

**Using AI and machine learning for predictive analytics in estimating the growth in the demand of delivery services post COVID-19**



**Submitted by**

**STUDENT NAME**

**STUDENT ID**

Table of Contents

[Background (overview of topic and motivation) 2](#__RefHeading___Toc3558_3973182502)

[Aim 4](#__RefHeading___Toc18959_1336096292)

[Objectives 4](#__RefHeading___Toc3562_3973182502)

[Project Overview 5](#__RefHeading___Toc18961_1336096292)

[Scope (what will I do? How it work?) 5](#__RefHeading___Toc18963_1336096292)

[Audience (Who is it for?) 5](#__RefHeading___Toc18965_1336096292)

[Summary of existing approaches (e.g. Competitive analysis, if appropriate) 5](#__RefHeading___Toc3570_3973182502)

[Brief summary of existing literature (e.g. Annotated bibliography or initial literature review with brief summary of sources) 7](#__RefHeading___Toc18967_1336096292)

[13.1.Research of software development process (e.g. agile approach, requirements, design, implementation) 9](#__RefHeading___Toc1146_99293359)

[13.2.Technology (implementation tools and resources, such as hardware and software) 17](#__RefHeading___Toc18976_1336096292)

[13.3.Version Management plan (e.g. Git repository or shared drive) 17](#__RefHeading___Toc3578_3973182502)

[15.1. Project Management: activity, schedule (i.e. Gannt or other, showing activities, deadlines), data management, Deliverables 24](#__RefHeading___Toc18969_1336096292)

[15.2. Risk: risk analysis as informed by current progress. Resolved risks and the success of the mitigation strategy. Changes to the project plan as the results of risks. Future risks 26](#__RefHeading___Toc18971_1336096292)

[15.3. Professional issues: identification and discussion of relevant legal, social, ethical, and environmental issues in the context of the project. 26](#__RefHeading___Toc18973_1336096292)

[18.1. A. Links to the data sources: 32](#__RefHeading___Toc18289_1336096292)

# 8. GLOSSARY

AI – artificial intelligence

CAGR - Compound Annual Growth Rate

CD - concept drift

CRM - customer relationship management

DB -database

DS - demand Services

EIA - Error Intersection Approach

HW - hardware

ML – machine learning

PA - Predictive Analitycs

PC – personal computer

SCD - Supply Chain Data

SQL – structure query language

SPA - System for Predictive Analysis

SW - software

TF - TensorFlow

TS - Time Series

# 9. ABSTRACT

In the present research project “Using AI and machine learning for predictive analytics in estimating the growth in the demand of delivery services post COVID-19” the System of Predictive Analytics(SPA) is proposed, the challenges on its creation are considered. The SPA design is divided into several stages, among them Concept Drift for regression analysis unit that considers the data related to sudden change input is described. The implementation of this unit is presented. The overall recommendations for SPA design are also described.

***Keywords***—AI, Cloud Computing, Machine Learning, Prediction of Time Series, Predictive Analytics, Concept Drift, Regression

# 10. **Acknowledges**

# 11. Introduction

## Background (overview of topic and motivation)

According to the current estimations, the total size of predictive analytics market worldwide will grow up from USD 10.5 billion in 2021 to USD 28.1 billion by 2026, at a Compound Annual Growth Rate (CAGR) of 21.7% during the forecast period[]. One of the main factors, that propels such progress is the new emerging techniques of using AI and ML for making the predictions of these analytics more accurate and the appearance of the more complex models for business forecasts. Therefore, new acquisitions and product launches in this market create the demand for novel Predictive Analytics software and services.

The importance of predictive analytics as the part of business data analytics follows from its ability to predict forecasting trends and behaviors of the market. Thus, correct and reliable predictive analytics allows making correct business decisions like ordering of the cheapest accessories in the right time, producing of the most demanding goods, predicting the most essential services, etc. In order to be practical, the methods of this analytics should generate insights for the future with a controllable degree of precision.

The predictions can be based on historical data and different techniques such as statistical modeling, numerical equation solution and others. Current boom of artificial intelligence techniques and rapid growth of computing services creates the trend of using machine learning and AI in predictive analytics. The using of ML and AI techniques not only helps in calculation and improving of precision for the well-known methods, like Regression, Clustering, etc., but also became the source for designing new approaches for analytics by building, for example, complex neuron networks that successfully predicts future sales, prices, outcomes, etc[].

The mathematical possibility to make the correct prognosis is grounded on the intrinsic in the physical world property of the different parameters to become the same in the same conditions. However, the variety of the natural and economic factors and their unexpected change often breaks even the most successful prediction models.

One of such challenging factors, that were arisen to the whole world, and disrupt businesses, economies, and usual life of the people, was the spreading of COVID-19. The pandemic itself caused a great havoc not only by threatening life and health of the people, but impacted the way of functioning of the society institutions as well as the economic and trade relationships. The need of enterprises and markets to be resilient under the threat of the pandemic has forced the business managers and economists to review the existing prediction models and has created the demand for the new predictive and analytical models allowing to calculate the impact of the changes, that were caused by COVID-19.

One of the most important outcome of the governmental politics in situation of the struggling with the decease and multiple forms of lock-downs is the changes in the structure of the marketing and delivery of goods to the customers, that leads to the rapid alterations in the business of goods delivery. The pandemic and pandemic control measures affects these services deeply and lead to the change of the situation on the market. However, this impact was not always one-sided, and the problem of forecasting of demand of delivery services in post COVID-19 world is not as simple as it seems to be from the first glance, and some of the predictions that were produced at the start of the pandemic, were failed to be correct. For example, some prognosis supposed that the demand on the food delivery services would rapidly rise during lock-downs, but that forecast wasn’t correct for all the regions and various factors affects the real growth of demand services. Therefore, the investigation of the factors that affects demand services sales and discovery of the reliable model of this growth in post COVID-19 world is the important and valuable task, that is demanding good solution.

In this work, the task of prediction of sales in food delivery is researched. Usually, this task is solved on the base of Time Series prediction methods, i.e., the data of the sales along with the data, that needed for correct predictions, like distances from contractors to delivery points, traffic factors, prices of goods and services, are coming to the System for Predictive Analysis (SPA) chronologically in periodic times. There are many well-known solutions to the problem, that performed reasonably good in pre-pandemic times. These models could be implemented as the part of the computer-aided automated System for Predictive Analysis (SPA) for quick automated estimation of the prediction. Such systems could help business analytic to make correct management solutions.

However, the upheaval of the COVID-19 drastically changed the correctness and reliability of these models, therefore impaired the work of SPA. This challenge cause the realignment of the algorithms for SPA. For example, we need to detect automatically the change in data, that caused by important factors, such as stopping of important institution operation, cease of public transport work, lock-down in specific area, so on. The next item of further SPA algorithm work could be the tweaking of the corresponding parameters of the models to reestablish their accuracy in required metrics.

We propose to tackle the described issue by implementing of Concept Drift (CD) detection and designing of the software (SW) system for SPA that uses CD for improvement of a prediction accuracy in ML for delivery services.

## Aim

The aim of the project is to design overall structure of the computer system, that allows to design, implement, test, and experimenting on ML models for predictions of the sales of the Demand Services (DS) from the historical data of the sales, based on the Time Series(TS) type of information. These data includes information about the kind of services and characteristics of the customers, transport and welfare in the area, so on. Also, this system is capable for the user to modify these data and input some other crucial and hard for formalization parameters as the lock-down type, state of the lock-down in the area, etc.

## Objectives

The objectives of the proposed project:

- to design the structure of the SPA systems and their interactions with the databases (DB) for their operation in post COVID-19 market environment, to present its dataflow;

- to propose the unit that allows to input the data non-related to DB for the changes in prognosis detection;

- to propose and implement the updates in the structure of data for time-series (TS) prediction with Concept Drift detection methods for SPA as the unit of the system that tweaks predictions methods for different condition

- to present recommendation for the SW of SPA implementation and testing.

## Project Overview

### Scope (what will I do? How it work?)

The designed SPA algorithms and system allows operating with different methods of TS forecasting and analytical methods, like Autoregression, Moving Average, etc by using DB with historical data on the DS and information from Internet for more correct PA in times of change. The proposed SPA allows for the user (data analytic) to modify the data and input non database type of data for tweaking the performance of forecasting methods. Also, these models can be evaluated on the base of their correctness by various metrics on run-time data.

This system is also wielded with continuous and active learning SW means, that improve SPA performance by adjusting the parameters of the models on the run, CD detection algorithms are helping to adapt these learning models for different kinds of lock-downs and pandemic situations in automated mode.

### Audience (Who is it for?)

The proposed system is designed for the business analytics who are struggling to adapt their predictive models with the new post COVID-19 reality. The solution should help to create the new models and tweak the existing ones for the presence of the new data, that should be considered for evaluating the impact of COVID-19 on predictions of sales. Also, the proposed SPA SW methods are supposed to suppress this negative effects of existing approaches in predictive analysis.

# 12.Background review

## Summary of existing approaches (e.g. Competitive analysis, if appropriate)

There are some commercial systems in the current SW market, that claim themselves as automated systems for predictive analytics []. Despite the fact that some of these systems are popular, they have some shortcomings, besides, of course, the obvious one – that they are non-free. For example, the most of them do not allow for the user to tweak the algorithms of forecast, limiting the approaches of predictions by the predefined set of classical methods, not all of them detect anomalies and shifts in data, so on. Also, the tasks of preparing dataset and linking with online DB are remained on the shoulders of their user, therefore leaving the issues regarded in the present paper mostly unsolved.

From the other side, there are some open source and commercial SW systems, designed especially for food delivery services and other types of delivery services business[]. These software applications could be tweaked for further usage in PA, but these features are not naturally intrinsic for them, and they needed to be seriously modified for such exploitation. In particular, the module for time-series (TS) analysis is required for this kind of the purpose.

Comparing to the mentioned before tasks, there are much more high-quality SW solutions, even free ones, for the different TS analysis tasks[]. We can include to these kinds of SW also means of well-known frameworks and programming languages like Mathlab, Python TF and Scilearn, R, Statistica, etc., that have the tools for working with TS and the rich libraries for various forecasts algorithms. But these software products can only be considered as the tool for the implementation of the SPA, because it has only useful blocks for SPA and have neither methods for working with DB, nor methods for business type problems description. Also, there is no means for Concept Drift analysis in these products yet.

There are some examples of drift-aware systems with different algorithmic traits and use cases. For examples, **simple** and **smart retraining** systems – that just retrain models on detected CD, Learn++ that updates voting weights of each model on CD, Streaming Ensemble Algorithm (SEA), Dynamic Weighted Majority (DWM) / AddExp and some others[].

These approaches have their advantages and shortcomings, and it is not clear which one of them is the most suitable for the considered task. However, the Error Intersection Approach (EIA) [] showed a good result for the similar task, therefore we believe it can be regarded as a good starting point.

There are also some open SW frameworks, that implement basic CD detection and analysis algorithms (e.g. CD\_MOA(Bifet,2013) and [*https://github.com/rsdevigo/pyAdwin*](https://github.com/rsdevigo/pyAdwin)). These practical frameworks could be used for testing purposes and as the base implementations.

## Brief summary of existing literature (e.g. Annotated bibliography or initial literature review with brief summary of sources)

The literature on usage of Automation, ML and TS techniques for predictive analytics is huge [], however it is constantly updated by the novel approaches, e.g. Automated ML algorithms[], Active Learning[], etc.

Nevertheless, most part of these articles and books are concentrated mostly on the mathematical components of the proposed system, while the scientific approach to the SW design of such systems is rarely considered there. One of the exceptions is the paper of Yüksel A. et al[], which studies the problem of design for the predictive system of spam detection, however their approach is focused on text data and ML concepts, that are not fully transferred to business analytics data. Another up-to-date research in this field is Krauß, 2020,[] that investigate the problem of scaling ML-based projects from scientific solution to production usage. This work propose the pipeline for the deployment of the SW ML systems and shows the possibilities and limits of applying AutoML in production.

The survey paper of Seyedan and Mafakheri, Fereshteh. (2020), [] presents modern methods and approaches for building of ML system for predictive analytics, the structure and content of databases in them, and demonstrate various solutions for controlling supply chains with the help of SPA-like systems.

There are also worth notion survey works of Nguyen,2017[], that display systematic approach to the literature review in the domain of Big Data for Supply Chains (SC) analysis and similar by the subject to the previous two papers, more modern paper of Tamym, 2020[], which concerned more in the aspects of SW/HW systems used SC analysis.

The work of Zhang, 2016 [], presents the methodology for creating ML system for failure prediction, that uses both ML and CD techniques for mission-critical IT systems using streaming operational logs, and enabling earlier failure predictions through the LSTM approach on historical data. The authors of this paper claim that their methods outperforms the other similar methods in metrics PR-AUC. This approach could be applied and to the problem of predictive analytics for business SPA. However, again, this method is bound to the text representations of the TS information. Therefore, for our task that uses numerical nature of the data, the elaboration of this approach and transferring to the SW application SPA designed for business forecasts, especially in the field of delivery services is needed.

Despite the emerging great amount of the works, investigating the impact of COVID-19 on statistical models[], and, particularly, on business models[] or urban traffic[], to the best of our knowledge there was no published work that research the problem of post COVID-19 analytics simultaneously with the modification of SW design for ML for predictive analysis and rearrangement of data flow in these systems with taking into account of data concerning pandemic. Therefore, complex theoretical research n this filed is almost definitely new, and important problem.

There is a vast amount of works dedicated to Concept Drift detection and correction[], including survey papers, that cover versatile methods in ML field. For example, Zliobait ̇et al (2015), describe different mathematical approaches for CD and also their practical implementations and tests approaches. Many of them regarding Time Series predictions and Business analytics trends forecasting[]. These papers, as the rule, dedicated to algorithmic methods of active learning and do not regard SW and HW structure design methods

M. Straat et al. 2022, describe how CD detection can improve the work of the neural networks, regression and clustering methods and provide the scheme of practical implementation of a framework that uses CD and predictive techniques simultaneously. However, they do not discuss SW implementation o these methods.

In the work of Lucas Baier et al.,2020, the method of Error Intersection Approach (EIA) is presented. They demonstrated some practical examples that showing effectiveness of using CD detection for the task of business prediction combined with regression and ML techniques. They prove that concept drift handling significantly outperforms a static model in their use-case of business data mining. They conclude the paper that another use-cases and another methods of CD and combinations of approaches to the design of similar systems are required further studying.

Another interesting case is discussed in the paper of Weingarten and Spinler, 2020,[] where it was shown how SPA and prediction data can be used to optimize delivery times for customers and propose the model for purchasing of online sales, but they have not considered drift in the data and issues with delivery services.

Zeynab Bahrami-Bidoni and Benoit Montreuil, 2021, shows the practical example of the system, that can use ML models and evaluation of correct predicting of behavior of urban logistics in the times of change. This paper mostly focuses, however, on the minimization of delivery time through predictive models of the customer, without calculating impact of changes because of COVID-19 and other reasons, so we believe that combining of mentioned approaches can lead to further improvement of regarded System of Predictive Analytics operation.

# 13.Methodology (Description of the research and the development methodology)

## 13.1.Research of software development process (e.g. agile approach, requirements, design, implementation)

The complexity and many-fold structure of the project of design of the System for Predictive Analytics in the demands of Delivery Services in post COVID-19 times leads to the complex and many-fold structure of the development process.

In the process of the research it was found, that while tho whole project is better to be developed in the methodology of Prototype Model as it is commonplace for the research project, but the separate parts of the project must be done using another types of the SW development life cycle.

Let’s discuss these parts of the project and their types in details.

1) Literature review

The first thing, needed to be done for creating of the proposed System for Predictive Analytics is to design the structure of the data that have to be processed in the system. Therefore, some research should be done to create optimal database structure.

For the fulfillment of this task, some research on open databases and paper review have to be performed. The result of this activity should produce the document, containing the structure of the DB.

The classical scheme of methodological approach to literature analysis, for example, to produce the survey papers, or methodologically classify bibliographic items could be described by the following list of steps:

step 1: Material collection, which is a structured process of search and

delimitation of articles;

step 2: Descriptive analysis, that is needed for generating general characteristics of the studied literature;

step 3: Category selection, where the construction of a classification framework based on a set of structural dimensions and analytic categories is created;

step 4: Material evaluation, which analyses the article based on the proposed classification framework and interprets the results.

The adaptation of this approach leads to the following interpretation within the framework of the current project of designing SPA for predictive analytics in estimating the growth in the demand of delivery services post COVID-19 world. This method resulted to the next stages.

**Material collection**

Before searching for required references, it is needed to identify an effective set of keywords that can capture the synthesis of existing literature related to our research topic. Our taxonomy of the keywords came up with the three groups:

Group 1: Words related to “Predictive analytics”: “Data analytics”, “Machine learning”, “Predictive analytics”, “Predictive analysis”, “Regression”, “ATMA”, “Time Series Forecasting”.

Group 2: Words related to “Demand in Delivery Services”: “Supply Chain”; “Delivery”; “Delivery Services”.

Group 3: Words related to “COVID-19 Data Analytics”: “COVID-19 big data”; “COVID shift in data”; “Delivery Services COVID”.

The reference searching was conducted based on possible combinations of power two between those three classes of keywords within the timeline from 2015 to the start of 2022 on well-known academic databases, i.e., Science Direct, Emeralds, Scopus, EBSCO, and IEEE Xplore. This particular timeline is chosen to suppress the non-relevant and not actual research on PA, that could be classical, but do not relate to post COVID-19 shifts in the researched domain.

Descriptive analysis

Descriptive analysis is devoted to classification of the obtained sources. However, as long as we're considering the particular task, these step was skipped, or more formally, performed on the first step of reference analysis.

Category selection

The goal of category selection step is to conceptualize the classification framework, by formal presentation of structural dimensions and analytic categories. The given problems constitute three structure dimensions of the framework which were stated in the first step. By eliminating irrelevant papers, the works, that are considered as the base solutions were selected.

Material evaluation

On this stage, only works that relevant to the topic of the project were selected. The corresponding methods, that can be used for design of SPA database and be implemented as the part of SW structure of PA system were examined, and corresponding solutions were laid as the basic for further investigation and analysis. Also, the reference analysis and bibliography titles were produced on this stage of the project.

The result as the final part of this first step of the project is Background Review of the present document.

The development of the ML models with whole industrial pipeline goes through the next stages:

- integration, that means design of the SW needed;

- preparation, that means preparation of the data to the required pipeline;

- modeling, that means the creation of the working mathematical model and both theoretical and practical evaluation of the model;

- model deployment, that consists of SW implementation of the given model, validation of it and testing of the overall system.

That approach and analysis of the corresponding references from the stage of the literature review leads us for the development of the current project to the next stages:

*- DB structure design;*

*- flexible dataflow design for SPA system, that includes the information connected to the pandemic or the other important information for PA;*

*- the enclosure of* *the methods of concept drift to the conceptual model of the proposed SPA structure;*

*- the recommendation of ML model implementation's development;*

*- the development of the recommendations for the testing and evaluating of the designed system.*

*Let’s consider describe here items in detail.*

2. Database structure design

Almost every automated system design started with definition of the data, it should process. So, the next step of the research was analysis of the types of data and creating of Entity-Relation Diagram.

Some research in this domain were conducted in the works of Seyedan, 2020[], Tamym,2020[], Nguyen, 2017[] and others.

Summarizing these surveys, Supply Chain Data (SCD) data are divided into the next groups:

1. Stores data (code, name, location, etc. of the store);
2. Customers data (id, name, mail, address, location of customer);
3. Product data (code, category, description, image, name, price, status of product);
4. Sales data (value in sales, value per customer, etc);
5. Orders data (id, customer id, product code, date, price, discount, amount and profit in order, etc);
6. Shipping data (date, mode, time of sipping, time of shipment);
7. Delivery data (status of delivery, late delivery risk).

These data could be easily interpreted as the tables in SQL database, normalized (we could set ids/codes types of attributes set as the keys for these tables), and so the natural solution was to keep them into SQL database. Also, the links and interconnection between tables are naturally described as relation-based. Thus, simple relational SQL DB seems fit for our purpose of SPA design. Open database SQLite was chosen as this database, because it is free, open-source, open-licensed, light-weight, and easy manageable SQL database application.

We may also note, that for the sake of simplicity we could rid off table with Shipping data, as long as we do not need to analyze full Supply Chain of the goods, and for the task of predictive analytics we can suppose, that amount of them are sufficient for every customer, and analytic do not need to know the parameters of shipping for managing the delivery.

Also, table of Sales is the aggregation table, thus, this table can be omitted from the total scheme of data and be present only in the output of the SPA.

The methodology to design and build of the database was chosen as Big Bang Model. The reason it was chosen, is that it does not have any defined process and does not require much planning and scheduling. In this current stage of research, there is no definite requirement analysis. The scope of the project is unclear, and the goals are vague, that’s why coding and development of the product performed only by current understanding. There is no testing team and no formal testing at this stage.

3. Flexible dataflow design for SPA system, that includes the information connected to the pandemic or the other important information for PA

The data structure, designed in the previous paragraph, is more or less trivial scheme of dataflow in the Customer Relationship Management (CRM) systems, however the more ambition goal of predicting sales in delivery services in post COVID times demands to take into the account more complex factors.

The new parameters that should be inserted into the model are geographical characteristics, economical data in the considered region and even political situation: anti-pandemic measures, demonstration and street riots so on.

The problem with these kinds of data is that they cannot be processed as the table data, that come as the regular time series, and yet we cannot count on them as the stable data, as they can be changed in time and drastically affect the PA results.

So we need data that are not stored in the DB of the system, yet they could be added as the part of the historical data and considered in PA.

This task requires the statement of the problem and the solution in the form of recommendation for the user, and the code for the input of data.

This stage was relatively ambiguous on the initial stage of the project, therefore Big Bang model is again used here.

4. Data mining for the project of SPA and data engineering

The execution of the task of accurate prediction in PA is tied to the correct fulfillment of the database for the SPA. Therefore, the next important step is the data mining for the DB of delivery services.

It is the definite issue to find a corresponding data, however some open databases, allowable for loading were found through googling.

After the initial formation of the database list, we need to transform the data to the format, required by the demands of the proposed SPA SW ML, and we face with the necessity to perform feature engineering.

Feature engineering is a process of the source data transformation into the format of a modeling dataset. It is needed, because the data are typically residing in a raw, non-aggregated form. Only transformed data then could be fed into ML algorithm cycle of training, validation and testing.

In the current test, as it is in many real-world cases, we have the database full of diverse tables, none of which is digestible by ML in the framework of the desired SPA system.

Moreover, we have the data from the different sources, and they are predictably presented in the plethora of different formats, and we need to transform them to uniform format digestible by the considered SPA software and hardware.

The given task is having to be done by the transformation of data to a format ready for ML by creating so-called features (aka Variables, Input Columns, factors) which roll the raw data-points up to the Target level.

This process often consumes ridiculous amounts of effort. This is exactly why there is a non-formal low of the data science, that this stage of data preparation takes up about 80%-90% of the project time.

The reason of that setback is that it is often manual, slow, error-prone and also completely open-ended because there is no limit to the number of features we can generate from the source data.

Each additional feature can increase the future model's performance - but can slow down the ML implementation significantly.

These factors force us to develop the strict set of features for SPA and implement them in the diagram.

4. The enclosure of the methods of Concept Drift to the conceptual model of the proposed SPA structure

Next step is the Concept Drift algorithms implementation is the detection of rapid changes in the obtained data. This could be done as a special program module on Python and tested with open implementations on sample databases. But, before this type of work should be done, the mathematical model should be built and evaluated. So, in the results part, we provide some calculations to present the proposed solution and its practical implementation.

In this stage of the project we have to build some SW, that could be practically tested, therefore we need some methodology to be used in order to make the tasks with the required level of industrial accuracy.

We choose to use Spiral Model in this step. Spiral model phases are followed in the iterations. The loops in the model represent the phases of the Spiral Model **are break down into** four phases:

* Planning
* Risk Analysis
* Engineering
* Evaluation

1) Planning **phase**: This stage includes requirement gathering, that is done by discovering of the demands to the SW on the process of research and its documentation. SW requirement specification document is created on the next phase.

2) Risk Analysis: The gathering of the discovered risks performed in the time of building of the prototype.

3) Engineering: Coding and initial testing of the SW Once the risk analysis is done, coding and testing are done.

4) Evaluation: On this stage we need to test the resulted code, evaluate the developed part of the systems and plan the next iteration.

**The** Spiral Model on this stage was chosen, because risk analysis on this step of the project is done extensively using the prototype models and any enhancement or change in the functionality can be easily implemented in the next iteration, thus we can use several iterations here, quite contrary from another parts of the project. But usage of this life cycle on this stage allows creating prototype here.

5. Recommendation of ML model implementation development

The stage, described above, must be finished with practical recommendations steps, that described how we can integrate this ML mathematical model into the proposed SPA. So, this stage results on the document for implementation of the SW for SPA.

6. Developing of the recommendations for the testing and evaluating of the designed system

All computer systems, that are designed to work with mission-critical data, have to be thoroughly and meticulously tested. The tests should be conducted on the data received on the first stage, by dividing them on Training, Testing and Validation sets. And, final testing, is checking the system on the real-world data on progress.

So, this step results with the description of practical approaches to tests and document the demands for the test of the proposed system as well as all of its components in one place as the corresponding document.

The final two stage are simple documenting, therefore we do not need to use special type of SW life cycle on them. However, as it was stated in the beginning of this chapter, the overall SW life cycle for the whole system is a Prototype Model.

The Prototype Model is a model in which the prototype is developed prior to the actual software.

This type of methodology have limited functional capabilities and inefficient performance when compared to the actual software. Some functions are not ready and used only to create prototypes. However, such approach creates the mechanism for understanding the customers’ needs.

The sense of the methodology is that when the requirement gathering is done, the quick design is created and the prototype, which is presented to the customer for further evaluation. Based on the reaction of the customer, the real working.

We believe that this type of the SW development cycle fits best for our task, because it uses the advantages of the Prototype Model:

* Prototype model reduces the cost and time of development as the defects are found much earlier.
* A missing feature or functionality or a change in requirement can be identified in the evaluation phase and can be implemented in the refined prototype.
* Involvement of a customer from the initial stage reduces any confusion in the requirement or understanding of any functionality.
* Software prototypes are built prior to the actual software to get valuable feedback from the customer. Feedbacks are implemented, and the prototype is again reviewed by the customer for any change. This process goes on until the model is accepted by the customer.

Further step will be collecting reviews from the customers and transform them into new working cycle for creating actual SW.

The results from each stage will be presented in the chapter, dedicated to the results of the project itself.

## 13.2.Technology (implementation tools and resources, such as hardware and software)

The list of required specialized resources and SW/HW and technologies for this proposed research work against the purpose is summarized in table 1.

|  |  |
| --- | --- |
| Required Resource | Usage |
| Access to Online Digital Libraries i.e. Google Scholar, IEEE Xplore, Science Direct, etc. | To collect the relevant research literature to perform an extensive review of literature pertinent to the proposed research area. |
| Cloud to store Database | Cloud to store Database |
| Github | Repository to store code |
| C++ | Programming language for the implementation |
| SQLite | Database for keeping data |
| Linux Ubuntu 20 | Open OS for working with Python and SQLite |
| PC | Computer for working |

Table 1: List of required specialized resources in proposed research work

## 13.3.Version Management plan (e.g. Git repository or shared drive)

The accurate and safe development requires the usage of backups and also control version tools.

The well known rule of conduct for the ML projects demands to keep the databases on the Server as well as on the HW of the client, and keeping the code in the repositories of Code Version Tools.

The database for the current project is SQLite. The tables of the DB could be stored on the local PC. However, it is needed to keep the data on the Server, or copy it to the cloud (like, for example, AWS). In order to provide sufficient level of data security and not to overuse it, we propose to perform backup every week.

The Code Version Tool was selected as Git – because it is free, and de facto standard tool for the version management. GitHub provides also free of charge usage of Git in the SW project, so we create login in the GitHub.com cite and use it as it is required for the fulfillment of the project.

# 14. Results: detailed documentation of results and testing. Critical evaluation and discussion of results, issues encountered, constraints, limitations, and originality

The results of each stage is shown as separate item in the list.

1. **Literature review**

The result of this stage is presented in the chapter Background Review

1. **Database structure design**

The remaining DB dependency tables can be described by the following relations:

**Order Table:**

**orderId** , date, amount, discount, time\_delivered

→ product\_id,

→ customer\_id

**Customer Table:**

**customerId,** name, region, phone, email, coordinates(address)

**Product Table:**

**productId,** category, name, description, price,

→ store\_id

**Store**

**storeId,** name, region, phone, address,

**Delivery:**

**delivery\_id, status, information**

→ **order\_id**

The tables have id keys, can be sorted, and all substantial parameters are defined, thus they are eligible to be presented in Basic Normal Form.

The processing of these data could be this way performed using SQL on selected DB.

***3. Flexible dataflow design for SPA system, that includes the information connected to the pandemic or the other important information for PA***

The proposed solution of this issue is the table which stores text data and numerical data simultaneously, and each mark of these data also contain information about the time when this particular event is happened (started or ended). For example, we can input, that in 17.04.2020 the partial lock-down is set, or for example that between 12-00 and 17-00 of 19.06.2021 there were serious traffic issues in the given urban area.

So the table and input form for these data looks like:

|  |
| --- |
| time\_id of the event |
| Event description |
| Estimated impact as number from 0 to 1 |

The connection and addendum of this information into ML prediction models comes with the parameter of time, e.g., these data inserted into the tine series of delivery services sales and designed ML algorithm, should be able to correctly recognize and process this data and with the help of CD algorithm correct the results of the model.

**4. Data mining for the project of SPA and data engineering**

The list of the database found are presented in the items below:

1. https://www.businessofapps.com/data/food-delivery-app-market/

1. https://www.statista.com/outlook/dmo/eservices/online-food-delivery/
2. https://www.ibisworld.com/global/market-research-reports/global-courier-delivery-services-industry/
3. https://microservices.io/refactoring/extract-service-delivery-service-step-2.html
4. https://www.sciencedirect.com/science/article/pii/S2352340919303609 : https://data.mendeley.com/datasets/m9z9hw4nsc/1/files/c8ca39ff-83f6-4bda-8311-5e044824b91c
5. https://www.pharmdata.co.uk/

This roster contains only easily found and downloadable links from Internet that requires only registration, and, of course, there is a possibility to find more professional and full data for investigation and PA of the sales in delivery services.

Not all the given links provides full set of data, that needed, also part of them concerns only partial data that are required for the given task of SPA design DB, thus, as usually in AI research the task of DB engineering and feature engineering emerges.

In our particular case, the feature engineering is simplified by the presence of the criteria for the data from the previous stages. Thus, our proposition is to perform this feature engineering is:

* to store all the data in CSV format, it can be easily done by using ordinary Office tools;
* to select data, required for the two previous stages by using language of programming (C++ or Python);
* to store the selected data using this language of programming in SQL database tables .dat.

**5. The enclosure of the methods of Concept Drift to the conceptual model of the proposed SPA structure**

The SW module for the regression analysis, that allows to predict the changes and tweaks the regression prediction is based on the modified model of the Bidoni, 2021, Weingarten, 2020, and Strat, 2016.

The development is carried out in the C ++ programming language. The program aims to simplify calculations when searching for dependencies using pairwise correlation-regression analysis.

Input of the data

The input data will be the prices of product and the price of using taxi services and the value of the uncertainty index (that value COVID impact or something else) at certain points in time. This data is recorded in the corresponding text files product.txt, taxi.txt and factor.txt respectively. Also in the code of the program by means of macros #define we set a constant MAXSDVIG (number of weeks for which we will carry out shift).

The user should enter n (how many pairs of input data we want to process), then the program reads the number from the file as a string and using the writing function "stringtolongdouble" that transforms the string into long double to the third decimal place.

In the future update of the system, we suppose to use ML function that will transform it according corresponding ML model like in the paper of *Zhang, 2016.*

The translated number is written to the array. After reading from the files, we have 5 arrays with data of type long double and the number of elements equal to n + MAXSDVIG.

Then, we perform the proposed CD tweaking algorithm.

The proposed algorithm is divided into the next stages:

1) **Step** **first: correlation search.**

Then, on the loop, the largest correlation coefficient is searching. A negative correlation coefficient is not appropriate because we are looking for a direct relationship, and we do not consider an inverse relationship. The variable bestkor, which after the end of the body of the cycle will contain the value of the highest correlation, assign a value of 0.

We make a shift of j weeks using the loop For, where 0 <= j <= MAXSDVIG. To do this, we take the array of elements from which we look for dependence without offset (the first n elements), and the array of elements that must be dependent on the elements of the first array, we take with an offset of j elements (starting from the j-th element and up to n + j) -th). The value of the variable j, which will be the largest correlation coefficient - will be the time of the manufacturer's response to rising prices. Find the correlation coefficients and write them in a two-dimensional array. Find the largest correlation coefficient by comparing the values ​​of the found correlation coefficients and the value of the variable bestkor. We write the correlation value in the variable maxkor and the value of the shift week j in the variable ned. Also using the functions "poiskb1" and "poiskb2" we find the regression coefficients, which are written in the variables bb1 and bb2 and the coefficient of elasticity using the function "elastic", which is written in the variable elast. Apart from the search for the relationships between product prices, we find the dependence of inflation on time with the initial condition η (0) = inflation at the first moment of the considered period.

2) **Second** **step: calculation of other coefficients and display of** **data**

At this stage, we display:

* the coefficient of elasticity,
* the matrix of correlation coefficients (using the "vuvodmatr" function),
* the regression equation ("vuvodfunc" function).

We also find the standard error (srkvad function) and the approximation coefficient (aprox function) and display it.

**3) Third step:Monte Carlo test.**

With the regression equation found, we perform the Monte Carlo test. The number of times of the test is determined by the constant at the beginning of the program. We find the variance and error in the percentage of regression coefficients.

After 100 Monte Carlo tests and all the the dependencies found, if the variance of the regression coefficients did not exceed a few percent, we can ignore the error.

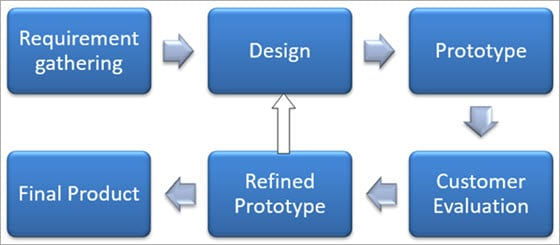
**4) Fourth stage: Runge–Kutta method**

We program the equation obtained in the previous section and look at the results at different values and try to solve the search problem by equating the shock function to zero. When the shock function begins to decline, so our assumptions are too general and need to be clarified by a factor of b in the first place.

The programming module link is paced int Appendix B.

**6. Recommendation of ML model implementation development**

ML model implementation development should be implemented according to the Prototype Model life cycle:



**7. Developing of the recommendations for the testing and evaluating of the designed system**

Before implementation of the complex system, we should confirm that the end-user experience will not negatively be affected at any point. It is important that the software overall ends up meeting the quality standards that were previously defined in the SRS document.

So the next stages of tests are needed to be done after the implementation of the SPA.

#### **1) Alpha and Beta Testing** ‍

Both are pre-release phase testing. Alpha is done at the early stage, while beta testing often involves willing customers in their environment.

‍2) Regulation Acceptance Testing ‍

Determines whether the software meets legal requirements.

‍3) Black Box Testing ‍

The content of a black box is unknown. Likewise, in this type of test, the testers know what an application should do, but are unaware of how it ought to do it or accomplish the task.

‍4) Production Readiness Testing ‍

As the name implies, it ascertains whether the software is ready for usage, with workflows that have been verified.

‍After testing, the overall design for the software will come together. Different modules or designs will be integrated into the primary source code through developer efforts, usually by leveraging training environments to detect further errors or defects.

# 15. Pro**fessional issues**

## 15.1. Project Management: activity, schedule (i.e. Gannt or other, showing activities, deadlines), data management, Deliverables

The development of the project consists of the next steps:

1) Literature review

2) Database structure design

3) Flexible dataflow design for SPA system, that includes the information connected to the pandemic or the other important information for PA

4) Research of the databases for predictive analytics on delivery services;

5) Implementation of the Concept Drift(CD) tweaking methods using C++;

6) Recommendation of ML model implementation development;

7) Developing of the recommendations for the testing and evaluating of the designed system, Testing coordination of DB unit and CD units.

The schedule of the proposed research work is summarized in table 2.

|  |  |
| --- | --- |
| Activity | Deadline |
| Literature review | 15.05.22 |
| Database structure design | 31.05.22 |
| Flexible dataflow design for SPA system, that includes the information connected to the pandemic or the other important information for PA | 16.06.22 |
| Research of the databases for predictive analytics on delivery services | 15.06.22 |
| Implementation of the Concept Drift(CD) tweaking methods using C++ | 15.07.22 |
| Recommendation of ML model implementation development | 15.08.22 |
| Developing of the recommendations for the testing and evaluating of the designed system, Testing coordination of DB unit and CD units | 31.09.22 |

Table 2: Schedule of the research work

The reference literature is stored in on the local disks.

Logging files and ML logs will be stored in PC.

The main result of the projects:

- Programming module for CD tweaking;

- Databases links for Predictive Analytics in post COVID-19 times;

- Literature review;

- Recommendation for the creation of SPA;

- Report on the database structure and dataflow for the system;

- Report on the unit and validation tests of the system units.

## 15.2. Risk: risk analysis as informed by current progress. Resolved risks and the success of the mitigation strategy. Changes to the project plan as the results of risks. Future risks

The initial idea fro the development of the project was the design and implementation of the system for predictive analytics on Python. But with the flow of the project, it was discovered that the aim is not achievable for the given period of time. The complexity of the SPA system has led to the shortening of this task to the part of mere design of the SPA and developing only the fixed part of the system components. The consequences of this, were changing and modifying project goals and plans.

Other issues of the project were concerned data mining issues. For example, the risk involved in accessing the data from a remote database can be that the data access rate might be too slow. The risk can be resolved by building a prototype of the data access subsystem, therefore we could lay this issue on the shoulder of clients within the framework of the project.

## 15.3. Professional issues: identification and discussion of relevant legal, social, ethical, and environmental issues in the context of the project.

Near the final stage of the development of the project, data protection issues arise. During collection of the data for SPA system, there were several computer security issues, that were shown off. For example, in the relation tables there is the table, which presence in the DB creates special security issues.

The table Customer contains not only attribute - the number of the customer and email, but also some sensible data – name, address, phone, and may also contain some other type of information, that user and/or legislation may do not admit not only to share, but even to store.

The permission to gather and collect these data can be granted by the user itself, upon the corresponding legislation of, for example, EU. But the storage and usage of them demands careful complying the legislation rules and the recommendations of computer security.

Those data, that are personal, must be stored carefully and securely, with accordance to legislation on this type of data of each particular region.

One of the possible solution to avoid these before mentioned complications is to reduce the amount of such personal information. In fact, some of these data have little or no impact on the PA results. For example, name, phone and email almost certainly have no correlation with the possible sales prediction results. While such data as civil status can have some impact on the model, but it is unlikely to be significant.

So, the reduction of these data could rule out the above-mentioned issues without degrading of the ML model.

However, some data, for example, the data of customer location may be crucial for the delivery service functioning. So, this leads us to the implementation into the SPA some forms of the security mechanism and apply cryptography protection of these kinds of data, as well as the supplement of SW defense for the whole SPA system and DB.

The partial solution to this problem with the legislation is to keep the data of customers locations in encrypted way while instead of the customer name, address and other personal information the designed SPA could store corresponding hashes. As we mentioned before, this way we do not decrease much the accuracy of the SPA performance and decrease the risks of data hijacking.

That approach, of course, do not completely remove the possibility of personal data leaks and other threats of computer era, but significantly mitigate them. To suppress these risks, the standard computer security measures are to look sufficient as far as we can see from this point.

# 16. Conclusion: summary of what was achieved and potential future work

The research on the project “Using AI and machine learning for predictive analytics in estimating the growth in the demand of delivery services post COVID-19” were divided into the following parts

The references in the field of AI creation for predictive analytics were analyzed with their relations for the topic of delivery service prediction, also the new bibliography, emerged in the field for data prediction in COVID-19 impact learning were studied. The method of concept drift detection was selected as the supplement for the System of Predictive Analytics (SPA) design.

The database structure for SPA was designed, its components were described. Also, the new component, that allows to create datflow in SPA to be more flexible dataflow is proposed. That part of SPA let information connected to the pandemic or the other important information for PA be input into SPA.

The list of the open databases for predictive analytics on delivery services, that could be used in the SPA is prepared (listed in Appendix A).

The implementation of the Concept Drift(CD) tweaking methods using C++ was conducted (Appendix B).

Recommendation of ML model implementation development and recommendations for the testing and evaluating of the designed system, was proposed and presented.

Potential future work on this topic could include implementation of other ways of Concept Drift detection and utilization, the tweaking of SPA by using of other methods like Active Learning, etc. And, certainly, the current status of the project can be improved by implementation and testing of the proposed methods in the full scale, for example, on Python TF.

# 17. Bibliography

[*https://www.marketsandmarkets.com/Market-Reports/predictive-analytics-market-1181.html*](https://www.marketsandmarkets.com/Market-Reports/predictive-analytics-market-1181.html)

[*https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ordering-in-the-rapid-evolution-of-food-delivery*](https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ordering-in-the-rapid-evolution-of-food-delivery)

[*https://www.cio.com/article/228901/what-is-predictive-analytics-transforming-data-into-future-insights.html*](https://www.cio.com/article/228901/what-is-predictive-analytics-transforming-data-into-future-insights.html)

*https://www.predictivelayer.com/*

*https://www.predictivetechnologysystems.com/*

*https://www.dymatrix.de/dymatrixblog/automated-predictive-analytics-nbo-uplift-models-scorecards*

*https://www.g2.com/categories/predictive-analytics*

*https://www.techtarget.com/searchbusinessanalytics/tip/6-top-predictive-analytics-tools*

*https://xpanse.ai/*

*https://itsourcecode.com/free-projects/database-design-projects/database-design-for-ordering-and-delivery-management-system/*

*https://www.ncss.com/software/ncss/time-series-and-forecasting-in-ncss/*

*https://www.g2.com/categories/time-series-intelligence*

*http://www.wessa.net/tsa.wasp*

*HARRI PREENJA, MOHAMMAD ALI. 2016. Predicting Time Series Data collected from Software Measurements with Machine Learning Approaches //https://gupea.ub.gu.se/handle/2077/44652  
Hyndman, Rob & Khandakar, Yeasmin. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software. 26. 10.18637/jss.v027.i03.*

*Oleg Ostashchuk. 2017. Time Series Data Prediction and Analysis*

https://github.com/abifet/adwin

[*https://github.com/rsdevigo/pyAdwin*](https://github.com/rsdevigo/pyAdwin)

*https://github.com/Western-OC2-Lab/PWPAE-Concept-Drift-Detection-and-Adaptation*

Kanoun, Karim & Schaar, Mihaela. (2015). Big-Data Streaming Applications Scheduling with Online Learning and Concept Drift Detection. 1547-1550. 10.7873/DATE.2015.0786.

Bifet, Albert & Read, Jesse & Pfahringer, Bernhard & Holmes, Geoff & Žliobaitė, Indrė. 2013. CD-MOA: Change Detection Framework for Massive Online Analysis. 92-103. 10.1007/978-3-642-41398-8\_9.

https://en.wikipedia.org/wiki/Predictive\_analytics

https://itchronicles.com/artificial-intelligence/predictive-analytics-vs-machine-learning-similarities-differences-and-how-they-both-benefit-business/

Krauß, Jonathan & Pacheco, Bruno & Zang, Hanno & Schmitt, Robert. (2020). Automated machine learning for predictive quality in production. Procedia CIRP. 93. 443-448. 10.1016/j.procir.2020.04.039.

Seyedan, Mahya & Mafakheri, Fereshteh. (2020). Predictive Big Data Analytics for Supply Chain Demand Forecasting: Methods, Applications, and Research Opportunities.

Nguyen, Truong & Zhou, Li & Spiegler, Virginia & Ieromonachou, Petros & Lin, Yong. (2017). Big data analytics in supply chain management: A state-of-the-art literature review. Computers & Operations Research. 10.1016/j.cor.2017.07.004.

Tamym, Lahcen & El Ouadghiri, Driss & Benyoucef, Lyes & Nait Sidi Moh, Ahmed. (2020). Big Data for Supply Chain Management in Industry 4.0 Context : A Comprehensive Survey.

William F. Ogilvie, Pavlos Petoumenos, Z. Wang, and H. Leather. 2017. Minimizing the cost of iterative compilation with active learning. 2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO) (2017), 245–256.

Yüksel, Asım & Çankaya, Ş Fuat & Üncü, Ismail. (2017). Design of a Machine Learning Based Predictive Analytics System for Spam Problem. Acta Physica Polonica A. 132. 500-504. 10.12693/APhysPolA.132.500.

*Zhang, Ke & Xu, Jianwu & Min, Martin & Jiang, Guofei & Pelechrinis, Konstantinos & Zhang, Hui. (2016). Automated IT system failure prediction: A deep learning approach. 1291-1300. 10.1109/BigData.2016.7840733.*

*Sheng, J., Amankwah-Amoah, J., Khan, Z., & Wang, X. (2020). COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions.  
British Journal of Management.* [*https://doi.org/10.1111/1467-8551.12441*](https://doi.org/10.1111/1467-8551.12441)

*J. Nidamanuri, A. Rohith, S. Pranjal and H. Venkataraman, "Covid-19 Impact and Implications on Traffic: Smart Predictive Analytics for Mobility Navigation," 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), 2022, pp. 812-817, doi: 10.1109/COMSNETS53615.2022.9668404.*

*Grybauskas, A., Pilinkienė, V. & Stundžienė, A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. J Big Data* ***8,*** *105 (2021). https://doi.org/10.1186/s40537-021-00476-0*

*https://www.iguazio.com/blog/concept-drift-deep-dive-how-to-build-a-drift-aware-ml-system/*

*Kadwe, Yamini & Suryawanshi, Vaishali. 2015. A Review on Concept Drift. IOSR Journal of Computer Engineering. 17. 20-26. 10.9790/0661-17122026.*

*Losing, V., Hammer, B., & Wersing, H. 2018. Incremental on-line learning: A review and comparison of state of the art algorithms. Neurocomputing, 275, 1261-1274.*

Žliobaitė, I., Pechenizkiy, M., & Gama, J. 2016. An Overview of Concept Drift Applications. //DOI:[*10.1007/978-3-319-26989-4\_4*](https://doi.org/10.1007/978-3-319-26989-4_4)

Baier, Lucas & Hofmann, Marcel & Kühl, Niklas & Mohr, Marisa & Satzger, Gerhard. 2020. Handling Concept Drifts in Regression Problems - the Error Intersection Approach.

Baier, Lucas & Reimold, Josua & Kühl, Niklas. (2020). Handling Concept Drift for Predictions in Business Process Mining. 76-83. 10.1109/CBI49978.2020.00016.

Zenisek, Jan & Holzinger, Florian & Affenzeller, Michael. (2019). Machine Learning based Concept Drift Detection for Predictive Maintenance. Computers & Industrial Engineering. 137. 106031. 10.1016/j.cie.2019.106031.

Straat, Michiel & Abadi, F. & Kan, Z. & Göpfert, Christina & Hammer, Barbara & Biehl, Michael. (2022). Supervised learning in the presence of concept drift: a modelling framework. Neural Computing and Applications. 34. 1-18. 10.1007/s00521-021-06035-1.

Weingarten, Jennifer & Spinler, Stefan. (2020). Shortening Delivery Times by Predicting Customers' Online Purchases: a Case Study in the Fashion Industry. 10.24251/HICSS.2020.159.

Bahrami Bidoni, Zeynab & Montreuil, Benoit. (2021). Predictive Demand Modeling for New Services in Hyperconnected Urban Parcel Logistics.

# 18. Appendices

## 18.1. A. Links to the data sources:

https://www.businessofapps.com/data/food-delivery-app-market/

https://www.statista.com/outlook/dmo/eservices/online-food-delivery/united-states

https://www.ibisworld.com/global/market-research-reports/global-courier-delivery-services-industry/

https://microservices.io/refactoring/extract-service-delivery-service-step-2.html

https://www.sciencedirect.com/science/article/pii/S2352340919303609 : https://data.mendeley.com/datasets/m9z9hw4nsc/1/files/c8ca39ff-83f6-4bda-8311-5e044824b91c

https://www.pharmdata.co.uk/pandemic\_delivery\_service.php

https://data.world/datasets/service-delivery

<https://www.ofwat.gov.uk/publication/service-and-delivery-report-2020-21-data/>

18.2 B. Link to the programming code

https://github.com/mecchmatProjects/AI\_project