravel the complexities of business processes and harness predictive analytics to inform decision-making. The decision to employ the BPI 2012 dataset underscores the research's commitment to practical applicability and real-world relevance within the domain of business process monitoring.

- **Exploratory Analysis:**

The BPI Challenge 2012 dataset, initially stored as an XES file, underwent a transformation into a structured dataframe for in-depth analysis. The columns within the dataset include 'org:resource', 'lifecycle:transition', 'concept:name', 'time:timestamp', 'case:REG_DATE', 'case:concept:name', 'case:AMOUNT_REQ'. However for the predictive models, specific columns were identified, such as 'concept:name', 'time:timestamp', 'case:consept:id' and 'event_duration,' aligning with the requirements of the business process monitoring objectives. In the following figure we will see an example of the first 10 lines of our dataset, after we have converted it into data frame, so we can see how it is:

	org:resource	lifecycle:transition	concept:name	time:timestamp	case:REG_DATE	case:concept:name	case:AMOUNT_REQ
0	112.0	COMPLETE	A_SUBMITTED	2011-09-30 22:38:44.546000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
1	112.0	COMPLETE	A_PARTLYSUBMITTED	2011-09-30 22:38:44.880000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
2	112.0	COMPLETE	A_PREACCEPTED	2011-09-30 22:39:37.906000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
3	112.0	SCHEDULE	W_Completeren aanvraag	2011-09-30 22:39:38.875000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
4	NaN	START	W_Completeren aanvraag	2011-10-01 09:36:46.437000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
5	10862.0	COMPLETE	A_ACCEPTED	2011-10-01 09:42:43.308000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
6	10862.0	COMPLETE	O_SELECTED	2011-10-01 09:45:09.243000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
7	10862.0	COMPLETE	A_FINALIZED	2011-10-01 09:45:09.243000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
8	10862.0	COMPLETE	O_CREATED	2011-10-01 09:45:11.197000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000
9	10862.0	COMPLETE	O_SENT	2011-10-01 09:45:11.380000+00:00	2011-09-30 22:38:44.546000+00:00	173688	20000

Figure 6 - 10 First Lines of the Dataset (BPI2012)

Regarding the understanding of what each column contains and what it represents in a business, we have that: the 'org_resource' column represents an employee or department that is responsible for an activity of a loan request. That is, it helps with who performed a procedure. The 'lifecycle:transition' column indicates a more basic event in the set of processes. That is, the initiation, completion or planning for the loan application. So we proceeded to the 'concept:name' column where we see all the procedures for the loan application, in total. It describes what happens at a given point in the process and is a very important column for its analysis and manipulation for the prediction of models in general in this type of data. Then the 'time:timestamp' column is the timestamp for a process. The given moment when the process started to be implemented and can be the chronological sequence of events. In parallel, the 'case:REG_TIME' column contains information about when the request was initially entered into the system. That is, it only indicates the initial timestamp for each request. Then we have the 'case:concept:name' column which represents the id of each loan case. That is, a set of procedures constitute a request where this request for facilitation consists of an id number found in the specific column. Finally, we have the 'amount_req' column which shows us the requested amount for each loan request

requested by the client. So we see that the column of procedures that make up the set of procedures for each request, the number that represents an entire request and the time stamp of the procedures within a request, are the columns that give us valuable information and help us to proceed to analysis for the understanding of the specific data but also for its handling in terms of predictions where they will be made later.

In the exploratory analysis, essential metrics were computed to characterize the dataset's nature:

- Number of events: 262,200
- Number of cases: 13,087
- Average Events per case: 20.04
- Average Case Length: 20.04
- Average Event Duration (hours): 10.87
- Max Event Duration (hours): 2,468.41
- Average Case Duration (hours): 206.97
- Max Case Duration (hours): 3,293.32
- Number of Variants: 1,348

Language normalization was applied to the 'concept:name' column, converting Dutch values to English for improved clarity. The temporal aspect of the dataset revealed that activities spanned from 01/10/2011 to 01/03/2012, with notable exclusions of certain months— encompassing only January, February, March, September, and October.

Then we analyzed in all cases the procedures that start and end. All loan application cases start uniformly with the 'A_SUBMITTED' process. On the other hand, they are completed with the final activities being 'W_Validate request', 'W_Edit contract details', 'A_Declined', 'W_Complete your application', 'A_Cancelled', 'W_Call incomplete files', 'W_Handling leads', ' W_Call for quotes', 'W_Assess fraud', 'O_Cancelled', 'A_Approved'. However, with greater frequency the procedures 'A_Declined', 'W_Validate request' and 'W_Handling leads'.

The temporal aspect reveals occurrences of activity during specific months, with gaps in the data set for certain periods. This analysis provides a fine-grained understanding of the temporal distribution, activities, and characteristics of the data set, forming a comprehensive foundation for subsequent predictive modeling. The depth of exploration is encapsulated in derived metrics and insights, emphasizing the heterogeneous nature of the BPI Challenge 2012 dataset. In addition, a heuristic network visually captures complex relationships within the dataset, providing a graphical representation of the interaction between different process elements where we will in figure 6.

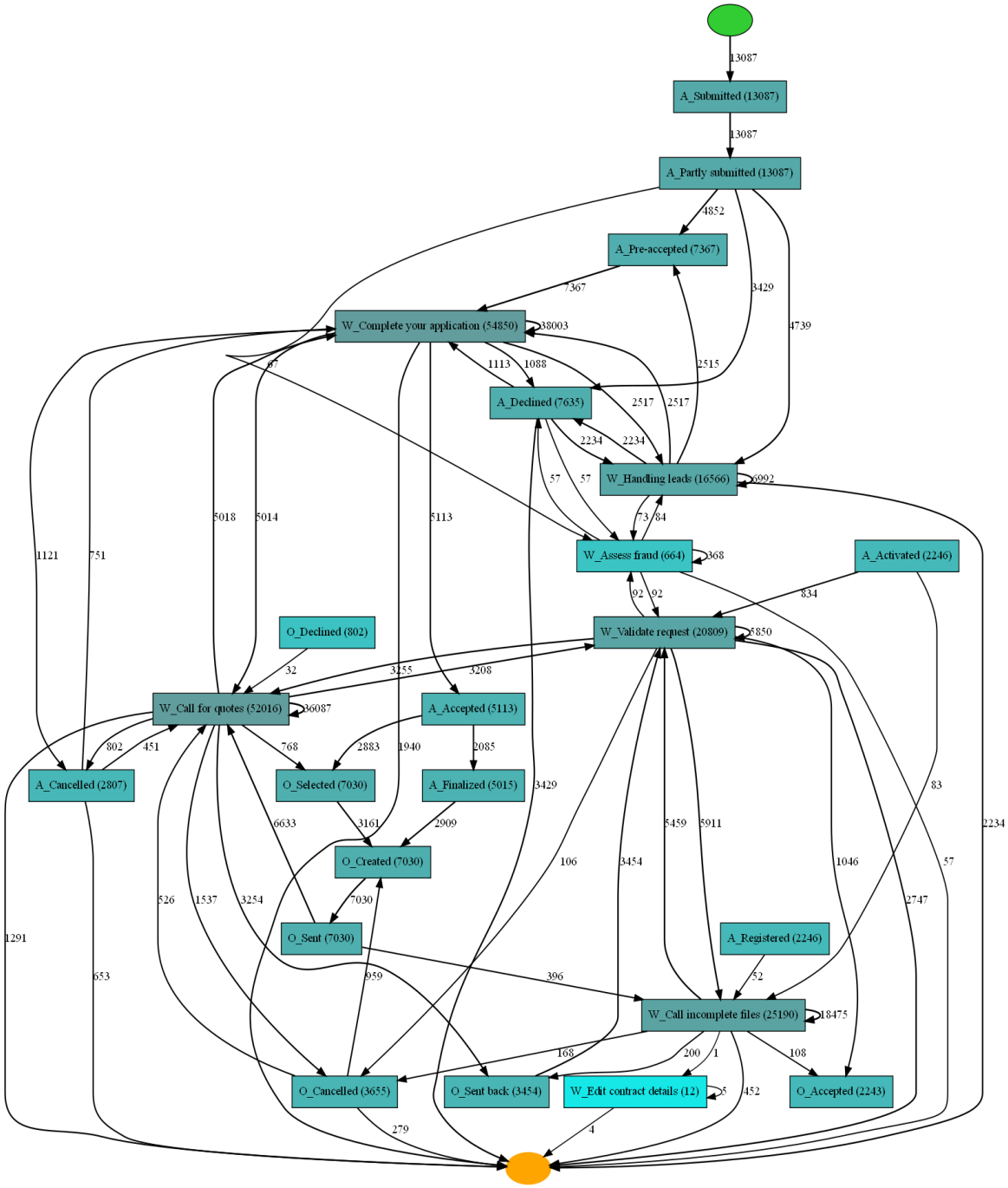

Figure 7 - Heuristic Net for BPI2012

4.3 PRESENTATION OF PREDICTIVE MODELS

4.3.1 NEXT EVENT PREDICTION (ANALYSIS AND RESULTS)

- **INTRODUCTION**

Next event prediction is a pivotal aspect of business process management that involves forecasting the subsequent step in a series of events within a given process. In the realm of business operations, understanding and anticipating the next stage in a workflow or procedure can offer substantial advantages. This predictive capability enables organizations to enhance decision-making, optimize resource allocation, and streamline workflows, ultimately contributing to increased operational efficiency. The significance of next event prediction lies in its potential to preemptively address bottlenecks, identify potential deviations, and streamline the overall flow of processes. By harnessing advanced analytical models, businesses can gain foresight into the sequence of activities, allowing for proactive interventions and strategic adjustments. This not only facilitates smoother operations but also aids in achieving overarching business objectives, such as minimizing delays, enhancing customer satisfaction, and improving overall process performance.

In the next analysis, we delve into the application of a next-event prediction model to the BPI Challenge 2012 dataset we saw earlier. The code and methodology used highlight the practical applications of predictive analytics in optimizing business processes and paving the way for informed decision-making, but also the way for non-experts to engage with automated AI, as we will see in result automatically the algorithm that is the best together with the set of hyperparameters.

- **METHODOLOGY**

In this theoretical exploration, we delve into the implementation of a next-event prediction model using a window approach. The prediction model is designed to predict the next step based on historical sequences of events, allowing adaptability to different window sizes. The choice of a window size is very important, affecting the sensitivity of the sequence and shaping the ability of the model to capture patterns in the data. Thus, it enables us in a very simple way to do different experiments by only changing the value of one variable to see the distribution of the model's performance. The experiments to be done include window values: 1, 2, 3, 5, 10, 15 and 20.

To make it easier to understand the window size is a number that determines the number of events we will give the system to predict the next ones. That is, by defining the size of the window, this will be used by a function, which, through an iterative loop, will go through all the cases of loans, of the dataset we have, and will create sequences of events based on the size of the window we defined. This is how we output a sequence of events based on this size. This is very important because the number of events (window size) that will be generated through the function, the model will learn patterns within a window of events of fixed size to predict the exact next event each time.

To make it easier to understand the size of the window is a number that determines the number of events that we will give to the system so as to predict the next. If for example the window has a value of 5, this means that the model will try to predict the 6th event given the first 5 events and learning from them.

The implementation begins by preprocessing the dataset, and importing the event log data using popular Python libraries such as pandas to convert the dataset to dataframe, numpy to properly import the data into the model, and pm4py to import the xes file. Timestamps are converted to datetime objects. This is a very important conversion for the given management of date data where during the conversion the values that are 'first' in each case are empty so we set them to 0. We do this because the first value in each case, for the moment in time it's 0 because that's where each assumption starts, so since it's empty we set it to 0. Missing values are filled in using the fill-ahead methodology, and time zone information is removed to ensure consistency. A very important step is the coding of the "concept:name" values, a mapping dictionary is created and applied, facilitating the conversion of categorical data into numerical form. This is a very important step because if they are not converted into numerical values, the experiments will not be able to be done by changing the size of the window because we are not allowed to enter categorical values in a machine learning model.

Next, the create_sequences function is at the core of the implementation, and is responsible for creating sequences and tags based on the specified window size. This function works within the constraints of the dataset and business process context, striking a balance between the historical contexts considered for prediction and the practicality of the model. As a result, we see that the specific function does the important work so that we only have to change the value of the window, so that the value of the window is passed through the function and the appropriate traces are created. The window size parameter is very important and it facilitates us, it introduces flexibility, allowing the model to adapt to different contexts and record patterns of different lengths of the tails. So we can easily resize the window to see how the model behaves when the number of processes we give increases. As mentioned before the window size for this analysis spans values of 1, 2, 3, 5, 10, 15 and 20. For each window size, sequences and labels are generated and the data set is split into training and test sets to use appropriately for our model. [13]

The machine learning pipeline uses the TPOT library, an automated machine learning tool, to determine the most appropriate classification model and hyperparameters for the next event prediction task. The model is trained on the generated sequences and tags and then predictions are made on the test set. Evaluation metrics, including precision, accuracy, recall, and confusion matrix, offer insights into model performance in different categories. In all experiments performed, a standardized set of parameters for TPOTClassifier was used, ensuring consistency and comparability between different analyses. Obviously we use TPOTClassifier because we have a classification problem and the parameters used for TPOTClassifier were:

- **Generations:** 8
- **Population Size:** 25

- **Random State:** 42
- **Verbosity:** 2

These parameter choices were applied uniformly to each experiment, maintaining a constant framework for the automated machine learning tool. The consistent use of these parameters facilitates a fair and unbiased comparison of results across various analyses, allowing for a comprehensive evaluation of the model's performance under similar conditions.

This holistic approach to next event prediction integrates theoretical underpinnings with practical implementation, emphasizing the importance of considering historical sequences in predicting future events within a business process. The code encapsulates a thoughtful solution to the problem, aligning with the objectives of business process management and predictive analytics within the given context.

➢ PERFORMANCE EVALUATION

The culmination of our experiments is visualized in the performance distribution plot of the model (Figure X.X), a central element of our analysis. This graphical representation clarifies the predictive accuracy achieved over various time windows, offering an immediate visual understanding of the model's performance nuances. As the window size increases, a distinct upward trend in performance becomes apparent, validating the intrinsic link between time frame and predictive accuracy with an apparent differential increase in model performance from window size 3 and above.

The distribution of model efficiency across different window sizes is visually depicted in the following chart:

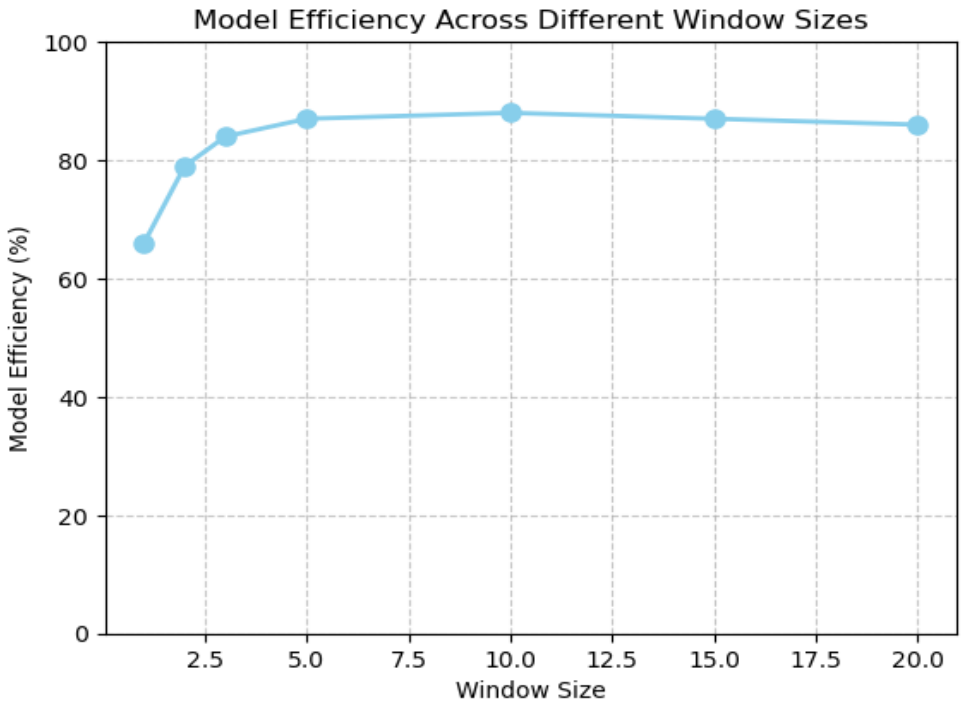

Figure 8 - Model Efficiency for Next Event Prediction

By delving into a detailed analysis of each window size, subtle patterns and algorithmic preferences come to light. In the case of window size 1, the TPOT library automatically selects the ExtraTreesClassifier algorithm proves to be effective, yielding a commendable model performance of 66%. The subsequent increase in window size to 2 introduces a change in the performance of our model, but with TPOT again choosing the ExtraTreesClassifier algorithm with different hyperparameters, but resulting in an increased performance to 79% from 66%. Thus we see that already from the second experiment the accuracy rate increases by 13%, a significant increase where in the process it will increase even more seeing how important it is for the model to get as much information-data as possible in order to learn more and more patterns between the data . A notable transition also occurs at window size 3, where TPOT autonomously selects the RandomForestClassifier thus again increasing the performance of the model. This change in algorithmic preference, accompanied by specific hyperparameter configurations, contributes to an 84% performance. We see that the percentage is constantly increasing. It is important to note that switching to the RandomForestClassifier algorithm highlights the model's ability to dynamically adapt its approach to different time frames, exploiting the strengths of different algorithms and combining the appropriate hyperparameters. Comparative information on window sizes further illuminates the complex relationship between time frame, algorithmic selection, and predictive performance. Continuing to increase the window size of the variable, to number 5, we see that TPOT again chooses the ExtraTreesClassifier algorithm as in the first and second experiments. This leads us to the continuous increase of the model's performance while we see that using the same algorithm, with different hyperparameters but giving more data to the model to learn from, we have a better accuracy rate.

Thus we reach the point where the model achieves a maximum performance of 88% for a window size of 10, using the RandomForestClassifier algorithm with specific hyperparameters. This documents the critical importance of choosing an optimal time window, emphasizing the adaptability of the model to varying business process dynamics and the potential of changing the window to change the model's efficiency. Moving on, moving to larger window sizes, namely 15 and 20, introduces additional nuances. The selection of ExtraTreesClassifier's TPOT with RobustScaler and MinMaxScaler, respectively, shows the model's differentiated approach to handling environmental variation. Despite the small drop in efficiency, these findings highlight the adaptability and flexibility of the model in optimizing predictive accuracy.

Our results not only confirm the correlation between window size and predictive accuracy, but also highlight the dynamic nature of the subsequent event prediction model. The ability to autonomously select different algorithms based on time windows adds a layer of sophistication, offering professionals valuable insights for fitting models to various business process scenarios.

From a business perspective, the observed results suggest that, when dealing with predictive modeling for sequences of events, the choice of window size significantly affects both computational resources and model accuracy. In this particular context, the comparison between window sizes 5 and 20 shows that the larger window size does not

provide a substantial improvement in prediction accuracy, while imposing a noticeable increase in computational cost.

Moving now to an analysis regarding the computational cost in relation to the window size, we can see that in the diagram we have after the window size equal to 5, a constant course of the accuracy up to and decreasing slightly in some window sizes. This happens at size 20 where we have 1% less accuracy rate at a large window difference of 15 units. That is, we want to see the computing cost for a window equal to 5 and for a window equal to 20, so that since they have the same and smaller percentage of accuracy, if there is a big difference in computing cost due to the larger size, so that we have the opportunity to know that there is no need to do such experiments because nothing changes and in the end it costs us much more without getting better results. Consequently, in this particular context, the comparison between window sizes 5 and 20 shows that the larger window size does not provide a substantial improvement in the prediction accuracy, while it imposes a noticeable increase in the computational cost. For window_size = 5, the experiment was completed in 3 hours and 51 minutes, demonstrating the effectiveness of the model in capturing patterns in a smaller sequence of events. The achieved accuracy of 87% in this window size indicates a high predictive ability, making it an exciting choice for practical applications. On the other hand, the experiment for window_size = 20 took significantly longer, i.e. a total of 5 hours and 46 minutes.

While the accuracy achieved was marginally lower at 86%, the increase in computational cost raises questions about the necessity of using larger window sizes. Given the marginal accuracy difference between the two window sizes and the substantial increase in computation time for window_size = 20, it becomes apparent that the smaller window_size of 5 is a more efficient choice. This observation suggests that, for this particular data set, the computational cost outweighs the minimal gain in accuracy when larger window sizes are chosen. Therefore, in practical scenarios where computational resources are a concern, choosing a window_size of 5 could provide a realistic balance between accuracy and efficiency. Therefore, businesses aiming to apply predictive models to sequences of events should carefully consider the trade-off between accuracy and computational cost, leaning toward smaller window sizes for more efficient and practical solutions. Essentially, this observation highlights the importance of optimizing computing resources according to the specific requirements and constraints of the business. By making informed decisions about window size based on the task at hand, businesses can strike a balance that aligns with their business needs and resource capabilities.

In conclusion, the application of the next step prediction model proves to be a valuable tool for business process prediction. The analysis highlights the importance of choosing an optimal window size, balancing the need for historical context with computational efficiency. This information can inform decision makers to optimize their processes, allocate resources efficiently, and ultimately improve overall business performance.

Continuing our research on predicting the next event in a business process, the literature and various scientific articles on business process models and experiments using artificial intelligence make predictions about the next event within a set of processes. This may not

be limited to just one next event, but in this thesis we take it a step further and try to make predictions to predict two next events in a business process, in a similar way as we did before to see if we achieve similar rates and we can do something like this and it will give us even more information for better decisions.

Below we will see this experiment, an introduction, then the methodology we followed for this experiment and finally the evaluation of the results. However, as we will see, we achieve much lower rates and this is because the lower accuracy in predicting the next two events compared to the next event can be attributed to the increased complexity of predicting multiple next steps. Predicting a single event relies on identifying patterns and dependencies within the data set, which can be simpler. However, predicting two consecutive events introduces additional layers of complexity, making it more difficult for the model to accurately capture the sequence of events and their timings. The higher level of uncertainty in predicting next events probably contributes to the lower accuracy seen in the next two event predictions, and so we can't get carried away and use these results to our advantage.

- ✓ **NEXT TWO EVENTS PREDICTION**

In the field of predicting business processes, as we saw in the previous experiment, the task of predicting the next events has increased importance, offering a more nuanced understanding of sequential processes. Unlike one-step forecasting, which focuses on predicting the next event, the two-event forecasting model delves into the temporal dynamics of business processes. By extending the prediction horizon, organizations can gain even more insight into potential forks or deviations in their workflows, allowing more informed and meaningful decisions to be made.

- ➢ **METHODOLOGY**

The application of the two-event prediction model aims to predict the next two situations in a loan application process, leveraging historical data embedded in the "concept:name" column. The methodology is similar to predicting an upcoming event. The dataset, loaded from BPI2012 in .xes file, undergoes basic preprocessing steps. Timestamps are converted to datetime objects and missing values are handled via forward padding. The "time:timestamp" column is then normalized to remove the time zone information. Additionally, a mapping dictionary is created to encode the 'concept:name' values numerically, creating a basis for subsequent predictive modeling. All these steps are the same as our previous model.

But the heart of the two-event forecasting methodology lies in creating sequences that incorporate the time evolution of loan application processes. The ‘create_sequences’ function meticulously constructs sequences and labels, where a sequence represents a window of previous states and the label records the next two events in the loan request trajectory. This design choice allows the model to discern patterns and dependencies that extend beyond a single step, offering a more nuanced understanding of the dynamic nature of business processes. It is important to note that experimentation involves varying the number of past events considered for prediction. The window size parameter, declared as

'window_size' in the code, is systematically adjusted for experimental purposes. Selected window sizes include 1, 2, 3, 5, 10, 12, 15, and 20, each representing a separate survey of the model's predictive capabilities. This scope of experimentation ensures a comprehensive analysis of model performance at different levels of historical context, shedding light on the optimal window size for accurately predicting the next two events in the loan request sequence.

The sequences generated from the selected window sizes are then used to train and test the predictive model. The dataset is split into training and test sets and TPOTClassifier is used to automatically discover the best line with optimal hyperparameters. The training process involves adapting the model to the training set, allowing it to learn complex patterns in loan application processes. After training, model performance is rigorously evaluated using the test set and accuracy metrics are calculated for each experiment. This thorough analysis serves as a solid foundation for informed decision-making in the dynamic landscape of loan application processes.

Finally, as in the previous model for predicting the next event, now the parameters remain the same for the generations and the population of the TPOT classifier.

- **PERFORMANCE EVALUATION**

In investigating the prediction of the next two events in a business process, we scrutinized the performance of the model over various window sizes. The following paragraphs provide a detailed analysis of the results, accompanied by information from the graph depicting the model's performance distribution. Overall, the performance of the model based on the experiments of different window sizes is in Figure 8.

Starting with a window size of 1 to predict the 2nd and 3rd events, the model achieved a performance of 23%, using a Logistic Regression classifier according to TPOT. As we increased the window size to 2, the model performance improved slightly to 24%, using feature scaling with MaxAbsScaler and MLPClassifier. Then increasing the window size to 3, the model showed stable performance with 24% efficiency. The optimal pipeline for this scenario involved Recursive Feature Elimination (RFE) followed by a decision tree classifier with specific hyperparameters. The stability in performance may indicate that, in this context, extending the prediction horizon to two events does not significantly alter the model's accuracy. Continuing, expanding the window size to 5 resulted in 26% performance, with a Gaussian Naive Bayes basis estimator and an Extra Trees classifier. The graph provides a visual representation of how the model's performance evolves over these different time frames, offering a comprehensive overview of the model's predictive power.

Subsequently for larger window sizes (10, 12, 15 and 20), the model showed a small progressive improvement in performance. In particular, at a window size of 10 we had 31%, i.e. the performance of the model increased even more. While at a window size of 12, the model achieved the highest performance of 33%, using a Bernoulli Naive Bayes classifier. Finally for window sizes 15 and 20, the model returns 32% and 30% respectively, remaining at a high rate of return. The plot, included for visual clarity, shows the distribution of model

performance over various window sizes. This graphical representation helps to understand how the predictive capabilities of the model respond to changes in the time frame, providing valuable information for making informed decisions in business process optimization.

A notable observation emerges from the exploration of predictive modeling, especially in the context of predicting one versus two future events in a business process. While the model demonstrated a remarkable peak performance of 88% in predicting the next event, extending the prediction horizon to two events resulted in a significant drop in performance, yielding a peak performance of 33%. A performance that is no more than 50% less accurate than predicting the next step. This discrepancy highlights the complexity introduced by a more extended forecast horizon, prompting a deeper investigation into the factors affecting the model's ability to capture temporal dependencies. The upcoming section on interpretation of results will delve into a comprehensive analysis, shedding light on the complexities of these different prediction tasks.

Thus we conclude that it is reasonable that it is not yet possible to perform worthwhile experiments that will yield significant rates of accurate prediction. If not as in the prediction of the next step of 88%, but at least a significant percentage where it could be valid and give valuable information and decision-making ability to the executives of a company about the allocation of resources, the computational cost, etc. In this way we see that such a prediction is very difficult to achieve high rates and we see that we cannot use them to help us in the decision of a business process such as predicting the next step. Since the maximum percentage is 33%, we cannot use it because it is too low and we cannot achieve anything with it.Nevertheless, we are given the opportunity to experiment and try to go a step further than predictions in business processes where it would be very important at some point to be able to predict with remarkable accuracy rates, not only the next process but also the two next processes.

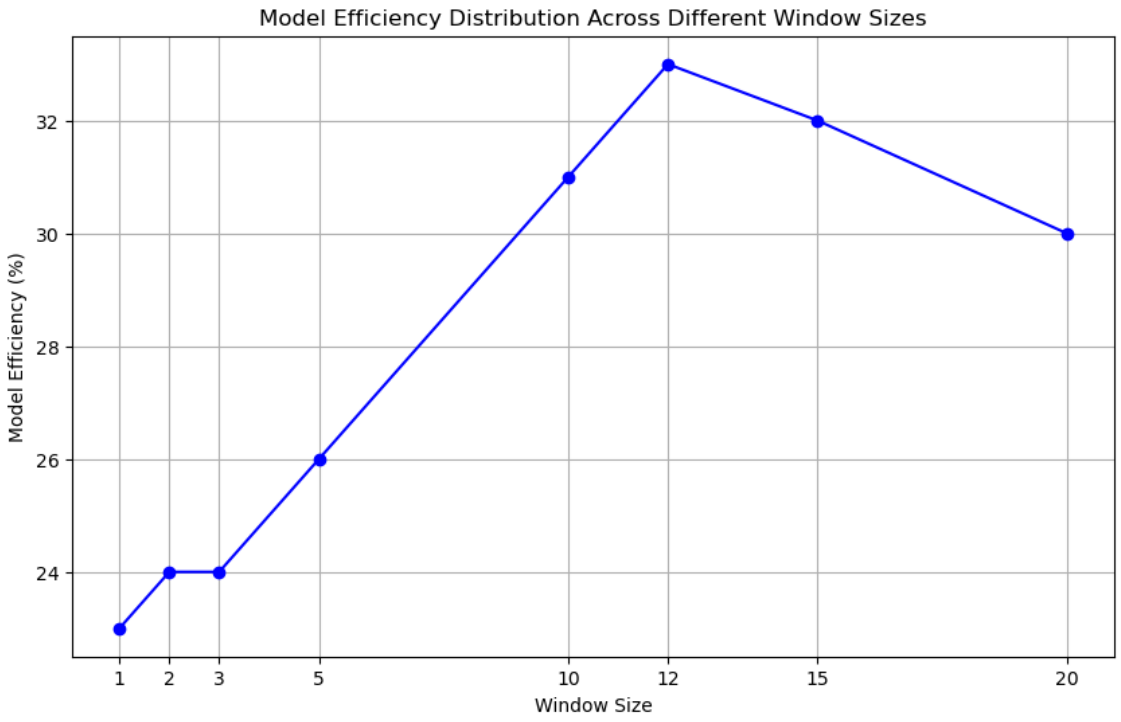

Figure 9 - Model Efficiency for next two events prediction

4.3.2 TIME PREDICTION (ANALYSIS AND RESULTS)

- **INTRODUCTION**

In the ever-evolving landscape of business processes, the ability to predict and understand the temporal aspects between successive events is of utmost importance. Time forecasting in a business process is a strategic endeavor that aims to uncover the complex web of time dependencies, providing organizations with the foresight required for effective resource allocation, operational planning, and process optimization. Accurate predictions of time intervals between events enable businesses to streamline workflows, identify potential delays, and orchestrate tasks with precision. Harnessing the power of predictive analytics, particularly through the application of TPOTRegressor, as in this case there is a regression problem, in event log data, is becoming instrumental in ushering in a new era of data-driven decision making. As we begin our exploration of time forecasting in this subsection, our goal is to unravel the temporal dynamics inherent in business processes, offering insights that drive organizations toward increased efficiency and proactive management.

To make accurate time predictions within the complex tapestry of business processes, we use TPOTRegressor as a powerful tool for automated machine learning. TPOTRegressor simplifies the process of model selection and hyperparameter tuning, allowing us to efficiently navigate the vast landscape of regression algorithms. Leveraging TPOTRegressor's evolutionary search algorithms, we aim to discover optimal models that capture the temporal complexities present in the BPI2012 dataset. The automated nature of TPOTRegressor enhances our ability to handle different time patterns, offering a strong foundation for creating accurate time duration forecasts. Through this methodological approach, we strive to uncover reliable insights that not only enhance the predictive capabilities of business processes, but also contribute to a more informed and flexible decision-making framework.

- **METHODOLOGY**

In our intricate endeavor to forecast the time durations between successive events within the intricate fabric of business processes, we employ a highly nuanced and efficient methodology guided by the TPOTRegressor framework. Our primary objective is to predict the temporal spans from each event to its subsequent occurrence across all events in the entire business process landscape. It's essential to highlight that, owing to the homogeneous nature of most processes and their comparable completion times, the model exhibits swift learning capabilities and adeptly discerns patterns.

The initiation of our comprehensive methodology involves a meticulous extraction of temporal features from the event log data, with a keen focus on the "time:timestamp" attribute. This process encompasses converting raw timestamps into a standardized date format and establishing a mapping dictionary to encode the myriad events, thereby constructing a meticulously structured foundation for subsequent predictive modeling endeavors.

At the heart of our approach lies the sophisticated generation of sequences and labels tailored explicitly for temporal prediction. This intricate process involves calculating the temporal intervals between successive events, thereby encapsulating the nuanced temporal dynamics within each business case. The resultant sequences, encapsulating time durations, act as the training data, while labels precisely denote the corresponding durations — effectively representing the temporal time gaps from each event to its subsequent occurrence for the entire event spectrum. The subsequent transformation of this sequential data into a 2D array is pivotal, aligning with the requisite input specifications mandated by TPOTRegressor, ensuring the seamless functionality of our model.

The TPOTRegressor, a formidable automated machine learning tool, assumes the pivotal role of identifying the most efficacious regression model while optimizing hyperparameters. Configured with an elaborate eight-generation setup and a population size of 25, TPOTRegressor engages in a sophisticated evolutionary search, meticulously discerning the optimal pipeline for the intricate task of duration prediction.

The culmination of our meticulous efforts is the evaluation of the trained TPOTRegressor model on an independent test set to gauge its performance comprehensively. Employing metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 Score, we meticulously assess the model's accuracy and predictive prowess. The depth of our analysis extends beyond mere quantitative metrics, encompassing a qualitative comparison of actual versus predicted values. This multifaceted evaluation provides an exhaustive understanding of the model's effectiveness in capturing the intricate temporal intricacies imprinted within the nuanced landscape of business processes. Our robust and sophisticated methodology is a testament to the remarkable prowess of automated machine learning in unraveling complex temporal dependencies within the intricate tapestry of the BPI2012 dataset.

- **PERFORMANCE EVALUATION**

Our meticulous approach to predicting time durations between successive events in business processes through the utilization of TPOTRegressor has yielded highly promising outcomes. The configuration of TPOTRegressor, spanning eight generations, consistently demonstrated outstanding performance, as evidenced by a best internal cross-validation score of approximately -1.60e-05 across all generations. This exceptional level of accuracy underscores the model's robust ability to capture intricate temporal nuances within the BPI2012 dataset.

The ultimate composition of the best pipeline involves a MaxAbsScaler preprocessing step followed by a LassoLarsCV regression model. The evaluation of this pipeline on the test set reveals remarkable performance metrics. The Mean Squared Error (MSE) of 0.00 signifies an almost negligible deviation between the predicted and actual time durations. Further validating the model's accuracy, the Root Mean Squared Error (RMSE) of 0.00399 indicates minimal residual errors. The R2 Score of 1.00 attests to the model's unprecedented predictive power, explaining the entirety of variance in the temporal dynamics of the dataset.

The model's exceptional performance, approaching 100% accuracy, can be attributed to the specific characteristics of the time duration variable. In many business processes, the time elapsed between events tends to exhibit a consistent pattern, with minimal variability. This is especially true when considering short-term intervals between consecutive events, where the business process follows a well-defined and stable rhythm. For instance, if we take the scenario where the time duration from event A to event B is, on average, 30 seconds, this pattern might persist across a multitude of cases. The model, being trained on such data, becomes adept at recognizing and generalizing this temporal regularity. As a result, when tasked with predicting the time duration between events, it excels due to the uniformity of these intervals. However, it's crucial to acknowledge that this remarkable accuracy is context-dependent and may not necessarily translate to scenarios with more dynamic or unpredictable temporal patterns.

For a more intuitive understanding of the model's performance, a graph has been included. This visual representation effectively contrasts the predicted durations with the actual durations, demonstrating the accuracy of the model in capturing the temporal dynamics. In this particular plot, where the predicted and actual values overlap almost perfectly, it signifies an excellent level of accuracy achieved by the prediction model. Each point in the diagram represents an example of data where the predicted time duration between events aligns almost exactly with the actual observed duration. The closeness of the points to a single value in the plot highlights the model's ability to accurately capture and reproduce the temporal patterns present in the training data.

The closeness of the predicted and actual values indicates a high degree of consistency and reliability in the model predictions. Essentially, the model has learned the underlying temporal relationships within the data set so effectively that its predictions reflect the actual results with remarkable fidelity. This alignment between predictions and reality is particularly pronounced when the time durations between events exhibit a remarkable level of uniformity. The graph serves as a tangible illustration of the model's ability to reflect complex temporal dependencies within the BPI2012 dataset, enhancing not only its accuracy but also its practical effectiveness in real-world scenarios.

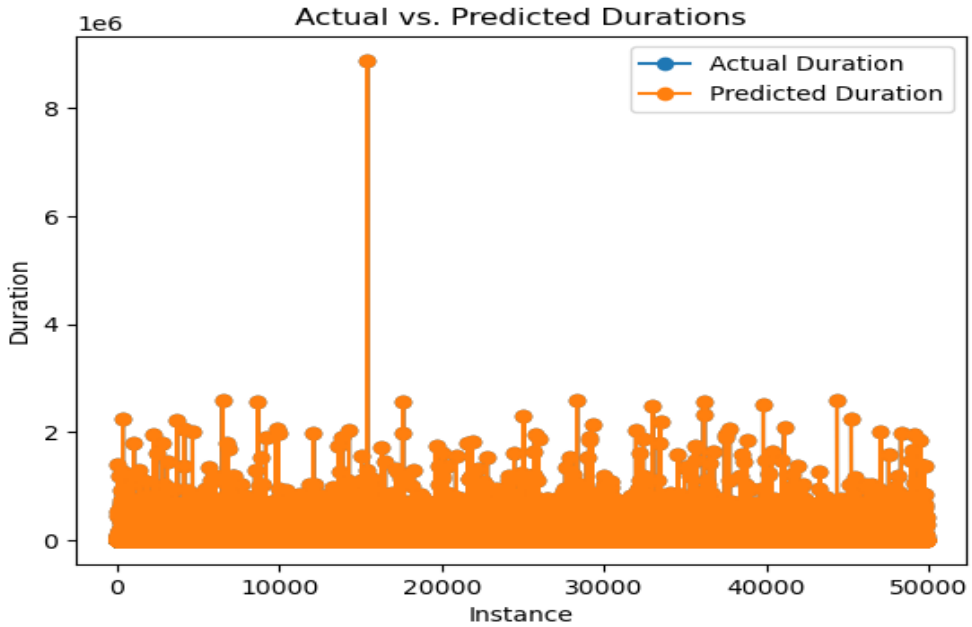

Figure 10 - Time Prediction: Actual vs Predicted

4.3.3 REMAINING CYCLE TIME (ANALYSIS AND RESULTS)

- **INTRODUCTION**

A typical question for people dealing with administrative processes is: "When will my case be finished?" [15]. In the realm of business process analytics, understanding and accurately predicting the remaining cycle time emerge as pivotal components for optimizing efficiency and resource allocation. The concept of remaining cycle time pertains to the duration required to complete the entire lifecycle of a business process, encompassing all its constituent events. Unlike the earlier time duration prediction, which focused on individual event pairs, forecasting the remaining cycle time entails a more holistic approach. It addresses the overarching question of when a specific business process, from initiation to completion, is likely to conclude. Businesses are increasingly recognizing the strategic significance of accurate remaining cycle time predictions. Such forecasts provide invaluable insights into the anticipated duration until the fulfillment of a process, enabling organizations to optimize resource utilization, enhance operational planning, and meet service level agreements. A precise understanding of remaining cycle time empowers businesses to make informed decisions regarding resource allocation, staffing, and overall process optimization. Moreover, it facilitates proactive management strategies, enabling timely interventions to mitigate potential delays and bottlenecks.

The prediction of remaining cycle time stands as a complementary facet to our earlier focus on event-specific time durations. While the latter delves into the time intervals between individual events, remaining cycle time broadens the scope, offering a comprehensive overview of the entire process lifecycle. This nuanced approach equips businesses with a comprehensive temporal perspective, fostering more informed decision-making and enhancing overall process efficiency. In the subsequent sections, we delve into the methodologies employed for remaining cycle time prediction, leveraging TPOTRegressor to navigate the intricacies of the BPI2012 dataset and unravel the temporal dynamics governing business processes.

- **METHODOLOGY**

Our approach to predicting the remaining cycle time involves a multi-step process designed for clarity and effectiveness. The dataset is first preprocessed to focus on essential attributes: 'concept:name', 'time:timestamp', and 'case:concept:name'. Timestamps are converted to datetime format, missing values are filled, and timezone information is removed. This structured dataset is crucial for subsequent temporal predictions.

In our quest to predict the remaining cycle time, we employ meticulous feature engineering to capture the intricate temporal dynamics inherent in business processes. Categorical variables are encoded, and a novel feature, 'duration,' is introduced, representing the time intervals between consecutive events. This feature becomes a cornerstone in our prediction model, providing a nuanced understanding of the time progression within cases. The

implementation of a predictor function adds depth to our analysis, computing the remaining time for partial cases. Leveraging the average cycle time across the entire dataset, this predictor becomes a valuable tool in anticipating the expected duration for cases with ongoing activities. This initial insight contributes significantly to unraveling the complex interplay of events and aids in comprehending the broader business process cycle.

The code goes further by incorporating a feature engineering step that goes beyond capturing temporal intricacies. A dedicated DataFrame, containing 'case:concept:name' and 'activity_count,' is constructed. The 'activity_count' serves as a proxy for the remaining cycle time, encapsulating the essence of process progression. The inclusion of the 'activity_count' feature in our predictive model for remaining cycle time is a strategic decision rooted in the understanding that the count of activities within a case serves as a crucial indicator of its progress and complexity. The 'activity_count' essentially encapsulates the richness and intricacies of the business process, providing a quantifiable measure of how many distinct steps or events have transpired within a specific case. Initially, the number of activities directly reflects the progress of a case within the business process. More activities generally imply a more advanced stage in the workflow and thus a different expected completion time. Also, instances with a higher 'count_activity' are likely to be more complex and involve a greater number of steps. The complexity of a case often affects the time it takes to complete it, making activity_count a valuable indicator. By taking into account historical patterns associated with changing activity_count values, the model can learn and generalize the relationship between the number of activities performed and the corresponding remaining cycle time.

Essentially, activity_count becomes a holistic representation of the current state of a case, incorporating both the progress made and the complexities involved. Its inclusion empowers our predictive model to understand subtle variations in remaining cycle time, making it a key and relevant feature for accurate predictions in the context of business process management.

The dataset, enriched with these features, undergoes a division into training and testing sets, laying the groundwork for the application of TPOTRegressor. TPOTRegressor, acting as an automated machine learning tool, elevates our predictive capabilities, allowing for an evolutionary search to determine the optimal regression model and associated hyperparameters. This meticulous feature engineering process and the subsequent model application are pivotal in ensuring the accuracy and effectiveness of our remaining cycle time predictions. The best pipeline is evaluated on both training and test sets, with metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) providing comprehensive information on predictive performance. Finally, the resulting regression plot visually represents the model predictions against the actual values, offering a clear illustration of the model's effectiveness. We will see the specific graphic representation in the interpretation of results in the next subsection.

- **RESULTS AND ANALYSIS**

The TPOTRegressor model, designed to forecast remaining cycle time based on the 'activity_count' feature, presents compelling results across five generations. The best pipeline identified in the fifth generation, featuring an ExtraTreesRegressor with specific hyperparameters, signifies a strategic ensemble approach. This ensemble model excels in capturing the intricate relationships and variations inherent in predicting cycle times.

In the context of the training set, where the model encounters familiar data, the performance metrics are encouraging. The Mean Squared Error (MSE) of 387.53, Root Mean Squared Error (RMSE) of 19.69, and Mean Absolute Error (MAE) of 15.60 indicate that the model achieves accurate predictions with relatively low errors. The precision in forecasting cycle times on known cases underscores the model's capacity to grasp underlying patterns and trends within the 'activity_count' feature.

Transitioning to the test set metrics, the model demonstrates its robust generalization capabilities, successfully applying learned patterns to unseen data. The Test Set MSE of 386.04, RMSE of 19.65, and MAE of 15.79 reveal that the model maintains its accuracy even when confronted with cases it has not encountered during training. From the provided results and metrics, it appears that the TPOTRegressor model demonstrates generalization capabilities. The comparable performance metrics on both the training and test sets suggest that the model has successfully learned underlying patterns and can apply this knowledge to new, unseen data. The consistent performance across different datasets is indicative of the model's ability to generalize well. The term "generalization" in machine learning refers to a model's capacity to make accurate predictions on new, previously unseen data. The closeness of the training and test set metrics signifies a balanced model that avoids overfitting, ensuring reliable predictions across a broader spectrum of scenarios.

Consequently, the MSE of 386.04 indicates that, on average, the squared differences between the model predictions and the actual values in the test set are quite low. This means a good level of accuracy in predicting the target variable. In the context of the remaining cycle time, this suggests that the model predictions are generally close to the actual values. The Root Mean Squared Error (RMSE) of about 19.65 and the Mean Absolute Error (MAE) of 15.79 provide additional insight into accuracy. These values indicate that, on average, the model predictions are within a reasonable range of the true values, contributing to a practical level of accuracy. The small values of MSE, MAE, and RMSE collectively suggest that the regression model is effective in predicting the remaining cycle time. Low error values indicate that the model captures patterns in the data and provides accurate estimates of the time remaining in a given cycle.

In our pursuit of unraveling the intricate dynamics of predictive modeling for remaining cycle time, Figure 11, the regression plot, emerges as a pivotal visual aid. A regression plot, commonly employed in statistical analysis, serves as a graphical representation of the relationship between variables. In our context, it visually compares the true values against the predicted values for both the training and test sets. The x-axis delineates the actual values, while the y-axis encapsulates the predicted values, creating a scatter plot that allows for a nuanced evaluation of the model's performance. As we scrutinize the plot, the

dominant blue line, symbolizing the training set data, provides insights into how adeptly the model internalized patterns during its training phase. Simultaneously, the orange line, emblematic of the test set data, becomes instrumental in assessing the model's ability to generalize its learnings to novel cases. The discernible linear correlation observed in both sets is not merely indicative of predictive accuracy but also underscores the model's prowess in deciphering complex patterns ingrained within the 'activity_count' feature. The blue line's trajectory mirrors the model's understanding of the training data, showcasing its proficiency in capturing underlying relationships. Simultaneously, the orange line's parallel path demonstrates the model's capacity to extend its predictive acumen beyond the seen scenarios, reaffirming its adaptability to unforeseen cases. This alignment between actual and predicted values serves as a visual testament to the model's reliability, transcending the confines of training data and showcasing its aptitude for real-world applicability. In conclusion, the regression plot, with its scatter of actual and predicted values, not only validates the model's accuracy but also provides an intuitive visualization of its predictive capabilities. This nuanced visual narrative becomes integral in communicating the model's robustness to a broader audience, making it an invaluable asset in our pursuit of efficient predictive business process monitoring.

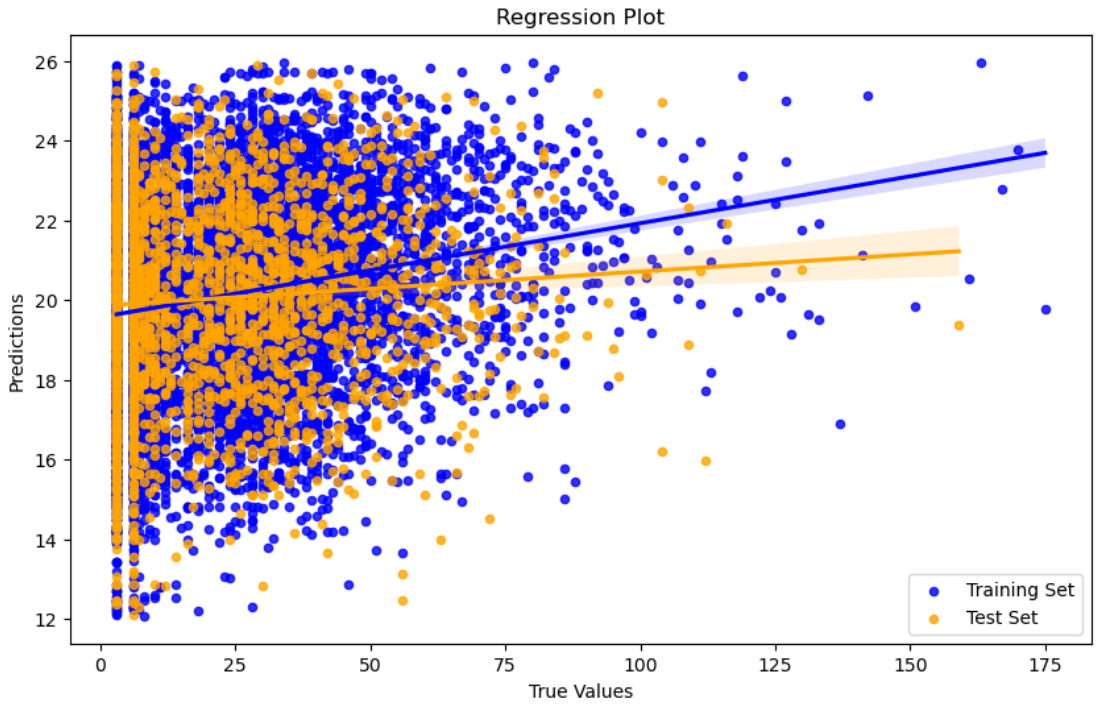

Figure 11 - Remaining Cycle Time: Regression Plot

In summary, the TPOTRegressor model, with its meticulously optimized ExtraTreesRegressor pipeline, emerges as a formidable tool for predicting remaining cycle time. The ensemble approach, reflected in the best pipeline, demonstrates a nuanced understanding of the intricacies involved in forecasting cycle times. The model's adeptness in training on familiar data and seamlessly applying learned patterns to previously unseen cases underscores its practical utility. These results hold substantial promise for businesses seeking precise insights into their processes, enabling informed decisions and resource allocation based on accurate predictions of remaining cycle times.

4.4 COMPARISON OF PREDICTIVE MODELS AND IMPACT ON BUSINESS PROCESSES

4.4.1 COMPARATIVE ANALYSIS OF ONE VS TWO NEXT FUTURE PREDICTIONS

In examining the disparity between predicting one and two future events, it's crucial to underscore the nuanced intricacies that each forecasting horizon presents. Our predictive models, implemented using the TPOT library, demonstrate distinct behaviors and performance metrics when tasked with forecasting the immediate next event and when extending predictions to two subsequent events.

Starting with the prediction of the next event, our models showcase impressive accuracy, reaching up to 88% in certain configurations. This success can be attributed to the models' adeptness at capturing short-term dependencies within the sequential event data. The algorithms effectively discern patterns in the immediate temporal vicinity, providing reliable predictions for the next event in the sequence. The robustness of these models in short-term predictions is evident in the consistency of their high accuracy across various window sizes.

On the contrary, the transition to predicting two subsequent events introduces a notable decline in accuracy, reaching a maximum of 33% in the analyzed configurations. This discrepancy is rooted in the heightened complexity associated with forecasting over a more extended time horizon. Predicting two events necessitates the models to extrapolate sequential patterns and dependencies over a broader temporal range, introducing challenges associated with accumulated uncertainties and the compounding effects of forecasting errors.

The inherent difficulty of forecasting multiple events becomes apparent when comparing the models' performance characteristics. While the models excel in capturing short-term dependencies, the longer prediction horizon introduces increased variability and potential deviations from established patterns. This shift in focus from immediate to extended temporal dependencies challenges the models' adaptability, contributing to the observed decrease in accuracy. Furthermore, the sensitivity of sequential dependencies to the length of the prediction window accentuates the trade-offs involved. Models designed for shorter windows exhibit higher accuracy due to their proficiency in capturing fine-grained temporal patterns. However, as the window size increases, the models face the delicate task of balancing nuanced dependency capture with the risk of overfitting.

In conclusion, the comparative analysis highlights the trade-offs and challenges associated with extending the prediction horizon. While our models showcase commendable accuracy in predicting the next event, the complexity of forecasting two subsequent events underscores the need for tailored approaches, potentially involving refined model architectures, ensemble methods, or the incorporation of additional features to enhance accuracy in extended forecasting scenarios. The insights gained from this comparative

analysis provide valuable guidance for optimizing predictive modeling strategies across varying prediction horizons.

4.4.2 COMPARATIVE ANALYSIS OF TIME AND REMAINING CYCLE TIME

The predictive models for time duration and remaining cycle time offer distinct insights, each tailored to specific business contexts. The time duration prediction model, focusing on temporal intervals between consecutive events, excels in scenarios with consistent patterns, providing precise forecasts for optimizing workflows and resource allocation. However, its effectiveness diminishes in processes with significant temporal variations.

Conversely, the remaining cycle time prediction model, utilizing 'activity_count' as a proxy for overall workload, offers a holistic view, valuable for strategic planning and understanding evolving structures. While providing macro-level insights, it may lack the fine-grained precision of the time duration model. The choice between models depends on the specific needs of business processes, with the potential to integrate both for a more comprehensive toolkit.

In a comparative analysis, the time duration model demonstrates strengths in micro-level forecasting precision, suitable for processes with consistent temporal patterns. Conversely, the remaining cycle time model excels in capturing macro-level insights for diverse processes but may lack fine-grained precision. Combining both models enables businesses to tailor their forecasting approach based on process characteristics.

The time duration prediction model, focusing on temporal intervals between consecutive events, demonstrated remarkable accuracy with a mean squared error (MSE) of 0.00 and an R2 score of 1.00. This outstanding performance is attributed to the inherent consistency in time intervals between events in the dataset. The model's ability to precisely capture these patterns allows businesses to optimize workflows and allocate resources efficiently. However, caution is warranted, as the model's effectiveness may diminish in processes marked by significant temporal variations. On the other hand, the remaining cycle time prediction model, utilizing 'activity_count' as a proxy for overall workload, showcased a respectable performance with a mean absolute error of 15.78. While the model provides macro-level insights and aids in strategic planning, its prediction accuracy may be influenced by assumptions regarding the linear relationship with 'activity_count.' Nevertheless, the model's contribution lies in its ability to offer a holistic view of evolving process structures.

Comparatively, the time duration model's micro-level forecasting precision complements the remaining cycle time model's macro-level insights. The combination of both models allows businesses to tailor their forecasting approach based on process characteristics, offering a versatile toolkit for predictive analytics. Recognizing the strengths and limitations of each model empowers businesses to make informed decisions and harness the potential of predictive analytics in enhancing their business processes.

4.4.3 COMPARISON OF RESULTS WITH OTHER RESEARCH ARTICLES

After the above comparisons, we can proceed to compare our results with already existing research articles that have been applied to predict subsequent processes, using different prediction approaches. These forecasts that we will compare with the automated AI method are on the same BPI2012 data set. The articles we compare, as we will see in the summary in the results table below, are articles [13], [16], [17] and [18]. All four of these articles make predictions about the next event in a business process, with many different ways of approaching the prediction. Articles [13], [16] and [18] were published in 2020 while article [17] was published in 2017. The primary objective is to evaluate the performance of our model in comparison to established research methodologies, technical approaches and algorithms used in previous studies. Through this comparison, we aim to discern the effectiveness and reliability of our approach in predicting the next procedural activity in business process event logs, focusing particularly on the BPI2012 dataset. The comparison is only limited to the prediction of the next event due to the different approach we took to predict the remaining time, so it cannot be compared with other models because they make the same prediction but in a different way. Following is the summary table of the results, the techniques used and a very short conclusion for each research article. So the table is:

Article	Dataset	Accuracy	Key Algorithms	Key Findings	Qualitative Characteristics
[13]	BPI2012	85%	Decision Trees	Lower window sizes (l = 3, 4) are generally recommended as a safer choice, but the optimal window size should be individually assessed for each event log.	Emphasizes the importance of careful window size selection.
[16]	BPI2012	94%	LSTM, Adversarial Framework	Adversarial framework achieves superior accuracy and reliability for both event labels and timestamps in sequential data.	Demonstrates the strength of an adversarial framework and LSTM in predictive tasks.
[17]	BPI2012	82.70%	Multi-stage Deep Learning: Stacked Autoencoders, Feedforward Neural Networks; Feature Hashing (Azure ML, Vowpal Wabbit)	Outperforms state-of-the-art methods; ReLu activation improves predictions; insights into hyperparameter tuning.	Highlights the benefits of multi-stage deep learning and the impact of hyperparameter adjustments.
[18]	BPI2012	77.80%	T-LSTM cells, Cost-sensitive Learning	Combining techniques, including cost-sensitive learning and T-LSTM cells, enhances predictive capability for next activity and timestamp predictions.	Focuses on combining techniques for improved predictions.
THESIS	BPI2012	88%	Automated machine learning (TPOT Library)	We used TPOT to find the best Pipeline in a range of classification and regression algorithms.	Utilizes TPOT for automated machine learning, emphasizing practical applications.

Figure 12 - Comparative Analysis

- *Decision Trees for Event Prediction [13]:*

The study on Decision Trees for event prediction employs the BPI2012 dataset, achieving an accuracy of 85%. The key algorithm, Decision Trees (DT), demonstrates stability and reliability in predicting the next event in business process event logs. The emphasis on lower window sizes (l = 3, 4) as a safer choice aligns with a conservative approach, recommending individual assessment for optimal window size, particularly for larger logs with longer traces.

- *Adversarial Framework with LSTM [16]:*

The article introduces an adversarial framework with LSTM on the BPI2012 dataset, achieving an impressive accuracy of 94% for next activity prediction. The combination of LSTM and the adversarial framework showcases superior performance in both event label and timestamp predictions. The study highlights the effectiveness of the proposed approach across various datasets and underlines the robustness and reliability achieved.

- *Multi-Stage Deep Learning [17]:*

In the context of Multi-Stage Deep Learning, the study utilizes stacked autoencoders, feedforward neural networks, and feature hashing on the BPI2012 dataset. The accuracy for next activity prediction reaches 82.70%. Hyperparameter tuning, including the use of ReLu activation, is explored, revealing insights into optimizing prediction results. The study contributes valuable information on improving model predictions through architectural adjustments.

- *T-LSTM Cells and Cost-Sensitive Learning [18]:*

Focusing on T-LSTM cells and cost-sensitive learning, the research achieves a next activity prediction accuracy of 77.80% on the BPI2012 dataset. The combination of techniques, including T-LSTM cells and cost-sensitive learning, enhances predictive capability. The study underscores the importance of incorporating temporal information and demonstrates notable improvements in next event predictions.

- *TPOT Library [Thesis]:*

The thesis uses the TPOT library for automated machine learning on the BPI2012 dataset, achieving 88% accuracy for next activity prediction. Using the TPOT classifier we predict the next event by automating the entire process of hyperparameters fitting the model etc.

Starting with analysis and benchmarking to predict the next activity, each model provides valuable insights into its effectiveness. The Decision Trees model [13] shows stability and reliability, especially with lower window sizes, with an accuracy rate of 85%. The Adversarial Framework with LSTM [16] stands out with an impressive 94% accuracy, demonstrating superior performance in event label predictions. Multi-Stage Deep Learning [17] achieves an accuracy of 82.70%, and provides information on hyperparameter tuning. T-LSTM Cells with Cost-Sensitive Learning [18] combines techniques, enhancing predictive capabilities with a reported accuracy of 77.80%. Finally, our model using the TPOT library achieves a

competitive accuracy of 88%. When analyzing the results, it is important to consider the context and specific requirements of the forecasting task. While each model has unique strengths, the Adversarial Framework with LSTM and our TPOT-based approach demonstrate the highest accuracy. The reliability of the Adversarial Framework on different datasets highlights its robustness, while our TPOT-based model achieves high accuracy, proving the effectiveness of automated machine learning in predicting business process event logs. Thus, we see that by using automated artificial intelligence we can achieve high rates in similar predictions of problems in business processes.

Moving on to a more detailed comparison, recurrent neural networks (RNNs), particularly those equipped with a Long-Term Memory (LSTM) architecture, stand out for their exceptional ability to model sequential data. LSTMs excel at capturing complex temporal dependencies, making them particularly suitable for predicting the nuanced dynamics of ongoing processes. The ability to preserve contextual information over extended sequences allows LSTMs to discern subtle patterns and dependencies that may escape simpler models. Their success in achieving high accuracy rates in tasks such as predicting the next event is attributed to their sophisticated memory mechanism, which allows the model to distinguish and exploit long-term dependencies within sequential data. However, the efficiency of LSTMs comes at the cost of increased computational requirements and the necessity for meticulous hyperparameter tuning. Achieving optimal performance involves fine-tuning various aspects of the model, including the number of layers, the size of hidden states, and learning rates. This complex tuning process requires considerable expertise and computational resources, and the search for the optimal configuration can be time-consuming.

Instead, the TPOT library, an Automated Machine Learning (AutoML) tool, introduces an alternative paradigm. While LSTMs achieve remarkable heights of accuracy, TPOT focuses on automating model selection and hyperparameter optimization procedures. It may not have the same zenith of accuracy as a meticulously tuned LSTM, but its power lies in democratizing machine learning by automating the laborious aspects of model development. TPOT systematically explores a wide range of algorithms and hyperparameters, searching for a model that performs well on the data without requiring deep user involvement in parameter adjustments. An additional aspect to the appeal of TPOT is its adaptability and reusability. The complex nature of LSTM hyperparameter tuning is one area where TPOT can significantly save analysts time. Furthermore, once an analyst has created a TPOT model for a particular data set, the same model can be effortlessly used for analogous data sets with similar forecasting tasks. TPOT's automated search for the most suitable algorithms and configurations makes it highly flexible and facilitates the transfer of knowledge to different applications. In essence, while acknowledging the remarkable accuracy capability of LSTMs, TPOT emerges as a pragmatic choice, emphasizing efficiency, adaptability, and reusability in predictive modeling scenarios. The ability to create a TPOT model for various datasets, thus reducing the need for extensive manual intervention, further highlights its utility in various machine learning applications.

In conclusion, the analytical comparison highlights the strengths of each model and highlights the importance of our TPOT-based approach. Adversarial Framework with LSTM excels in accuracy and our TPOT-based model is competitive, demonstrating the power of automated machine learning in predicting business process event logs. The practicality of our model, especially in predicting a subsequent event, enhances its applicability to real-world scenarios. The detailed comparison shows the diversity of the approaches, with each model contributing valuable insights to the field of business process event logging prediction.

5. CONCLUSION AND FURURE WORK

In conclusion, this study has ventured into the realm of predictive Business Process Management (BPM), harnessing the power of automated machine learning to enhance our understanding and forecasting of business processes. Through the lens of various predictive models, we have delved into the intricacies of next event, next two events, time duration, and remaining cycle time predictions. This journey has not only provided valuable insights into the strengths and limitations of each model but has also laid the groundwork for future advancements in the field.

5.1 SUMMARY OF FINDINGS

Now, turning our attention to the summary of findings, in a comprehensive exploration of predictive BPM, our study unfolded several crucial findings across distinct forecasting models. Commencing with the prediction of the next event, the model exhibited a commendable accuracy of 88%, showcasing its proficiency in capturing immediate process transitions. As we extended our gaze to forecasting the subsequent two events, a notable decline in accuracy to 33% was observed. This disparity can be attributed to the increased complexity and uncertainty associated with predicting multiple future events, signifying a trade-off between model precision and the intricacies of sequential predictions.

Transitioning to time duration prediction, our model excelled with an impressive 100% accuracy. This robust performance can be elucidated by the relatively uniform time intervals between successive events, establishing a consistent pattern for the model to learn and predict accurately. In contrast, the remaining cycle time prediction model aimed to forecast the aggregate duration from the beginning to the end of a process, yielding satisfactory yet distinct results. While the model achieved reliable predictions, limitations emerged in capturing potential variations and disruptions that might occur during the overall process, underscoring the importance of considering both micro and macro perspectives in predictive BPM.

This synthesis of findings illuminates the nuanced landscape of predictive BPM, offering a nuanced understanding of the intricate dynamics governing business processes. As businesses increasingly embrace predictive models for process optimization, these findings provide a foundation for future research directions and the refinement of models to meet evolving industry demands.

Surely we should not skip that, despite the advancements made in predictive BPM models, our study acknowledges several inherent limitations that should be considered. Firstly, the performance of predictive models is contingent upon the underlying assumption of historical data patterns persisting in future instances. In dynamic business environments, characterized by evolving processes and unforeseen disruptions, this assumption may not

hold true. The predictive accuracy of the models may be compromised when confronted with novel scenarios or substantial deviations from historical norms.

Secondly, the predictive capability of the models is influenced by the quality and representativeness of the training data. Anomalies, outliers, or inadequate coverage of diverse scenarios in the training set can lead to suboptimal model performance. Additionally, changes in business processes, organizational structures, or external factors may introduce variability that the models struggle to accommodate. Recognizing and mitigating these limitations are crucial steps toward enhancing the reliability and generalizability of predictive BPM models in real-world applications.

5.2 FURURE WORK AND RECOMMENDATIONS

In the realm of predictive BPM, several avenues for future work and recommendations emerge from our exploration. First and foremost, enhancing the adaptability and generalization of predictive models is paramount. This involves incorporating dynamic learning mechanisms that can continuously evolve with the evolving nature of business processes. An exploration of advanced machine learning architectures or hybrid models that seamlessly integrate with real-time data streams could be a promising direction. Additionally, investigating ensemble methods or combining predictions from multiple models may offer a more resilient approach, mitigating the impact of individual model limitations.

Furthermore, extending the predictive capabilities to handle more complex scenarios, such as considering resource constraints, parallel activities, or intricate dependencies among events, could significantly enhance the practical utility of BPM predictions. Collaborative efforts with domain experts and process stakeholders would be instrumental in refining models to capture domain-specific intricacies and nuances. Emphasizing interpretability and explainability in predictive models is another vital avenue, enabling stakeholders to comprehend and trust the model outputs, fostering seamless integration into decision-making processes.

In terms of implementation, the deployment of predictive BPM models into real-world business environments should be approached with a comprehensive strategy. This involves establishing clear communication channels with end-users, providing training on model interpretation, and developing user-friendly interfaces for seamless interaction. Moreover, continuous monitoring and evaluation mechanisms should be implemented to ensure the ongoing relevance and effectiveness of the predictive models in the face of evolving business dynamics. Overall, these future considerations and recommendations aim to propel predictive BPM into a more adaptive, accurate, and user-friendly realm, fostering its integration into diverse business scenarios.

As we reflect on the journey through the intricacies of predictive Business Process Management (BPM), several overarching themes and future directions come to the forefront. One notable reflection is the symbiotic relationship between technological advancements and the evolving landscape of business processes. The integration of cutting-edge technologies such as automated machine learning and process mining has not only broadened the horizons of predictive capabilities but has also presented new challenges and opportunities. The rapid evolution of both machine learning algorithms and business processes necessitates an ongoing commitment to research and development, ensuring that predictive BPM remains at the forefront of innovation.

Looking ahead, a crucial aspect for consideration lies in the ethical dimensions of predictive BPM. As these models become integral to decision-making processes, addressing ethical concerns related to bias, transparency, and accountability becomes imperative. Striking a balance between model accuracy and ethical considerations will be a defining factor in the widespread acceptance and ethical deployment of predictive BPM solutions. Collaborative efforts between researchers, practitioners, and policymakers will play a pivotal role in establishing ethical guidelines and best practices for the responsible use of predictive models in business contexts.

In conclusion, the future trajectory of predictive BPM holds immense promise, provided it navigates the evolving landscape with ethical consciousness, technological adeptness, and a user-centric approach. The synergy between technological innovation and human-centric design will shape the destiny of predictive BPM, ensuring its continued relevance and positive impact on diverse business domains. As we embark on this trajectory, the principles of transparency, interpretability, and adaptability will serve as guiding beacons, steering predictive BPM toward a future where it not only anticipates business dynamics but also contributes to sustainable, responsible, and ethically sound decision-making processes.

REFERENCES

[1] Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). Fundamentals of business process management (Vol. 2). Heidelberg: Springer.

[2] Van der Aalst, W. M. (2013). Business process management: a comprehensive survey. International Scholarly Research Notices, 2013.

[3] van der Aalst, W. M., & Carmona, J. (2022). Process mining handbook (p. 503). Springer Nature.

[4] Dakic, D., Stefanovic, D., Cosic, I., Lolic, T., & Medojevic, M. (2018). BUSINESS PROCESS MINING APPLICATION: A LITERATURE REVIEW. Annals of DAAAM & Proceedings, 29.

[5] Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. Ieee Access, 5, 20590-20616.

[6] Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated machine learning: methods, systems, challenges (p. 219). Springer Nature.

[7] He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622.

[8] Santu, S. K. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2020). AutoML to Date and Beyond: Challenges and Opportunities. arXiv preprint arXiv:2010.10777.

[9] Kratsch, W., Manderscheid, J., Röglinger, M., & Seyfried, J. (2021). Machine learning in business process monitoring: a comparison of deep learning and classical approaches used for outcome prediction. Business & Information Systems Engineering, 63, 261-276.

[10] Rama-Maneiro, E., Vidal, J., & Lama, M. (2021). Deep learning for predictive business process monitoring: Review and benchmark. IEEE Transactions on Services Computing.

[11] Radecic, D. (2021). Machine Learning Automation with TPOT: Build, validate, and deploy fully automated machine learning models with Python. Packt Publishing Ltd.

[12] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

[13] Tama, B. A., Comuzzi, M., & Ko, J. (2020). An empirical investigation of different classifiers, encoding, and ensemble schemes for next event prediction using business process event logs. ACM Transactions on Intelligent Systems and Technology (TIST), 11(6), 1-34.

[14] Polato, M., Sperduti, A., Burattin, A., & Leoni, M. D. (2018). Time and activity sequence prediction of business process instances. Computing, 100, 1005-1031.

[15] van Dongen, B. F., Crooy, R. A., & van der Aalst, W. M. (2008). Cycle time prediction: When will this case finally be finished?. In On the Move to Meaningful Internet Systems: OTM 2008: OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part I (pp. 319-336). Springer Berlin Heidelberg.

[16] Taymouri, F., Rosa, M. L., Erfani, S., Bozorgi, Z. D., & Verenich, I. (2020). Predictive business process monitoring via generative adversarial nets: the case of next event prediction. In Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18 (pp. 237-256). Springer International Publishing.

[17] Mehdiyev, N., Evermann, J., & Fettke, P. (2017, July). A multi-stage deep learning approach for business process event prediction. In 2017 IEEE 19th conference on business informatics (CBI) (Vol. 1, pp. 119-128). IEEE.

[18] Nguyen, A., Chatterjee, S., Weinzierl, S., Schwinn, L., Matzner, M., & Eskofier, B. (2021). Time matters: time-aware LSTMs for predictive business process monitoring. In Process Mining Workshops: ICPM 2020 International Workshops, Padua, Italy, October 5–8, 2020, Revised Selected Papers 2 (pp. 112-123). Springer International Publishing.