**Open Ended Group Task Project (Python Programming)**

1. **Predicting user’s next location**

Predicting the user’s most probable next location (next summer vacation, holidays, etc.) becomes an important requirement to make decisions for future services. These services include healthcare applications, network management, travel management, and so on. Working on this AI project model will help you to understand the Lempel Ziv (LZ) algorithm, Markov Model (MM), Neural Networks (NNs), Bayesian Networks, and Association rules.

1. **Detecting YouTube comment spam**

The popularity of YouTube not only attracted genuine viewers but spammers as well. As a result, there is an increase in unwanted spam videos and comments. Here comes the importance of an AI-based YouTube spam comment detection model. In this AI project, you will be focusing on text and words and classify internet comments as spam or not spam. The spam detection model can be accomplished by using bag-of-words and random forest techniques. You can also predict positive and negative reviews with the Word2Vec approach and the k-nearest neighbour classifier in addition to spam detection.

1. **Identifying the genre of a song**

In this AI project, you will be identifying the genre of a song using an artificial neural network. You will be using Librosa (python library) to extract features from the song and Mel-frequency cepstral coefficients (MFCC) to detect the music genre.

1. **Shock front classification**

This AI project detects shock fronts in computational fluid mechanics (CFD) simulations. The presence of shock results in additional complexities in fluid mechanics, and hence, it is necessary to detect and handle shock fronts to deal with fluid mechanics problems. Here, you will be using supervised algorithms for classification such as classification trees (RPART), linear discriminant analysis (LDA), naive Bayes (NB).

1. **Predicting bird species**

Birds are ecological indicators, and they respond quickly to environmental changes. Hence, it is important to classify birds to understand the problems in ecology. Domain experts can classify birds manually, but manual classification has become a tedious and time-consuming process due to the tremendous increase in amounts of data. Here comes the importance of artificial intelligence-based classification. This classification-based **AI project** can be approached in two ways. If you are a beginner, you can use a random forest to predict bird species. If you are looking for an intermediate level, you can use a convolution neural network.

1. **Identifying handwritten mathematical symbols**

In this **Artificial Intelligence project**, you will be using a convolution neural network (CNN) to detect handwritten mathematical symbols. The HASYv2 dataset is the input to the neural network; it contains 168,000 images from 369 different classes.

1. **Scotch whisky classification**

Scotch whiskey is famous for its distinct flavours. In this **simple AI project**, you will classify scotch whiskeys based on their flavour characteristics. Here, we will use datasets of scotch whiskeys from several distilleries and cluster it based on the flavours. Please check the below link for datasets.

1. **Investigate Enron**

Enron is one of the largest energy companies in America that collapsed overnight. This AI project investigates Enron fraud activities with the help of the emails sent by their former senior executives. It has 500 thousand emails from their former employees. Check the below link for the Enron database.

1. **Automatic system for detecting trends in fashion**

Coolhunter has gained significant importance in the fashion world. They take advantage of social media platforms to understand new trends in fashion. But, due to irrelevant information, it becomes a challenging task to predict fashion trends. This **AI-based project** filters relevant information from the irrelevant one.

1. **Web pattern navigation profiling for online marketing campaigns**

Each time when users search for information on the internet, they leave an invisible blueprint of their preferences. These preferences are recorded based on their browsing behavior in a specific sequence of domains. Here, segments of user groups are created based on their browsing habit or social media opinions.

In this **Artificial Intelligence project**, you will learn a new perspective in collecting user preferences. Here, different navigation profiles are extracted based on the consecutive sequence of domain visiting order and the route followed within a certain socio-demographic profile. Here, you will define an algorithm to extract frequent contiguous sequences and also use Bonferroni and. FDR corrections to retrieve socio-demographic characteristics

1. **Food attribute classification using a multi-scale convolutional network**

This **AI-based project** classifies the diverse array of food based on cuisine and its flavors. Here, we create a deep learning model based on a multi-scale convolutional network. The food attribute dataset – Yummly48k – is taken from the website [Yummly](https://www.yummly.com/). In addition to the multi-scale convolutional network, it also uses Negative Log-Likelihood (NLL) for the model creation.

1. **Diet4You**

Maintaining a healthy lifestyle plays a key role in preventing the cause of chronic diseases. The right amount of nutrition is a must to maintain a healthy lifestyle, but due to a poor diet plan, a major chunk of the population is suffering from undernutrition.

Diet4You is an intelligent decision support system (IDSS) that uses different techniques to tailor a personalized menu planner. It not only considers the prescription provided by the nutritionist but also considers various other factors such as the nutritional guidelines that are to be followed, the person’s characteristics, health status, habits, food preferences, allergies, and so on. This **AI project** combines advanced techniques such as Knowledge Engineering, Case-Based Reasoning (CBR), and Data Analysis. Diet4You consists of two main modules: NPG module – tailor a nutrition plan for a specific person – and PMP module – a nutrition plan for a specific period.

1. **Unlocking phone using faceID**

It is one **artificial intelligence project** that uses face biometrics to unlock a phone. Using deep learning, the AI application can extract image features. It mainly uses two types of neural networks: Convolution neural networks and Deep autoencoders network. And it is a four-step process. They are- face detection, face alignment, face extraction, and face recognition.

1. **Gmail’s Smart Reply**

Gmail’s smart reply uses a machine-learning algorithm to suggest replies to email. It is based on a novel thinking hierarchy where each hierarchical model can learn, remember, and recognize a sequential pattern. While responding, it considers whether it is a positive gesture or a negative gesture. It uses technologies such as long-short-term-memory (LSTM) recurrent neural networks and semantics.

1. **Uber’s ride-sharing app**

Uber is one of the biggest cab service providers. It uses artificial intelligence to make predictions about market demand, provide better customer experience, find the best route for drivers, detect fraud, etc. Uber AI uses various techniques such as forecasting, demand modeling, dynamic pricing, and so on.

1. **Forecasting earthquake aftershock locations with AI-assisted science**

Earthquakes cause massive destruction in the entire world. It initially occurs as a mainshock and is then followed by a set of aftershocks. The timing and size of aftershocks can be identified using empirical laws, but forecasting the locations remains a challenging part. This Google **AI project** applies deep learning to identify the location where the aftershock might occur. The project uses the information of 118 major earthquakes reported around the world. Here, it uses a neural network to analyze the static stress change of mainshock and aftershock locations.

1. **XTREME**

XTREME is a pre-trained multilingual model that includes 40 typologically diverse languages and comprises tasks that need reasoning. XTREME is one of Google’s **ai projects** that use natural language processing – sentence classification, structured prediction, sentence retrieval, and question answering.

1. **MEENA**

MEENA is a chatbot that handles a wide variety of conversational topics and humanizes computer interaction, improves foreign language practice, and so on. It is an end-to-end trained neural conversational model with a single Evolved Transformer encoder block and 13 Evolved Transformer decoder blocks. These blocks help them to respond sensibly by minimizing the perplexity and the uncertainty in prediction.

1. **COVID19 Chatbots**

As a beginner, you can start by creating chatbots. You’re recommended to start by creating a simple version. Several chatbots are available on almost every company website. You can check them out and identify the basic structure and can build your very own chatbots with a similar kind of structure. Once you have completed, try creating one, which is more tedious. Also, try working on different niche chatbots. Artificial intelligence gives you the freedom to open up your wings and helps you put your ideas into action.

1. **E-commerce Customer Recommendation System**

Have you ever wondered how an e-commerce platform comes with suggestions that you need? You might have seen ads popping up on your social media accounts related to the searches you have made.  This all is done through the customer recommendation system.

You can develop your very own customer recommendation system. To start with, you can use your browsing history. Both behavioral data and implicit data is required. There is considerable scope for recommendation system developers in the tech world. If you crack this and learn how to upskill the way to come up with a better version of the E-commerce customer recommendation system, then there is a better chance for you to end up having a career in a renowned company.

1. **Stock Market Predictor**

If you are good at numbers, then this one is pretty much for you. Have you come across a stock market predictor before? If not, do check them out. They are known for their accuracy based on mathematical assumptions and present circumstances. You can even get to know whether your predictor works or not within no time by keeping stock prediction cycles small. There is a massive value & demand for such systems. This project will help you to make a career in finance if mathematics is your cup of tea.

1. **Sentiment Predictor**

Consumer behavior is something that every online business is targeting. To know how a consumer reacts to a post will better the chances of them buying a product. Artificial intelligence can be used to identify the sentiments of consumers. You can come up with a sentiment predictor that can help analyze in what state of mind the consumer is.

This project will help you create an impression among Tier-I companies. Almost every online business has started implementing this. Now is the time to showcase your skill by implementing this project and grabbing hold of the opportunity.

1. **Classification using AI**

If you have an idea of coming up with a facial monitoring system or retinal identification system any time in the future, you can start small by coming with a flower classification system. You may be aware that Iris flowers are of different shapes and sizes. Based on this shape, size, and color, they can be classified as Versicolor, Virginica, and Setosa.

Using AI, come up with a system that can classify these flowers easily. As we have said before, if you’re looking to build a facial or retinal recognition system, this is where you can start. There is considerable scope for such AI developers. This project will stand as a testimonial on how equipped you are.

1. **Activity Recognition system**

You would have already come across stuff like smart-watches, bands, etc. Did you know that they use artificial intelligence to attain accuracy in determining your heartbeat and the number of calories you have lost based on the walking you have done?

An activity recognition system would be a good project. Analyze how the product works and build your very own smart activity recognition system with a similar kind of structure. You’ll get to learn what algorithms are used in such applications and let us tell you that you’ll enjoy the process. Initially, start with a simple algorithm; once done, do go for a complex one. There is considerable scope for such devices, so do give it a try.

1. **Wine Quality Analyzer**

Using a particular set of data, you can determine the quality of the wine. You might be aware of the fact that the older the wine, the better it becomes. Several considerations are to be taken into account before validating the quality of the wine. The pH content, the percentage of alcohol, the amount of acidity are some of the few criteria to be taken into consideration.

Using artificial intelligence, you can test these factors and conclude which one is the best wine. The same thing is implemented for testing the fertility of the soil by architectures using AI. You can initially start with wine to get an unobstructed exposure to how the algorithm works. You’ll find that there are more than 4000 odd sets of data that you have to consider. This project will surely hone your AI skills.

1. **Object Detector**

Have you heard of the term deep neural network? Neural networks are used by top-notch companies to carry out a different set of operations like face recognition, translation, etc. The object detector is somewhat similar to the flower classification system yet a little complex. It helps you detect a particular object with the help of artificial intelligence. If you are looking to make more of an impact with your project, then this smart object detection is the one for you.

Try developing a basic algorithm structure to start with. Once you have completed, try creating one, which is more tedious. Object detection is the same method that artificially intelligent robots implement. Virtual Reality and 3D augmentation also work on the same principle. Make the best use of this opportunity to learn and to upgrade such a skill.

1. **Recommender Engine with longtail**

Had it ever got you thinking while watching a video or a show on YouTube or Netflix, how similar videos pop up based on your preferences. How about creating an engine that can do the same task? Based on the behavioral and implicit activity, the algorithm can decide on your preferences and show similar content. Instead of binge-watching, you could build your very own recommender engine. To start with, you can use your browsing history. Both behavioral data, and implicit data, is required. It definitely will turn heads around.

When it comes to job opportunities, recommender engine developers are in demand. With several training institutes going digital, everyone is looking for an AI developer to come up with a recommendation system on their websites.

1. **Sales Predictor**

Supermarkets are a place where there is a surplus number of products. How they manage to keep track of the sales of every product is beyond our imagination. That is where a sales predictor comes in handy. It helps you monitor stocks that come in daily and products that are sold out.

Sales Predictor will turn out to be one heck of a project. You have to come up with an algorithm on how many products are being sold daily and predict the sales of that product on a weekly or monthly basis.

To start with, you can find vast sets of Big Mart sales on the internet. Try working on those data and create a simple algorithm to predict the sales. You can also discover how many sales of a particular product happened in a year in their sales report released in 2013.

The sales predictor is definitely a project that would create a better first impression. If you find it simple, try a sophisticated algorithm and check whether it works. You’ll get to learn what algorithms are used in such applications and let me tell you’ll enjoy the process. Almost every online grocery store has started implementing this. Do give it a try.

1. **Sentiment analysis for depression based on social media post**

This topic is so sensitive to be considered nowadays and in urgent need to do something about it. There are more than 264 million individuals worldwide who are suffering from depression. Depression is the main cause of disability worldwide and is a significant supporter of the overall global burden of disease and nearly 800,000 individuals consistently bite the dust because of suicide every year. Suicide is the second driving reason for death in 15–29-year-olds. Treatment for depression is often delayed, imprecise, and/or missed entirely.

Internet-based life gives the main edge chance to change early melancholy mediation services, especially in youthful grown-ups. Consistently, roughly 6,000 Tweets are tweeted on Twitter, which relates to more than 350,000 tweets sent for each moment, 500 million tweets for every day, and around 200 billion tweets for each year.

As indicated by the Pew Research Center, 72% of the public uses some sort of internet-based life. Datasets released from social networks are important to numerous fields, for example, human science and brain research. But the supports from a specialized point of view are a long way from enough, and explicit methodologies are desperately out of luck.

By analyzing linguistic markers in social media posts, it’s possible to create a deep learning model that can give an individual insight into his or her mental health far earlier than traditional approaches.

1. **Sports match video to text summarization using neural network**

So, this project idea is basically based on getting precise summary out of Sports match videos. There are sports websites that tell about highlights of the match. Various models have been proposed for the task of extractive text summarization, but neural networks do the best job. As a rule, Summarization alludes to introducing information in a brief structure, concentrating on parts that convey facts and information, while safeguarding the importance.

Automatically creating an outline of a game video gives rise to the challenge of distinguishing fascinating minutes, or highlights, of a game.

So, one can achieve that using some deep learning techniques like 3D-CNN (three-dimensional convolutional networks), RNN (Recurrent neural network), LSTM (Long short term memory networks) and also through Machine learning algorithms by dividing the video into different sections and then applying SVM (Support vector machines), NN(Neural Networks), k-means algorithm.

1. **Handwritten equation solver using CNN**

Among all the issues, handwritten mathematical expression recognition is one of the confounding issues in the region of computer vision research. You can train Handwritten equation solver by handwritten digits and mathematical symbols using Convolutional Neural Network (CNN) with some image processing techniques. Developing such a system requires training our machines with data, making it proficient to learn and make the required prediction.

1. **Business meeting summary generation using NLP**

Ever got stuck in a situation, where everyone wants to see a summary not full reports. Well, I face it during my school and college days where we spend a lot of time preparing a whole report, but the teacher only has time to read the summary.

Summarization has risen as an inexorably helpful way to tackle the issue of data over-burden. Extracting information from conversations can be of very good commercial and educational value. This can be done by feature capture of the statistical, linguistic, and sentimental aspects with the dialogue structure of the conversation.

Manually changing the report to a summed-up form is too time taking, isn’t that so? But one can rely on Natural Language Processing (NLP) techniques to achieve that.

Text summarization using deep learning can understand the context of the entire text. Isn’t it a dream come true for all of us who need to come up with a quick summary of a document!!

1. **Facial recognition to detect mood and suggest songs accordingly**

The human face is an important part of an individual’s body, and it particularly plays a significant role in knowing a person’s state of mind. This eliminates the dreary and tedious task of manually isolating or grouping songs into various records and helps in generating an appropriate playlist based on an individual’s emotional features.

People tend to listen to music based on their mood and interests. One can create an application to suggest songs for users based on their mood by capturing facial expressions.

Computer vision is an interdisciplinary field that helps convey a high-level understanding of digital images or videos to computers. computer vision components can be used to determine the user’s emotion through facial expressions.

1. **Finding out habitable exo-planet from images captured by space vehicles like Kepler**

In the most recent decade, over a million stars were monitored to identify transiting planets. Manual interpretation of potential exoplanet candidates is labor-intensive and subject to human mistake, the consequences of which are hard to evaluate. Convolutional neural networks are fit for identifying Earth-like exoplanets in noisy time-series data with more prominent precision than a least-squares strategy.

1. **Image regeneration for old, damaged reel picture**

I know, how time- consuming and painful it is to get back your old, damaged photo in the original form as it was earlier. So, this can be done using deep learning by finding all the image defects (fractures, scuffs, holes), and using Inpainting algorithms, one can easily discover the defects based on the pixel values around them to restore and colorize the old photos.

1. **Music generation using deep learning**

Music is an assortment of tones of various frequencies. So, the Automatic Music Generation is a process of composing a short piece of music with the least human mediation. Recently, Deep Learning engineering has become the cutting edge for programmed Music Generation.

1. **Marketing analytics:**AI systems learn from, analyze, and measure marketing efforts. These solutions track media activity and provide insights into PR efforts to highlight what is driving engagement, traffic, and revenue. As a result, companies can provide better and more accurate marketing services to their customers.

Besides PR efforts, AI-powered marketing analytics can lead companies to identify their customer groups more accurately. By discovering their loyal customers, companies can develop accurate marketing strategies and also retarget customers who have expressed interest in products or services before. Feel free to read more about marketing analytics with AI from this article.

1. **Personalized Marketing:**The more companies understand their customers, the better they serve them. AI can assist companies in this task and support them in giving personalized experiences for customers.

As an example, suppose you visited an online store and looked at a product but didn’t buy it. Afterward, you see that exact product in digital ads. More than that, companies can send personalized emails or special offers and recommend new products that go along with customers’ tastes.

1. **Context-Aware Marketing**: You can leverage machine vision and Natural Language Processing to understand the context where your ads will be served. With context-aware advertising, you can protect your brand and increase marketing efficiency by ensuring your message fits its context, making static images on the web come alive with your messages.

To learn more about AI use cases in marketing, you can check out our complete guide on the topic.

1. **Sales Forecasting:**AI allows automatic and accurate sales forecasts based on all customer contacts and previous sales outcomes. Automatically forecast sales accurately based on all customer contacts and previous sales outcomes. Give your sales personnel more sales time while increasing forecast accuracy. Hewlett Packard Enterprise indicates that it has experienced a 5x increase in forecast simplicity, speed, and accuracy with Clari’s sales forecasting tools.
2. **Lead generation:**Use a comprehensive data profile of your visitors to identify which companies your sales reps need to connect. Generate leads for your sales reps leveraging databases and social networks
3. **Sales Data Input Automation:**Data from various sources will be effortlessly and intelligently copied into your CRM. Automatically sync calendar, address book, emails, phone calls, and messages of your salesforce to your CRM system. Enjoy better sales visibility and analytics while giving your sales personnel more sales time.
4. **Predictive sales/lead scoring:**Use Artificial Intelligence to enable predictive sales. Score leads to prioritize sales rep actions based on lead scores and contact factors. Sales forecasting is automated with increased accuracy thanks to systems’ granular access to lead scores and sales rep performance. For scoring leads, these systems leverage anonymized transaction data from their customers, sales data of this specific customer. For assessing contact factors, these systems leverage anonymized data and analyze all customer contacts such as email and calls.
5. **Sales Rep Chat/ Email Bot:**Chatbots are ideal to answer first customer questions. If the chatbot decides that it cannot adequately serve the customer, it can pass those customers to human agents. Let 24/7 functioning, intelligent, self-improving bots handle making initial contacts to leads. High value, responsive leads will be called by live agents, increasing sales effectiveness.
6. **Sales Rep Response Suggestions:**AI will suggest responses during live conversations or written messages with leads. Bots will listen in on agents’ calls suggesting best practice answers to improve sales effectiveness
7. **Sales Rep Next Action Suggestions**: Your sales reps’ actions and leads will be analyzed to suggest the next best action. This situation wise solution will help your representatives to find the right way to deal with the issue. Historical data and profile of the agent will help you to achieve higher results. All are leading to more customer satisfaction.
8. **Sales Content Personalization and Analytics:**Preferences and browsing behavior of high priority leads are analyzed to match them with the right content, aimed to answer their most important questions. Personalize your sales content and analyze its effectiveness allowing continuous improvement.
9. **Retail Sales Bot**: Use bots on your retail floor to answer customer’s questions and promote products. Engage with the right customer by analyzing the profile. Computer vision will help you to provide the right action depending on the characteristics and mimics of the customer.
10. **Meeting Setup Automation (Digital Assistant):**Leave a digital assistant to set up meetings freeing your sales rep’s time. Decide on the targets to prioritize and keep your KPI’s high.
11. **Prescriptive Sales**: Most sales processes exist in the mind of your sales reps. Sales reps interact with customers based on their different habits and observations. Prescriptive sales systems prescribe the content, interaction channel, frequency, price based on data on similar customers**.**
12. **Customer Sales Contact Analytics:**Analyze all customer contacts, including phone calls or emails, to understand what behaviors and actions drive sales. Advanced analytics on all sales call data to uncover insights to increase sales effectiveness
13. **Sales Call Analytics**: Advanced analytics on call data to uncover insights to increase sales effectiveness. See how well your conversation flow performs. Integrating data on calls will help you to identify the performance of each component in your sales funnels.
14. **Sales attribution:**Leverage big data to attribute sales to marketing and sales efforts accurately. See which step of your sales funnel performs better. Pinpoint the low performing part by the insights provided by analysis.
15. **Sales Compensation:** Determine the right compensation levels for your sales personnel. Decide on the right incentive mechanism for the sales representatives. By using the sales data, provide objective measures, and continuously increase your sales representatives’ performance.
16. **Analytics Platform**: Empower your employees with unified data and tools to run advanced analyses. Quickly identify problems and provide meaningful insights.
17. **Analytics Services**: Satisfy your custom analytics needs with these e2e solution providers. Vendors are there to help you with your business objectives by providing turnkey solutions.
18. **Automated Machine Learning (autoML)**: Machines helping data scientists optimize machine learning models. With the rise of data and analytics capabilities, automation is needed in data science. AutoML automates time consuming machine learning tasks, enabling companies to deploy models and automate processes faster.
19. **Geo-Analytics Platform**: Enables analysis of granular satellite imagery for predictions. Leverage spatial data for your business goals. Capture the changes in any landscape on the fly.
20. **Conversational Analytics**: Use conversational interfaces to analyze your business data. Natural Language Processing is there to help you with voice data and more. Automated analysis of reviews and suggestions.
21. **Real-Time Analytics**: Real-Time Analytics for your time-sensitive decisions. Act timely and keep your KPI’s intact. Use machine learning to explore unstructured data without any disruptions.
22. **Image Recognition and Visual Analytics**: Analyze visual data with advanced image and video recognition systems. Meaningful insights can be derived from the data piles of images and videos.
23. **E-Commerce Analytics**: Specialized analytics systems designed to deal with the explosion of e-commerce data. Optimize your funnel and customer traffic to maximize your profits.
24. **Social Listening & Ticketing**: Leverage Natural Language Processing and machine vision to identify customers to contact and respond to them automatically or assign them to relevant agents, increasing customer satisfaction. Use the data available in social networks to uncover whom to sell and what to sell.
25. **Intelligent Call Routing**: Route calls to most capable agents available. Intelligent routing systems incorporate data from all customer interactions optimizing customer satisfaction. Based on the customer profile and your agent’s performance, make it possible to provide the right service with the right agent. Reach superior net promoter scores. Feel free to read our article about intelligent call routing if you want to learn more.
26. **Call Classification**: Leverage Natural Language Processing to understand what customer is trying to achieve enabling your agents to focus on higher value-added activities. Before channeling the call, detect the nature of your customers’ needs and let the right department handle the problem. Enhance efficiency with higher satisfaction rates.
27. **Voice Authentication**: Authenticate customers without passwords leveraging biometry to improve customer satisfaction and reduce issues related to forgotten passwords. Their unique voice id will be their most secure key for accessing confidential information. Instead of the last four digits of SSN, customers will gain access by using their voice.
28. **Call Intent Discovery**: Leverage Natural Language Processing and machine learning to estimate and manage customer’s intent (e.g., churn) to improve customer satisfaction and business metrics. Sentiment analysis through the customer’s voice level and pitch. Detect the micro-emotions that drive the decision-making process.
29. **Customer Service Response Suggestions**: Bots will listen in on agents’ calls suggesting best practice answers to improve customer satisfaction and standardize customer experience. Increase upsells and cross-sells by giving the right suggestion. Responses will be standardized, and the best possible approach will serve the benefit of the customer.
30. **Chatbot**: Chatbots can understand more complicated queries as AI algorithms improve. Thus, businesses understand their customers better since chatbots collect information from customers while interacting with them and spot their weaknesses. There are other benefits like 24/7 availability and reduced costs, as bots can handle more tasks as they learn more. All these benefits significantly improve the customer satisfaction of businesses.

The automotive industry is one of the areas that use chatbots. While a significant portion of leads to car dealers come from online channels, high conversion rates are vital for these companies. For example, Kia observes three times more conversions through its chatbot Kian, compared to its website. Kian’s availability to answer complex questions is a dominant factor for achieving high conversion rates.

1. **Customer Service Chatbot (Self – Service Solution)**: Build your own 24/7 functioning, intelligent, self-improving chatbots to handle most queries and transfer customers to live agents when needed. Reduce customer service costs and increase customer satisfaction. Reduce the traffic on your existing customer representatives and make them focus on the more specific needs of your customers.
2. **Call Analytics**: Advanced analytics on call data to uncover insights to improve customer satisfaction and increase efficiency. Find patterns and optimize your results. Analyze customer reviews through voice data and pinpoint, where there is room for improvement. Sestek indicates that ING Bank observed a 15% increase in sales quality score and a 3% decrease in overall silence rates after they integrated AI into their call systems.
3. **Survey & Review Analytics**: Leverage Natural Language Processing to analyze text fields in surveys and reviews to uncover insights to improve customer satisfaction and increase efficiency. Automate the process by mapping the right keywords with the right scores. Make it possible to lower the time for generating reports.

Protobrand states that they used to do review analytics manually through the hand-coding of the data, but now it automates much of the analytical work with Gavagai. This helps the company to collect larger quantitative volumes of qualitative data and still complete the analytical work in a timely and efficient manner. You can read more about survey analytics from our related article.

1. **Customer Contact Analytics**: Advanced analytics on all customer contact data to uncover insights to improve customer satisfaction and increase efficiency. Utilize Natural Language Processing for higher customer satisfaction rates
2. **Chatbot Analytics**: Analyze how customers are interacting with your chatbot. See the overall performance of your chatbot. Pinpoint its shortcomings and improve your chatbot. Detect the overall satisfaction rate of your customer with the chatbot.
3. **Chatbot testing**: Semi-automated and automated testing frameworks facilitate bot testing. See the performance of your chatbot before deploying. Save your business from catastrophic chatbot failures. Detect the shortcomings of your conversational flow.
4. **Data Visualization**: Visualize your data for better analytics and decision making. Let the dashboards speak. Convey your message more easily and more esthetically. Our website has more information about data visualization if you are interested.
5. **Data Management & Monitoring**: Keep your data high quality for advanced analytics. Adjust the quality by filtering the incoming data. Save time by automating manual and repetitive tasks.
6. **Data Integration**: Combine your data from different sources into meaningful and valuable information. Data traffic depends on multiple platforms. Therefore, managing this huge traffic and structuring the data into a meaningful format will be important. Keep your data lake available for further analysis. You can also take a look at our in-depth guide about data integration if you are interested.
7. **Data Preparation Platform**: Prepare your data from raw formats with data quality problems to a clean, ready to analyze format. Use extract, transform, and load (ETL) platforms to fine-tune your data before placing it into a data warehouse.
8. **Data Cleaning & Validation Platform**: Avoid garbage in, garbage out by ensuring the quality of your data with appropriate data cleaning processes and tools. Automate the validation process by using external data sources. Regular maintenance cleaning can be scheduled, and the quality of the data can be increased. Feel free to read our in-depth guide about data cleaning if you want to have more information.
9. **Data Transformation**: Transform your data to prepare it for advanced analytics. If it is unstructured, adjust it for the required format.
10. **AppDev:** App development platforms for your custom projects. Your in-house development team can create original projects for your specific business needs. These platforms will help your team with the necessary tools.
11. **Data Labeling**: Unless you use unsupervised learning systems, you need high quality labeled data. Label your data to train your supervised learning systems. Human-in-the-loop systems auto label your data and crowdsource labeling data points that cannot be auto-labeled with confidence.
12. **Synthetic Data:** Computers can artificially create synthetic data to perform certain operations. The synthetic data is usually used to test new products and tools, validate models, and satisfy AI needs. Companies can simulate not yet encountered conditions and take precautions accordingly with the help of synthetic data. They also overcome the privacy limitations as it doesn’t expose any real data. Thus, synthetic data is a smart AI solution for companies to simulate future events and consider future possibilities. You can have more information on synthetic data from our related article
13. **Hiring:**Hiring is a prediction game: Which candidate, starting at a specific position, will contribute more to the company? Machine’s better data processing capabilities augment HR employees in various parts of hiring such as finding qualified candidates, interviewing them with bots to understand their fit or evaluating their assessment results to decide if they should receive an offer
14. **Performance Management**: Manage your employees’ performance effectively and fairly without hurting their motivation. Follow their KPI’s on your dashboard and provide real-time feedback. This would increase employee satisfaction and lower your organization’s employee turnover. Actualize your employee’s maximum professional potential with the right tools.
15. **HR Retention Management**: Predict which employees are likely to churn and improve their job satisfaction to retain them. Detect the underlying reasons for their motive for seeking new opportunities. By keeping them at your organization, lower your human capital loss.
16. **HR Analytics**: HR analytics services are like the voice of employee analysis. See your people analytics and make better people decisions. Gain actionable insights and impactful suggestions for higher employee satisfaction.
17. **Digital Assistant**: Digital assistants are mature enough to replace real assistants in email communication. Include them in your emails to schedule meetings. They have already scheduled hundreds of thousands of meetings. Use the power of artificial intelligence in your day-to-day activities. Your own on-demand powerful AI-backed assistant is helping you 24/7.
18. **Employee Monitoring**: Monitor your employees for better productivity measurement. Provide objective metrics to see how well they function. Forecast their overall performance with the availability of massive amounts of data.
19. **Building Management**: Sensors and advanced analytics improve building management. Integrate IoT systems in your building for lower energy consumption and many more. Increase the available data by implementing the right data collection tools for effective building management.
20. **Analytics & Predictive Intelligence for Security**: In 2014, Kaspersky Lab said it had detected 325,000 new malware files every day. Analyze data feeds about the broad cyber activity as well as behavioral data inside an organization’s network to come up with actionable insights to help analysts predict and thwart impending attacks. Integrate external data sources the watch out for global cyber threats and act timely. Keep your tech infrastructure intact or minimize losses.
21. **Knowledge Management**: Enterprise knowledge management enables effective and effortless storage and retrieval of enterprise data, ensuring organizational memory. Increased collaborative work by ensuring the right individuals works with the right data. Seamless organizational integration through knowledge management platforms.
22. **Natural Language Processing Library/ SDK/ API**: Leverage Natural Language Processing libraries/SDKs/APIs to quickly and cost-effectively build your custom NLP powered systems or to add NLP capabilities to your existing systems. An in-house team will gain experience and knowledge regarding the tools. Increased development and deployment capabilities for your enterprise.
23. **Image Recognition Library/ SDK/ API:**Leverage image recognition libraries/SDKs/APIs to build your custom image processing systems quickly and cost-effectively or to add image processing capabilities to your existing systems.
24. **Secure Communications**: Protect employee communications like emails or phone conversations with advanced multilayered cryptography & ephemerality. Keep your industry secrets safe from corporate espionage.
25. **Deception Security**: Deploy decoy-assets in a network as bait for attackers to identify, track, and disrupt security threats such as advanced automated malware attacks before they inflict damage. Keep your data and traffic safe by keeping them engaged in decoys. Enhance your cybersecurity capabilities against various forms of cyber attacks
26. **Autonomous Cybersecurity Systems**: Utilize learning systems to respond to security threats efficiently and instantaneously, often augmenting the work of security analysts. Lower your risk of human errors by providing greater autonomy for your cybersecurity. AI-backed systems can check compliance with standards.
27. **Smart Security Systems**: AI-powered autonomous security systems. Functioning 24/7 for achieving maximum protection. Computer vision for detecting even the tiniest anomalies in your environment. Automate emergency response procedures by instant notification capabilities.
28. **Machine Learning Library/ SDK/ API**: Leverage machine learning libraries/SDKs/APIs to quickly and cost-effectively build your custom learning systems or to add learning capabilities to your existing systems.
29. **AI Developer**: Develop your custom AI solutions with companies experienced in AI development. Create turnkey projects and deploy them to the specific business function. Best for companies with limited in-house capabilities for artificial intelligence.
30. **Deep Learning Library/ SDK/ API**: Leverage deep learning libraries/SDKs/APIs to quickly and cost-effectively build your custom learning systems or to add learning capabilities to your existing systems.
31. **Developer Assistance**: Assist your developers using AI to help them intelligently access the coding knowledge on the web and learn from suggested code samples. See the best practices for specific development tasks and formulate your custom solution. Real-time feedback provided by the huge history of developer mistakes and best practices.
32. **AI Consultancy**: Provides consultancy services to support your in-house AI development, including machine learning and data science projects. See which units can benefit most from AI deployment. Optimize your artificial intelligence spending for the best results from the insight provided by a consultant.
33. **Robotic Process Automation (RPA)**: Digitize your processes in weeks without replacing legacy systems, which can take years. Bots can operate on legacy systems learning from your personnel’s instructions and actions. Increase your efficiency and profitability ratios. Increase speed and precision, and many more.

In a McKinsey report, RPA becomes a promising new development in business automation that offers a potential ROI of 30–200 percent in the first year. To learn more, feel free to read our article about RPA.

1. **Robotic Process Automation (RPA) Implementation**: Implementing RPA solutions requires effort. Suitable processes need to be identified. If a rules-based robot will be used, the robot needs to be programmed. Employees’ questions need to be answered. That is why most companies get some level of external help. Generally, outsourcing companies, consultants, and IT integrators are happy to provide temporary labor to undertake this effort. Today, companies like Argos Labs offer low-code RPA solutions to provide easy RPA implementation.
2. **Process Mining**: Leverage AI algorithms to mine your processes and understand your actual processes in detail. Process mining can provide fastest time to insights about your as-is processes as demonstrated in case studies.
3. **Predictive Maintenance**: Predictively maintain your robots and other machinery to minimize disruptions to operations. Implement big data analytics to estimate the factors that are likely to impact your future cash flow. Optimize PP&E spending by gaining insight regarding the possible factors.
4. **Manufacturing Analytics**: Also called industrial analytics systems, these systems allow you to analyze your manufacturing process from production to logistics to save time, reduce cost, and increase efficiency. Keep your industry effectiveness at optimal levels.
5. **Inventory & Supply Chain Optimization**: Leverage machine learning to take your inventory& supply chain optimization to the next level. See the possible scenarios in different customer demands. Reduce your stock, keeping spending, and maximize your inventory turnover ratios. Increase your impact factor in the value chain.
6. **Robotics**: Factory floors are changing with programmable collaborative bots that can work next to employees to take over more repetitive tasks. Automate physical processes such as manufacturing or logistics with the help of advanced robotics. Increased your connected systems by centralizing the whole manufacturing process. Lower your exposures to human errors.
7. **Collaborative Robot**: Cobots provide a flexible method for automation. Cobots are flexible robots that learn by mimicking human workers’ behavior. Smart engineering systems for solutions still requiring human oversight.
8. **Cashierless Checkout**: Self-checkout systems have many names. They are called cashierless, cashier-free, or automated checkout systems. They allow retail companies to serve customers in their physical stores without the need for cashiers. Technologies that allowed users to scan and pay for their products have been used for almost a decade now, and those systems did not require great advances in AI. However, these days we are witnessing systems powered by advanced sensors and AI to identify purchased merchandise and charge customers automatically.
9. **Invoicing:** Invoicing is a highly repetitive process that many companies perform manually. This causes human errors in invoicing and high costs in terms of time, especially when a high volume of documents needs to be processed. Thus, companies can handle these repetitive tasks with AI, automate invoicing procedures, and save significant time while reducing invoicing errors. The company avoids re-invoicing costs with AI tools, as well. For example, Hypatos also indicates that automated invoice capture can lead up to 90% cost saving in your invoice-related processes.

Elekta has reduced its costs and increased its number of processed invoices from 50,000 to 120,000 with the same number of staff by automating its invoicing procedure with Medius Flow. You can also read our article about invoice automation.

1. **Self-Driving Cars**: From mining to manufacturing, self-driving cars/vehicles are increasing the efficiency and effectiveness of operations. Integrate them into your business for greater efficiency. Leverage the power of artificial intelligence for complex tasks.
2. **Vehicle Cybersecurity**: Secure connected and autonomous cars and other vehicles with intelligent cybersecurity solutions. Guarantee your safety by hack-proof mechanisms. Protect your intelligent systems from attacks.
3. **Vision Systems**: Vision systems for self-driving cars. Integrate vision sensing and processing in your vehicle. Achieve your goals with the help of computer vision.
4. **Driving Assistant**: Required components and intelligent solutions to improve rider’s experience in the car. Implement AI-Powered vehicle perception solutions for the ultimate driving experience.
5. **Fraud Detection**: Leverage machine learning to detect fraudulent and abnormal financial behavior, and/or use AI to improve general regulatory compliance matters and workflows. Lower your operational costs by limiting your exposure to fraudulent documents.
6. **Insurance & InsurTech**: Leverage machine learning to quote optimal prices, manage claims effectively, and improve customer satisfaction while reducing costs. Detect your customer’s risk profile and provide the right plan.
7. **Financial Analytics Platform**: Leverage machine learning, Natural Language Processing, and other AI techniques for financial analysis, algorithmic trading, and other investment strategies or tools.
8. **Travel & expense management**: Use deep learning to improve data extraction from receipts of all types including hotel, gas station, taxi, grocery receipts. Use anomaly detection and other approaches to identify fraud, non-compliant spending. Reduce approval workflows and processing costs per unit.
9. **Credit Lending & Scoring**: Use AI for robust credit lending applications. Use predictive models to uncover potentially non-performing loans and act. See the potential credit scores of your customers before they apply for a loan and provide custom-tailored plans.
10. **Billing**: Leverage accessible billing services that remind your customers to pay. Increase your loan recovery ratios. Use automated invoice systems for your business.
11. **Robo-Advisory**: Use AI chatbot and mobile app assistant applications to monitor personal finances. Set your target savings or spending rates for your own goals. Your finance assistant will handle the rest and provide you with insights to reach financial targets.
12. **Regulatory Compliance**: Use Natural Language Processing to quickly scan legal and regulatory text for compliance issues and do so at scale. Handle thousands of paperwork without any human interaction.
13. **Data Gathering**: Use AI to efficiently gather external data such as sentiment and other market-related data. Wrangle data for your financial models and trading approaches.
14. **Debt Collection**: Leverage AI to ensure a compliant and efficient debt collection process. Effectively handle any dispute and see your success right in debt collection.
15. **Patient Data Analytics**: Analyze patient and/or 3rd party data to discover insights and suggest actions. Greater accuracy by assisted diagnostics. Lower the mortality rates and increase patient satisfaction by using all the diagnostic data available to detect the underlying reasons for the symptoms.
16. **Personalized Medications and Care**: Find the best treatment plans according to patient data. Provide custom-tailored solutions for your patients. By using their medical history, genetic profile, you can create a custom medication or care plan.
17. **Drug Discovery**: Find new drugs based on previous data and medical intelligence. Lower your R&D cost and increase the output — all leading to greater efficiency. Integrate FDA data, and you can transform your drug discovery by locating market mismatches and FDA approval or rejection rates.
18. **Real-Time Prioritization and Triage**: Prescriptive analytics on patient data enabling accurate real-time case prioritization and triage. Manage your patient flow by automatization. Integrate your call center and use language processing tools to extract the information, priorate patients that need urgent care, and lower your error rates. Eliminate error-prone decisions by optimizing patient care.
19. **Early Diagnosis**: Analyze chronic conditions leveraging lab data and other medical data to enable early diagnosis. Provide a detailed report on the likelihood of the development of certain diseases with genetic data. Integrate the right care plan for eliminating or reducing the risk factors.
20. **Assisted or Automated Diagnosis & Prescription:**Suggest the best treatment based on the patient complaint and other data. Put in place control mechanisms that detect and prevent possible diagnosis errors. Find out which active compound is most effective against that specific patient. Get the right statistics for superior care management.
21. **Pregnancy Management**: Monitor mother and fetus health to reduce mothers’ worries and enable early diagnosis. Use machine learning to uncover potential risks and complications quickly. Lower the rates of miscarriage and pregnancy-related diseases.
22. **Medical Imaging Insights**: Advanced medical imaging to analyze and transform images and model possible situations. Use diagnostic platforms equipped with high image processing capabilities to detect possible diseases.
23. **Healthcare Market Research**: Prepare hospital competitive intelligence by tracking market prices. See the available insurance plans, drug prices, and many more public data to optimize your services. Leverage NLP tools to analyze the vast size of unstructured data.
24. **Healthcare Brand Management and Marketing**: Create an optimal marketing strategy for the brand based on market perception and target segment. Tools that offer high granularity will allow you to reach the specific target and increase your sales.
25. **Gene Analytics and Editing**: Understand gene and its component. Predict the impact of gene edits. Before using gene therapy, use models the uncover what are the possible outcomes and find are the other solutions.
26. **Device and Drug Comparative Effectiveness**: Analyze drug and medical device effectiveness. Rather than just using simulations, test on other patient’s data to see the effectiveness of the new drug, compare your results with benchmark drugs to make an impact with the drug.
27. **Forecast a big hypermarket’s sales on 2 major holidays – Christmas and Thanksgiving.**

The hypermarket has various departments, so you should predict which departments are impacted by the holiday events and what is the scope of the impact. The 2 holidays are times when the hypermarket store makes the highest sales. By predicting sales for these events, the store wants to ensure that there is enough product quantity to meet the upcoming demand. You can use the historical datasets of the store.

1. **Study the factors contributing to air pollution in a given city.**

Perform an analytical study of the air quality data for a given city, famous with its bad quality index. Identify the key factors that contribute to the rise in air pollution across the key locations in the city. The air pollution is one of the most serious problems in the world. So, this is one of the hottest data science projects and case studies.

1. **Study the likelihood of a worsening or improving diabetes in children under 12 years.**

You need to perform a predictive analysis of the data for children under 12 years diagnosed with diabetes. The purpose is to determine possible patient outcomes. You also can-do diagnostic analytics (that includes major contributing factor examination) to determine why a diabetes outcome happened.

1. **Optimize the product price of an online gluten-free grocery store.**

Let’s say the store sell 100 products online. The purpose is to use analytics to identify the right pricing which maximizes the revenue earned. This is a hard level data science project as it includes business approach, analytical methodology, and logic.

1. **Estimate the minimum age of the universe by looking at data about stars.**

For the purpose look at data about the stars. This data is collected into groups called globular clusters. Then you can come up with your own opinion for the minimum age of our universe. This is one of the very cool data science project ideas for beginners and students.

1. **Identify the common skills and qualifications of the top-performing employees in a company.**

This is one of the interesting data science project ideas in the HR area that aim to help businesses find their best talents. Create algorithms or a system that examine personality skills from the ten best performers of a company. Then use these skills for new hiring criteria for the next hiring program. This way the employee’s performance will be improved, and the company’s profitability will be increased.

1. **Why do some humans react to the same drugs differently than others?**

The people responses mainly are a result of genetic mutations. You will need to examine a database to identify how a genetic mutation contributes to a response to certain drugs. This is a typical medical and healthcare data science project problem.

1. **Do an image mining project.**

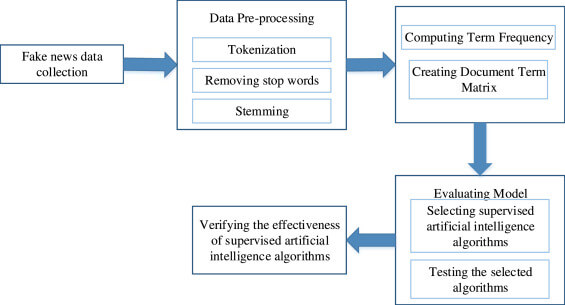
Let’s say a company produces paper designs and they have several hundreds of designs which vary in color, patterns, and etc. You need to create a predictive model to find out the probability of success of the new designs. You need to use their historical datasets to discover which designs have been more successful in the past, what colors are hot sellers, and etc.

1. **Do data mining for increasing and predicting sales in tourism.**

There is a ton of information and insights that are hidden in the tourism big data. You can use data clustering and the other techniques to understand where and when tourists like to go, what they like at each location, how tourists are traveling between spots.

1. **WhatsApp news detection**

With the advent of the technological revolution, it is easier for users to have access to the internet which increases the probability of fake news to spread like a wildfire.  In this project, you will learn how to classify news into Real or Fake. Also in current times, this will be one of the best data mining projects for project submissions. You will use Passive Aggressive Classifier to perform the above function.

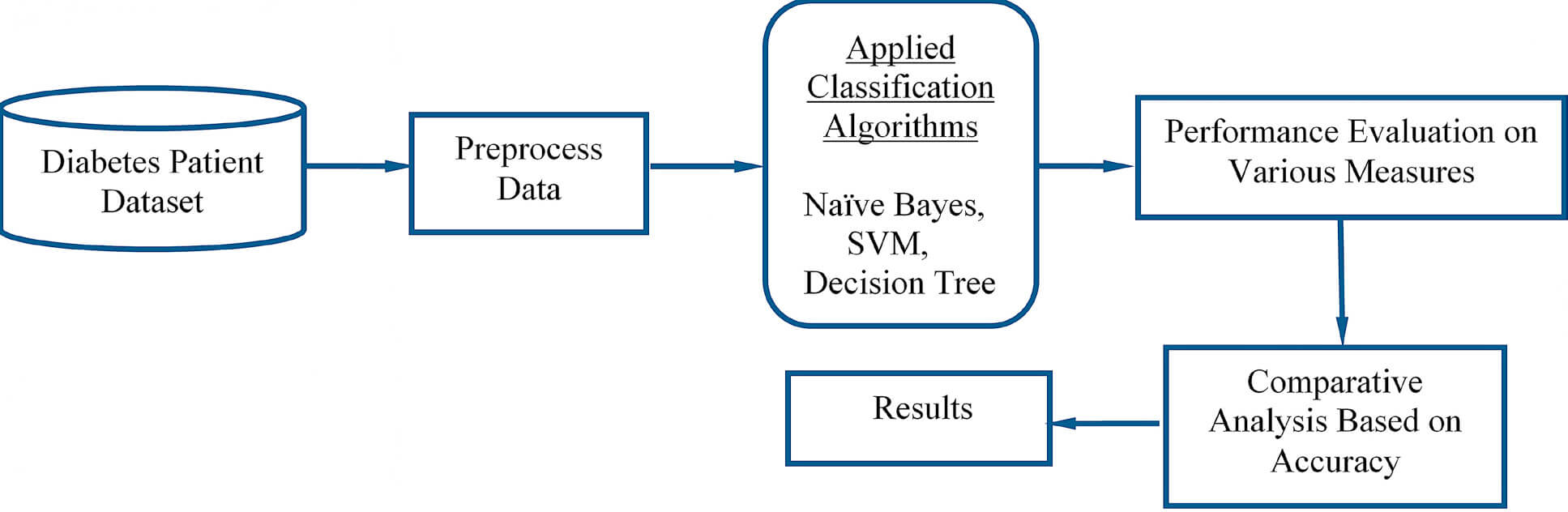


1. **Detecting Phishing website**

In recent times, technological advancement created a way for the development of e-commerce sites and most of the users started shopping online for which they have to provide their sensitive information like bank details, username, password, etc. Fraudsters and cybercriminals use this opportunity and create fake sites that look similar to the original to collect sensitive user data. In this data mining project, you will develop an algorithm to detect phishing sites based on the characteristics like security and encryption criteria, URL, domain identity, etc.

1. **Diabetes prediction**

Diabetes is one of the most common and hazardous diseases on the planet. It requires a lot of care and proper medication to keep the disease in control. In this data mining project, this project teaches you to develop a classification system to detect whether the patient has diabetes or not. As part of this project, you will learn about the Decision tree, Naive Bayes, SVM calculations, etc. Find the dataset: here.



1. **House price prediction**

In this data mining project, you will utilize data science techniques like machine learning to predict the house price at a particular location. This project finds applications in real estate industries to predict house prices based on the previous data for example the location and size of the house and facilities near the house. Find the dataset: here.

1. **Credit Card Fraud Detection**

With the increase in online transactions, credit card frauds have also increased. Banks are trying to handle this issue using data mining techniques. In this data mining project, we use python to create a classification problem to detect credit card fraud by analyzing the previously available data. We have made this credit card fraud detection project using machine learning here.

1. **Detecting Parkinson’s disease**

Data mining techniques are widely utilized in the healthcare industry to provide quality treatment by analyzing the patient’s medical records. In this data mining project, you will learn to predict Parkinson’s disease using python. The project works with UCI ML Parkinson’s dataset. Find more information about the project dataset: here.

1. **Anime recommendation system**

This is one of the favorite data mining project ideas among students. This project data set contains information on user preference data from 73,516 users on 12,294 anime. Each user is able to add anime to their completed list and give it a rating and this data set is a compilation of those ratings. The aim of the project is to create an efficient anime recommendation system based only on user viewing history. Find the dataset: here.

1. **Mushroom Classification**

This dataset contains details of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each mushroom species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter category is combined with the poisonous one. The facts suggest that there is no simple rule to determine if the mushroom is edible; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy. Find more information about the data: here.



1. **Solar Power Generation Data**

This data has been extracted from two solar power plants in India over a 34 day period. It has two pairs of files - each pair has one power generation dataset and one sensor reading dataset. The power generation datasets are extracted from the inverter level - each inverter has multiple lines of solar panels attached to it. And the sensor data is extracted from a plant level - a single array of sensors optimally located at the plant.

These are concerns at the solar power plant -

Can we predict the power generation for the next couple of days?

Can we identify the importance of panel cleaning/maintenance?

Can we identify faultily or suboptimally performing equipment?

1. **Heart Disease Prediction**

Heart disease is one of the most common diseases. It needs a lot of care by the doctor to get diagnosed. In this data mining project, you will learn to develop a system to detect whether the patient is suffering from heart disease or not. In this project, you will learn about the Decision tree, Naive Bayes, SVM calculations, etc. Find the dataset: here.

1. **Fraud Detection in Monetary Transactions**

Detecting fraudulent transactions is a very significant use case in today’s scenario of digitized monetary transactions. In order to address this problem, a Synthetic Data is generated using PaySim Simulator and it is made available at Kaggle. The data contains transaction details like transaction type, amount of transaction, customer initiating the transaction, old and new balance in Origin i.e., before and after transaction respectively and same as in Destination Account along with the target label, is fraud. So, based on the transaction details, a Classification Model can be developed that can detect fraudulent transactions.

1. **Adult Census Income Prediction**

The US Census Data is made available at the UCI Machine Learning Repository. The Dataset contains variables like age, work-class, hours per week, sex, etc. including other variables that can foretell whether the annual income of an individual is greater than 50K dollars or not. This is a Classification Problem for which a Machine Learning model can be trained to predict the Income Level of an individual.

1. **Titanic Survival Prediction**

In order to get started with Data Mining, this is the go-to project. A Titanic Dataset is created by Kaggle and a competition for the same is being hosted in this link. The data contains explanatory variables like Passenger details like Class, Gender, Age, Fare, etc. that are responsible for predicting whether a passenger will survive the Titanic Disaster or not with Survived (0/1) as the target variable. So, the Project Expectation is to build a Classification ML Model that predicts the probable survival of the passenger in Titanic.

1. **Web Data Mart Informatica (Power Center, IDE, IDQ) Project**

Project Description: The main aim and ultimate goal of this Web data mart Data Warehousing project is to make the anonymous web traffic information into meaningful analytical information. This allows measurement of what people say, how they feel, and most importantly, how they actually respond. This information is the foundation of personalized one-to-one marketing techniques, allowing the business to target specific audiences with customized products and services that directly solve their problems.

1. **Cartoonify Image with Machine Learning**

Transform images into its cartoon. Yes, the objective of this machine learning project is to CARTOONIFY the images. Thus, you will build a python application that will transform an image into its cartoon using machine learning libraries.

1. **Iris Flowers Classification Project**

The iris flowers have different species, and you can distinguish them based on the length of petals and sepals. This is a basic project for machine learning beginners to predict the species of a new iris flower.

1. **Emojify – Create your own emoji with Python**

The objective of this machine learning project is to classify human facial expressions and map them to emojis. You will build a convolution neural network to recognize facial emotions. Then you will map those emotions with the corresponding emojis or avatars

1. **Loan Prediction using Machine Learning**

The idea behind this ML project is to build a model that will classify how much loan the user can take. It is based on the user’s marital status, education, number of dependents, and employments. You can build a linear model for this project.

1. **Housing Prices Prediction Project**

The dataset has house prices of the Boston residual areas. The expense of the house varies according to various factors like crime rate, number of rooms, etc. It is a good ML project for beginners to predict prices on the basis of new data.

1. **MNIST Digit Classification Machine Learning Project**

Project idea – The MNIST digit classification python project enables machines to recognize handwritten digits. This project could be very useful for computer vision. Here you need to use MNIST datasets to train the model using Convolutional Neural Networks.

1. **Stock Price Prediction using Machine Learning**

Project idea – There are many datasets available for the stock market prices. This machine learning beginner’s project aims to predict the future price of the stock market based on the previous year’s data.

1. **Titanic Survival Project**

Project idea – This will be a fun project to build as you will be predicting whether someone would have survived if they were in the titanic ship or not. For this beginner’s project, you will use the Titanic dataset that contains real data of the survivors and people who died in the Titanic ship.

1. **Wine Quality Test Project**

Project idea – In this project, you can build an interface to predict the quality of the red wine. It will use the chemical information of the wine and based on the machine learning model; it will give you the result of wine quality.

1. **Fake News Detection Project**

Project idea – Fake news spreads like a wildfire and this is a big issue in this era. You can learn how to distinguish fake news from a real one. You can use supervised learning to implement a model like this.

1. **Music Genre Classification Machine Learning Project**

Project Idea: The idea behind this python machine learning project is to develop a machine learning project and automatically classify different musical genres from audio. You need to classify these audio files using their low-level features of frequency and time domain.

1. **Bitcoin Price Predictor Project**

bitcoin price predictor - ML project idea Project idea – The bitcoin price predictor is a useful project. Blockchain technology is increasing and there are many digital currencies rising. This project will help you predict the price of the bitcoin using previous data.

1. **Uber Data Analysis Project**

Project idea – The project can be used to perform data visualization on the uber data. The dataset contains 4.5 millions of uber pickups in the new York city. This much data needs to be represented beautifully in order to analyze the rides so that further improvements in the business can be made.

1. **Personality Prediction Project**

Project idea – The Myers Briggs Type Indicator is a personality type system that divides a person into 16 distinct personalities based on introversion, intuition, thinking and perceiving capabilities. You can identify the personality of a person from the type of posts they put on social media.

1. **Handwritten Character Recognition**

Project Idea: In this machine learning project, you will detect & recognize handwritten characters, i.e., English alphabets from A-Z. You are going to achieve this by modeling a neural network.

1. **Xbox Game Prediction Project**

Project idea – The data generated by people while searching can be used to predict the interest of the users. The BestBuy consumer electronics company has provided the data of millions of searches from users, and you will predict the Xbox game that a user will be most interested to buy. This will be used to recommend games to the visitors.

1. **Credit Card Fraud Detection Project**

Project idea – Companies that involve a lot of transactions with the use of cards need to find anomalies in the system. The project aims to build a fraud detection model on credit cards. You need to use the transaction and their labels as fraud or non-fraud to detect if new transactions made by the customer are fraud or not.

1. **Sign Language Recognition with Machine Learning**

Project Idea: A lot of research has been done to help people who are deaf and dumb. In this sign language recognition project, you create a sign detector that detects sign language. This can be very helpful for the deaf and dumb people in communicating with others

1. **Barbie with Brains Project**

Project idea – Kid toys like barbie have a predefined set of words that they can speak repeatedly. You can use machine learning methods to give the barbie some brain. It will be more engaging when a toy can understand and speak with different sentences.

1. **Customer Segmentation using Machine Learning**

Project idea – Customer segmentation is a technique in which we divide the customers based on their purchase history, gender, age, interest, etc. It is useful to get this information so that the store can get help in personalize marketing and provide customers with relevant deals. With the help of this project, companies can run user-specific campaigns and provide user-specific offers rather than broadcasting same offer to all the users.

1. **Sentiment Analysis using Machine Learning**

Project idea – Sentiment analysis is the process of analyzing the emotion of the users. You can categorize their emotions as positive, negative or neutral. It is a great project to understand how to perform sentiment analysis and it is widely being used nowadays. This is one of the most popular machine learning projects. The reason behind this is every company is trying to understand the sentiment of their customers if customers are happy, they will stay.

1. **Enron Investigation Project**

Project idea – The Enron company collapsed in 2000 but the data was made available for investigation. The database has 500,000 emails of real employees who worked in the company, so the data is very useful to perform data analytics and many data scientist use this dataset.

1. **Speech Emotion Recognition Machine Learning Project**

Project idea – This is one of the best machine learning projects. The speech emotion recognition system uses audio data. It takes a part of speech as input and then determines in what emotions the speaker is speaking.

You can identify different emotions like happy, sad, surprised, angry, etc. This project could be helpful for identifying customer emotions during the call with the call centre.

1. **Hing Illegal Fishing ProjecCatct**

Project idea – This is an interesting machine learning project. There are many ships, boats on the oceans and it is impossible to manually keep track of what everyone is doing. It will be an amazing project that can identify illegal poaching of animals and catch fishing activities through satellite and Geolocation data. The Global Fishing Watch is offering real-time data for free, that can be used to build the system.

1. **Online Grocery Recommendation using Collaborative Filtering**

Project idea – Collaborative filtering is a great technique to filter out the items that a user might like based on the reaction of similar users. A grocery recommendation system would be a great project to make customers realize what they would like in their baskets. It is good for those who are planning to start the Grocery Store.

1. **Movie Recommendation System Using Machine Learning**

Project idea – Recommendation systems are everywhere, be it an online purchasing app, movie streaming app or music streaming. They all recommend products based on their targeted customers. A movie recommendation system is an excellent project to enhance your portfolio.

1. **Automatic License Number Plate Recognition System**

Project idea: The objective of this machine learning project is to detect and recognize the license number plate of a vehicle and read the license numbers printed on the plate. This could be a good application for security scans, traffic monitoring, etc.

1. **Image Segmentation with Machine Learning**

Project Idea: Predict location as well as class to which each object in the image belongs. Image segmentation results in granular level information about the shape of an image and thus an extension of the concept of Object Detection.

1. **Spoiler Blocker Extension**

AI Project Idea – When a good movie or show comes, people always spoil the fun of others by spoiling it. We can create a browser extension that will block out all the mentions of your favourite show that you don’t want to get spoiled. You can replace the mentions with a cute picture of cats.

1. **Spam Classifier**

AI Project Idea – Every day we get dozens of email notifications and most of them are spam.

Build a tool that can classify the emails as spam or non-spam based on the content of the email alone.

1. **A data mining framework to analyze road accident data**

One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data. In this paper, we proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster-based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further a trend analysis has also been performed for each cluster and EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

1. **Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud**

Big sensor data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity it is difficult to process using on-hand database management tools or traditional data processing applications. Cloud computing provides a promising platform to support the addressing of this challenge as it provides a flexible stack of massive computing, storage, and software services in a scalable manner at low cost. Some techniques have been developed in recent years for processing sensor data on cloud, such as sensor-cloud. However, these techniques do not provide efficient support on fast detection and locating of errors in big sensor data sets. For fast data error detection in big sensor data sets, in this paper, we develop a novel data error detection approach which exploits the full computation potential of cloud platform and the network feature of WSN. Firstly, a set of sensor data error types are classified and defined. Based on that classification, the network feature of a clustered WSN is introduced and analyzed to support fast error detection and location. Specifically, in our proposed approach, the error detection is based on the scale-free network topology and most of detection operations can be conducted in limited temporal or spatial data blocks instead of a whole big data set. Hence the detection and location process can be dramatically accelerated. Furthermore, the detection and location tasks can be distributed to cloud platform to fully exploit the computation power and massive storage. Through the experiment on our cloud computing platform of U-Cloud, it is demonstrated that our proposed approach can significantly reduce the time for error detection and location in big data sets generated by large scale sensor network systems with acceptable error detecting accuracy.

1. **Big data, big knowledge: big data for personalized healthcare.**

The idea that the purely phenomenological knowledge that we can extract by analyzing large amounts of data can be useful in healthcare seems to contradict the desire of VPH researchers to build detailed mechanistic models for individual patients. But in practice no model is ever entirely phenomenological

or entirely mechanistic. We propose in this position paper that big data analytics can be successfully combined with VPH technologies to produce robust and effective in silico medicine solutions. In order to do this, big data technologies must be further developed to cope with some specific requirements that emerge from this application. Such requirements are working with sensitive data; analytics of complex and heterogeneous data spaces, including non-textual information; distributed data management under security and performance constraints; specialized analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales; and specialized analytics to define the “physiological envelope” during the daily life of each patient. These domain-specific requirements suggest a need for targeted funding, in which big data technologies for in silico medicine becomes the research priority.

1. **Deduplication on Encrypted Big Data in Cloud**

Cloud computing offers a new way of service provision by re-arranging various resources over the Internet. The most important and popular cloud service is data storage. In order to preserve the privacy of data holders, data are often stored in cloud in an encrypted form. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. Traditional deduplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few of them can be readily deployed in practice. In this paper, we propose a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. We evaluate its performance based on extensive analysis and computer simulations. The results show the superior efficiency and effectiveness of the scheme for potential practical deployment, especially for big data deduplication in cloud storage.

1. **Processing Geo-Dispersed Big Data in an Advanced MapReduce Framework**

Big data takes many forms, including messages in social networks, data collected from various sensors, captured videos, and so on. Big data applications aim to collect and analyze large amounts of data, and efficiently extract valuable information from the data. A recent report shows that the amount of data on the Internet is about 500 billion GB. With the fast increase of mobile devices that can perform sensing and access the Internet, large amounts of data are generated daily. In general, big data has three features: large volume, high velocity and large variety [1]. The International Data Corporation (IDC) predicted that the total amount of data generated in 2020 globally will be about 35 ZB. Facebook needs to process about 1.3 million TB of data each month. Many new data are generated at high velocity. For example, more than 2 million emails are sent over the Internet every second.

1. **Recent Advances in Autonomic Provisioning of Big Data Applications on Clouds**

CLOUD computing [1] assembles large networks of virtualized ICT services such as hardware resources (such as CPU, storage, and network), software resources (such as databases, application servers, and web servers) and applications. In industry these services are referred to as infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Mainstream ICT powerhouses such as Amazon, HP, and IBM are heavily investing in the provision and support of public cloud infrastructure. Cloud computing is rapidly becoming a popular infrastructure of choice among all types of organizations. Despite some initial security concerns and technical issues, an increasing number of organizations have moved their applications and services in to “The Cloud”.

These applications range from generic word processing software to online healthcare. The cloud

system taps into the processing power of virtualized computers on the back end, thus significantly speeding up the application for the user, which just pays for the used services

1. **Privacy Preserving Data Analysis in Mental Health Research**

The digitalization of mental health records and psychotherapy notes has made individual mental health data more readily accessible to a wide range of users including patients, psychiatrists, researchers, statisticians, and data scientists. However, increased accessibility of highly sensitive mental records threatens the privacy and confidentiality of psychiatric patients. The objective of this study is to examine privacy concerns in mental health research and develop a privacy preserving data analysis approach to address these concerns. In this paper, we demonstrate the key inadequacies of the existing privacy protection approaches applicable to use of mental health records and psychotherapy notes in records-based research. We then develop a privacy-preserving data analysis approach that enables researchers to protect the privacy of people with mental illness once granted access to mental health records. Furthermore, we choose a demonstration project to show the use of the proposed approach. This paper concludes by suggesting practical implications for mental health researchers and future research in the field of privacy-preserving data analytics.

1. **BFC: High-Performance Distributed Big-File Cloud Storage Based On Key-Value Store**

Nowadays, cloud-based storage services are rapidly growing and becoming an emerging trend in data storage field. There are many problems when designing an efficient storage engine for cloud-based systems with some requirements such as big-file processing, lightweight meta-data, low latency, parallel I/O, deduplication, distributed, high scalability. Key-value stores played an important role and showed many advantages when solving those problems. This paper presents about Big File Cloud (BFC) with its algorithms and architecture to handle most of problems in a big-file cloud storage system based on key value store. It is done by proposing low-complicated, fixed-size meta-data design, which supports fast and highly concurrent, distributed file I/O, several algorithms for resumable upload, download and simple data deduplication method for static data. This research applied the advantages of ZDB - an in-house key value store which was optimized with auto-increment integer keys for solving big-file storage problems efficiently. The results can be used for building scalable distributed data cloud storage that support bigfile with size up to several terabytes.

1. **Performance Analysis of Scheduling Algorithms for Dynamic Workflow Applications**

In recent years, Big Data has changed how we do computing. Even though we have large scale infrastructure such as Cloud computing and several platforms such as Hadoop available to process the workloads, with Big Data there is a high level of uncertainty that has been introduced in how an application processes the data. Data in general comes in different formats, at different speed and at different volume. Processing consists of not just one application, but several applications combined to form a workflow to achieve a certain goal. With data variation and at different speed, applications execution and resource needs will also vary at runtime. These are called dynamic workflows. One can say that we can just throw more and more resources during runtime. However, this is not an effective way as it can lead to, in the best case, resource wastage or monetary loss and in the worst case, delivery of outcomes much later than when it is required. Thus, scheduling algorithms play an important role in efficient execution of dynamic workflow applications. In this paper, we evaluate several most commonly used workflow scheduling algorithms to understand which algorithm will be the best for the efficient execution of dynamic workflows.

1. **PaWI: ParallelWeighted Itemset Mining by means of MapReduce**

Frequent itemset mining is an exploratory data mining technique that has fruitfully been exploited to extract recurrent co-occurrences between data items. Since in many application contexts items are enriched with weights denoting their relative importance in the analyzed data, pushing item weights into the itemset mining process, i.e., mining weighted itemsets rather than traditional itemsets, is an appealing research direction. Although many efficient in-memory weighted itemset mining algorithms are available in literature, there is a lack of parallel and distributed solutions which are able to scale towards Big Weighted Data. This paper presents a scalable frequent weighted itemset mining algorithm based on the MapReduce paradigm. To demonstrate its actionability and scalability, the proposed algorithm was tested on a real Big dataset collecting approximately 34 millions of reviews of Amazon items. Weights indicate the ratings given by users to the purchased items. The mined itemsets represent combinations of items that were frequently bought together with an overall rating above average.

1. **Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing**

The unprecedented growth in mobile device adoption and the rapid advancement of mobile technologies & wireless networks have created new opportunities in mobile marketing and adverting. The opportunities for Mobile Marketers and Advertisers include real-time customer engagement, improve customer experience, build brand loyalty, increase revenues, and drive customer satisfaction. The challenges, however, for the Marketers and Advertisers include how to analyze troves of data those mobile devices emit and how to derive customer engagement insights from the mobile data. This research paper addresses the challenge by developing Big Data Mobile Marketing analytics and advertising recommendation framework. The proposed framework supports both offline and online advertising operations in which the selected analytics techniques are used to provide advertising recommendations based on collected Big Data on mobile user's profiles, access behaviors, and mobility patterns. The paper presents prototyping solution design as well as its application and certain experimental results.

1. **Review Based Service Recommendation for Big Data**

The unprecedented growth in mobile device adoption and the rapid advancement of mobile technologies & wireless networks have created new opportunities in mobile marketing and adverting. The opportunities for Mobile Marketers and Advertisers include real-time customer engagement, improve customer experience, build brand loyalty, increase revenues, and drive customer satisfaction. The challenges, however, for the Marketers and Advertisers include how to analyze troves of data those mobile devices emit and how to derive customer engagement insights

from the mobile data. This research paper addresses the challenge by developing Big Data Mobile Marketing analytics and advertising recommendation framework. The proposed framework supports both offline and online advertising operations in which the selected analytics techniques are used to provide advertising recommendations based on collected Big Data on mobile user's profiles, access behaviors, and mobility patterns. The paper presents prototyping solution design as well as its application and certain experimental results.

1. **Secure Sensitive Data Sharing on a Big Data Platform**

Users store vast amounts of sensitive data on a big data platform. Sharing sensitive data will help enterprises reduce the cost of providing users with personalized services and provide value-added data services. However, secure data sharing is problematic. This paper proposes a framework for secure sensitive data sharing on a big data platform, including secure data delivery, storage, usage, and destruction on a semi-trusted big data sharing platform. We present a proxy re-encryption algorithm based on heterogeneous ciphertext transformation and a user process protection method based on a virtual machine monitor, which provides support for the realization of system functions. The framework protects the security of users’ sensitive data effectively and shares these data safely. At the same time, data owners retain complete control of their own data in a sound environment for modern Internet information security.

1. **Load Balancing for Privacy-Preserving Access to Big Data in Cloud.**

In the era of big data, many users and companies start to move their data to cloud storage to simplify data management and reduce data maintenance cost. However, security and privacy issues become major concerns because third-party cloud service providers are not always trusty. Although data contents can be protected by encryption, the access patterns that contain important information are still exposed to clouds or malicious attackers. In this paper, we apply the ORAM algorithm to enable privacy-preserving access to big data that are deployed in distributed file systems built upon hundreds or thousands of servers in a single or multiple geo-distributed cloud sites. Since the ORAM algorithm would lead to serious access load unbalance among storage servers, we study a data placement problem to achieve a load balanced storage system with improved availability and responsiveness.

Due to the NP-hardness of this problem, we propose a low-complexity algorithm that can deal with large-scale problem size with respect to big data. Extensive simulations are conducted to show that our proposed algorithm finds results close to the optimal solution, and significantly outperforms a random data placement algorithm.

1. **Enabling Efficient Access Control with Dynamic Policy Updating for Big Data in the Cloud** Due to the high volume and velocity of big data, it is an effective option to store big data in the cloud, because the cloud has capabilities of storing big data and processing high volume of user access requests. Attribute-Based Encryption (ABE) is a promising technique to ensure the end-to-end security of big data in the cloud. However, the policy updating has always been a challenging issue when ABE is used to construct access control schemes. A trivial implementation is to let data owners retrieve the data and re-encrypt it under the new access policy, and then send it back to the cloud. This method incurs a high communication overhead and heavy computation burden on data owners. In this paper, we propose a novel scheme that enabling efficient access control with dynamic policy updating for big data in the cloud. We focus on developing an outsourced policy updating method for ABE systems.

Our method can avoid the transmission of encrypted data and minimize the computation work of data owners, by making use of the previously encrypted data with old access policies. Moreover, we also design policy updating algorithms for different types of access policies. The analysis show that our scheme is correct, complete, secure and efficient.

1. **MRPrePost-A parallel algorithm adapted for mining big data.**

With the explosive growth in data, using data mining techniques to mine association rules, and then to find valuable information hidden in big data has become increasingly important. Various existing data techniques often through mining frequent itemsets to derive association rules and access to relevant knowledge, but with the rapid arrival of the era of big data, Traditional data mining algorithms have been unable to meet large data's analysis needs. In view of this, this paper proposes an adaptation to the big data mining parallel algorithms-MRPrePost. MRPrePost is a parallel algorithm based on Hadoop platform, which improves PrePost by way of adding a prefix pattern, and on this basis into the parallel design ideas, making MRPrePost algorithm can adapt to mining large data's association rules. Experiments show that MRPrePost algorithm is more superior than PrePost and PFP in terms of performance, and the stability and scalability of algorithms are better.

1. **Privacy Preserving Data Analytics for Smart Homes.**

A framework for maintaining security & preserving privacy for analysis of sensor data from smart homes, without compromising on data utility is presented. Storing the personally identifiable data as hashed values withholds identifiable information from any computing nodes. However, the very nature of smart home data analytics is establishing preventive care. Data processing results should be identifiable to certain users responsible for direct care. Through a separate encrypted identifier dictionary with hashed and actual values of all unique sets of identifiers, we suggest re-identification of any data processing results. However, the level of re-identification needs to be controlled, depending on the type of user accessing the results. Generalization and suppression on identifiers from the identifier dictionary before re-introduction could achieve different levels of privacy preservation. In this paper we propose an approach to achieve data security & privacy throughout the complete data lifecycle: data generation/collection, transfer, storage, processing and sharing.

1. **Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine- grained Updates.**

Cloud computing opens a new era in IT as it can provide various elastic and scalable IT services in a pay-as-you-go fashion, where its users can reduce the huge capital investments in their own IT infrastructure. In this philosophy, users of cloud storage services no longer physically maintain direct control over their data, which makes data security one of the major concerns of using cloud. Existing research work already allows data integrity to be verified without possession of the actual data file.

When the verification is done by a trusted third party, this verification process is also called data auditing, and this third party is called an auditor. However, such schemes in existence suffer from several common drawbacks. First, a necessary authorization/authentication process is missing between the auditor and cloud service provider, i.e., anyone can challenge the cloud service provider for a proof of integrity of certain file, which potentially puts the quality of the so-called ‘auditing-as- a service’ at risk; Second, although some of the recent work based on BLS signature can already support fully dynamic data updates over fixed-size data blocks, they only support updates with fixed- sized blocks as basic unit, which we call coarse-grained updates. As a result, every small update will cause re-computation and updating of the authenticator for an entire file block, which in turn causes higher storage and communication overheads. In this paper, we provide a formal analysis for possible types of fine-grained data updates and propose a scheme that can fully support authorized auditing and fine-grained update requests. Based on our scheme, we also propose an enhancement that can dramatically reduce communication overheads for verifying small updates. Theoretical analysis and experimental results demonstrate that our scheme can offer not only enhanced security and flexibility, but also significantly lower overhead for big data applications with a large number of frequent small updates, such as applications in social media and business transactions.

1. **KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data** Applications Service recommender systems have been shown as valuable tools for providing appropriate recommendations to users. In the last decade, the amount of customers, services and online information has grown rapidly, yielding the big data analysis problem for service recommender systems. Consequently, traditional service recommender systems often suffer from scalability and inefficiency-cy problems when processing or analyzing such large-scale data. Moreover, most of existing service recommender systems present the same ratings and rankings of services to different users without considering diverse users' preferences, and therefore fails to meet users' personalized requirements. In this paper, we propose a Keyword-Aware Service Recommendation method, named KASR, to address the above challenges. It aims at presenting a personalized service recommendation list and recommending the most appropriate services to the users effectively. Specifically, keywords are used to indicate users' preferences, and a user-based Collaborative Filtering algorithm is adopted to generate appropriate recommendations. To improve its scalability and efficiency in big data environ- ment, KASR is implemented on Hadoop, a widely adopted distributed computing platform using the MapReduce parallel processing paradigm. Finally, extensive experiments are conducted on real-world data sets, and results demonstrate that KASR significantly im-proves the accuracy and scalability of service recommender systems over existing approaches.
2. **Cost Minimization for Big Data Processing in Geo-Distributed Data Centers.**

The explosive growth of demands on big data processing imposes a heavy burden on computation, storage, and communication in data centers, which hence incurs considerable operational expenditure to data center providers. Therefore, cost minimization has become an emergent issue for the upcoming big data era. Different from conventional cloud services, one of the main features of big data services is the tight coupling between data and computation as computation tasks can be conducted only when the corresponding data is available. As a result, three factors, i.e., task assignment, data placement and data movement, deeply influence the operational expenditure of data centers. In this paper, we are motivated to study the cost minimization problem via a joint optimization of these three factors for big data services in geo-distributed data centers. To describe the task completion time with the consideration of both data transmission and computation, we propose a two-dimensional Markov chain and derive the average task completion time in closed form. Furthermore, we model the problem as a mixed-integer non-linear programming (MINLP) and propose an efficient solution to linearize it. The high efficiency of our proposal is validated by extensive simulation-based studies.

1. **Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework.** The buzz-word big-data refers to the large-scale distributed data processing applications that operate on exceptionally large amounts of data. Google’s MapReduce and Apache’s Hadoop, its open-source implementation, are the defacto software systems for big-data applications. An observation of the MapReduce framework is that the framework generates a large amount of intermediate data. Such abundant information is thrown away after the tasks finish, because MapReduce is unable to utilize them. In this paper, we propose Dache, a data-aware cache framework for big-data applications. In Dache, tasks submit their intermediate results to the cache manager. A task queries the cache manager before executing the actual computing work. A novel cache description scheme and a cache request and reply protocol are designed. We implement Dache by extending Hadoop. Testbed experiment results demonstrate that Dache significantly improves the completion time of MapReduce jobs.
2. **ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application.**

Spurred by service computing and cloud computing, an increasing number of services are emerging on the Internet. As a result, service-relevant data become too big to be effectively processed by traditional approaches. In view of this challenge, a Clustering-based Collaborative Filtering approach (ClubCF) is proposed in this paper, which aims at recruiting similar services in the same clusters to recommend services collaboratively. Technically, this approach is enacted around two stages. In the first stage, the available services are divided into small-scale clusters, in logic, for further processing. At the second stage, a collaborative filtering algorithm is imposed on one of the clusters. Since the number of the services in a cluster is much less than the total number of the services available on the web, it is expected to reduce the online execution time of collaborative filtering. At last, severalexperiments are conducted to verify the availability of the approach, on a real dataset of 6,225 mashup services collected from ProgrammableWeb.