PROPOSTA DE SOLUÇÃO PARA UTILIZAÇÃO DE REGISTRO DE OCORRÊNCIAS POLICIAIS PARA AUXILIAR NO PROCESSO DECISÓRIO GERENCIAL NO ÂMBITO DA SEGURANÇA PÚBLICA

Gláucio Bezerra Rocha[1]
Gilberto Farias de Sousa Filho[2]

RESUMO

Considerando que a segurança pública no Brasil, área de grande importância da sociedade atual, é um setor crítico nos Estados brasileiros, e sendo de alta relevância pelo fato está diretamente ligada ao maior bem jurídico do ser humano, a vida, faz necessário, a realização de estudo que permita mitigar o aumento da criminalidade por meio de uso de inteligência policial. Nesse cenário, em que a ciência oferece diversas metodologias e ferramentas para auxiliar na solução e gestão de problemas, dentre elas estão Business Intelligence, Inteligência Artificial, Séries Temporais, Problemas de Localização e Análise Criminal. Nesse contexto, onde é notório uma gestão de segurança pública sem os recursos adequados, esse trabalho propõe a utilização dos registros de crimes ocorridos na cidade de João Pessoa, fornecidos pela Polícia Civil da Paraíba, para implementação de uma ferramenta de inteligência de negócio, como também o fornecimento, por meio do uso do séries temporais, de padrões de comportamentos, assim como a aplicação de modelos estatísticos matemáticos para uma previsão de ocorrências criminais, finalizando com a submissão de algoritmos que objetivam resolver o problema de localização para uma melhor distribuição de viaturas policiais.

Palavras-chave: Inteligência de Negócio; Series Temporias; Problema de Localização; Análise Criminal; Segurança Pública.

ABSTRACT

Considering that public security in Brazil, an area of great importance in today's society, is a critical sector in Brazilian states, and being of high relevance because it is directly linked to the greatest legal good of the human being, life, it is necessary to carry out of study that allows mitigating the increase in crime through the use of police intelligence. In this scenario, in which science offers several methodologies and tools to assist in the solution and management of problems, among them are Business Intelligence, Artificial Intelligence, Time Series, Location Problems and Criminal Analysis. In this context, where public security management without adequate resources is notorious, this work proposes the use of records of crimes that occurred in the city of João Pessoa, provided by the Civil Police of Paraíba, to implement a business intelligence tool, such as also the provision, through the use of time series, of behavior patterns, as well as the application of mathematical statistical models for a prediction of criminal occurrences, ending with the submission of algorithms that aim to solve the location problem for a better distribution of police cars.

Keywords: Business Intelligence; Temporary Series; Location Problem; Criminal Analysis; Public security.

[1] Vinculação. E-mail: glauciobr@gmail.com
[2] Vinculação. E-mail: gilberto@ci.ufpb.br

1 INTRODUÇÃO

A segurança pública no Brasil, área de grande importância da sociedade atual, é um setor crítico nos Estados brasileiros. Em Ferreira e Rigueira (2013) constata que a segurança vive no limite de crises decorrentes problemas estruturais, uma mídia pesada utilizada para denegrir sua imagem, e quanto mais a criminalidade aumenta os Estados não consegue conter e controlar os problemas relacionados à segurança. É uma área de alta relevância pelo fato está diretamente ligada ao maior bem jurídico do ser humano, a vida. É responsável por proteger a vida, cessar e punir criminosos.

No Estado da Paraíba a realidade da segurança pública não é diferente. Pouca evolução percebe-se na estruturação da segurança pública, principalmente no tocante a Tecnologia da Informação como ferramenta de gestão para subsidiar a atividade de análise criminal na produção de inteligência policial. Observa-se a ausência de vários indicadores e/ou informações pertinentes e relevantes para a gestão de segurança pública que não estão acessíveis nos dias atuais para os gestores. Saber quais os crimes de maior incidência, bairros mais perigosos, gêneros das vítimas, horários mais propícios do acontecimento de crime, quais locais estão relacionados a determinados tipos de crimes, previsão de quantitativo de crimes em determinada regiões para uma atuação mais planejada de efetivo policial, dentre outros questionamentos que, não estão acessíveis, onde o gestor da segurança pública de posse desse conhecimento poderá gerir com maior eficácia um setor de tamanha relevância. A Polícia Civil, instituição ligada diretamente a segurança pública, e responsável por apurar todos os crimes ocorridos em território paraibano, no desenvolvimento de sua função de polícia judiciária, é responsável pelo registro, por meio das informações das vítimas, de todas as ocorrências de crimes que ocorreram. Esses registros, que são formalizados em Boletim de Ocorrência Policial, contém todos os dados sobre o fato criminoso sofrido pela vítima, e caso bem utilizados e submetidos a uma metodologia de análise criminal pode responder muitas perguntas hoje que estão sem respostas, impactando diretamente na gestão e colhendo resultados mais assertivos no combate à criminalidade.

Nesse cenário, a ciência oferece diversas metodologias e ferramentas para auxiliar na solução e gestão de problemas, dentre elas estão Business Intelligence, Inteligência Artificial, Séries Temporais, Problemas de Localização e Clusterização de Dados.

Com base em todo o contexto apresentado, onde é notório uma gestão de segurança pública sem os recursos adequados para sua gestão, esse trabalho apresenta uma proposta para desenvolvimento de uma solução que utilizará milhares de registros de ocorrências policiais reais que atualmente são ignorados, subutilizados, mais precisamente da cidade de João Pessoa, que não são utilizados para fins de estudo do fenômeno do comportamento da segurança pública local, para fornecer suporte a gestão pública. A proposta é projetar e implementar um sistema de Business Intelligence que servirá de suporte para tomada de decisão, externando, de forma intuitiva, os dados em painel gerencial (dashboard). Para gerar mais conhecimento policial, os dados das ocorrências serão clusterizados, mapeando regiões com seus respectivos índices de criminalidade, assim como serão apresentados a análise desses dados para que o gestor possa identificar padrões, tendência e sazonalidade por meio de séries temporais, como também submeter esses dados a modelos matemáticos estatísticos para a realização de previsão de ocorrências policias, para um melhor planejamento onde a presença estatal se faz mais necessária. Posteriormente, com a identificação das regiões com maiores índices de criminalidade e a previsão de crimes nessas regiões, esse estudo propõe aplicar técnicas de problema de localização (Problema de Localização de Máxima Cobertura Capacitada) que permitirá um melhor uso dos recursos e aparatos policiais, identificando e propondo locais que viaturas policiais ostensivas possam se situar para atender uma maior parte população, considerando melhor distância e tempo, com valores mínimos aceitáveis.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SÉRIES TEMPORAIS

Série temporal é definido por Garcia (2000) como uma sequência de dados ordenados cronologicamente. Esses dados são oriundos de eventos registrados sempre em função de uma variável tempo, podendo ser agrupados de forma diária, semanal, mensal, anual, etc. Esses eventos são vistos em diversas áreas de conhecimento, tais como: economia, meteorologia, saúde, comércio, etc. Morettin and Toloi (1981) trás alguns exemplos de séries temporais:

- Estimativa trimestral do Produto Nacional Bruto (PNB);
- Valores diários de temperatura;
- Índices diários da Bolsa de Valores;
- Quantidade mensal de chuva;
- Valores mensais de vendas de veículos;
- Registro de marés.

Outros exemplos de séries temporais apresentadas por Morettin and Toloi (2004) são:

- Valores mensais de consumo de energia elétrica;
- Emissão diária de poluentes de uma cidade;
- Índices do Produto Interno Bruto;
- Índices de custos de vida de uma cidade;

Como pode ser observado, as séries temporais estão bem mais presentes no dia-a-dia do que possamos imaginar. O objetivo de uma série temporal é analisar os dados registrados e identificar padrões, e a partir dessa análise poder entender alguns comportamentos, como exemplo, como um produto está sendo comercializado, se está sendo bem aceito, épocas do ano que se aumentam as vendas, assim como permitir a realização de previsões de vendas futuras. Morettin and Toloi (1981) apresenta alguns objetivos da análise de uma série temporal:

- Investigar o mecanismo que gerou a série temporal;
- Fazer previsões de valores;
- Descrever comportamentos;
- Procurar periodicidade relevantes nos dados.

Para Ehlers (2007), os objetivos de analisar uma série temporal, de modo geral, podem ser:

- Descrever as propriedades da série, tais como: tendência, sazonalidade, discrepâncias, alterações estruturais, etc;
- Predição, ou seja, com base em dados do passado prevê valores futuro.
- Controlar qualidade de processos;
- Explicar a variação de uma série temporal baseado em outra série temporal.

No tocante a predição de valores, Ehlers (2007) enfatiza que a previsão do futuro envolve o fator incerteza, onde as elas não são perfeitas, mas tem que buscar sempre reduzir os erros dessa previsão.

Para Ehlers (2007) a ordem dos dados apresentados em uma série temporal é determinante, assim como aponta que a principal característica a ser observada numa série temporal é que os dados vizinhos são dependentes, onde se faz necessário estudar, analisar e modelar essa dependência. Outras características apresentadas afirmam que as observações correlacionadas em séries temporais são difíceis de análise necessitando de técnicas específica para esse fim, assim como ratifica a dificuldade de lidar com observações perdidas e dados discrepantes devido à natureza sequencial.

Segundo Morettin and Toloi (2004), os modelos que descrevem séries temporais são processos estocásticos. Ehlers (2007) define processos estocásticos como uma coleção de variáveis aleatórias que são ordenadas no tempo e definidas como um conjunto de pontos que podem ser discreto ou contínuos. A definição de um modelo de classificação de série temporal vai depender de vários fatores, tais como: comportamento do fenômeno em análise ou conhecimento a priori sobre a natureza e do objetivo da análise. Uma das classificações de processos estocásticos são os processos estacionários ou não-estacionário. Um processo estacionário se desenvolve quando a origem do tempo não é importante na análise, apresentando uma constância, diferentemente do processo não-estacionário.

Morettin and Toloi (2004) classifica os tipos de modelos de séries temporais em dois, quais sejam: modelos paramétricos e os não paramétricos. Os modelos paramétricos são aqueles que os números de parâmetros considerados são finitos, já os não paramétricos envolvem um número infinitos de parâmetros.

Os modelos paramétricos toda a análise é realizada no domínio do tempo, e os modelos mais usados são os modelos de regressão, modelos auto-regressivo e de média móvel (ARMA), modelos auto-regressivo integrados e de médias móveis (ARIMA), modelos de memória longa (ARFIMA), modelos estruturais e modelos não-lineares. Os modelos não paramétricos mais usados são os a função de auto-covariância, ou autocorrelação, e sua transformada de Fourier, o espectro.

Quando falamos em série temporal temos que considerar alguns componentes desta que são fundamentais para seu entendimento, quais sejam: a sazonalidade, o ciclo da série, a aleatoriedade e a tendência (Garcia, 2000).

Para Garcia (2000) e Ehlers (2007), a sazonalidade diz respeito a um padrão identificado que se repete a cada p período de tempo idêntico, onde p é denominado de fator de sazonalidade. Um exemplo prático acontece na venda de protetores solar no período de verão, onde pode ser percebido um aumento nas vendas desses produtos. Ehlers (2007) apresenta dois tipos de sazonalidade, a Aditiva e a Multiplicativa. A sazonalidade aditiva a série temporal apresenta flutuações sazonais próxima a uma constância não considerando o nível global a série. Supondo que seja esperado um aumento de 1 milhão de reais de venda em um determinado mês do ano, esse valor deve ser somado a média anual naquele mês especificamente. Já a sazonalidade multiplicativa as flutuações sazonais vão depender do nível global as séries. Considerando o cenário de venda no qual espera-se um aumento de 10% em relação da média anual em um período específico, e seu valor dependerá do valor da média anual, por isso denominada de sazonalidade multiplicativa.

Garcia (2000) define com ciclo da série como a variação cíclica de um determinado padrão. Esse componente difere da sazonalidade por não ter um intervalo frequente dessa variação, apresentando uma desregularidade.

A tendência de uma série temporal mostra o crescimento ou declínio dos valores dessa série. É um componente que apresenta mudanças graduais em longo prazo. Para Ehlers (2007) a tendência em uma série pode apresentar vários padrões, quais sejam:

- Crescimento linear: onde espera-se um aumento nas vendas a cada período de 1 milhão de reais;
- Crescimento exponencial: onde espera-se um aumento nas vendas a cada período em um percentual de 30%, resultando em um fator de 1,3.
- Crescimento amortecido: onde espera-se um aumento nas vendas a cada período em um percentual sobre o aumento de vendas registrado no período anterior. Se no período x foi registrado um aumento de 1 milhão de reais que correspondeu a 70% de aumento, nos próximos períodos o aumento esperado será que 70% de 1 milhão de reais, resultando em 700 mil, e assim acontecerá nos demais períodos, mantendo o percentual, mas o aumento bruto sendo amortecido.

Garcia (2000) define a aleatoriedade de uma série temporal como um comportamento identificado que não se encontra explicação, não sendo possível realizar nenhum tipo de previsão, justificando até uma possível falha de previsão. Muitas vezes são resultantes de fatos inesperados como pandemias, catástrofes naturais etc.

2.2 PREVISÃO DE DADOS POR MEIO DE SÉRIES TEMPORAIS

Como já mencionado, as séries temporais são formadas por dados oriundo de eventos e dispostos em relação a variável tempo. Elas estão presentes nas mais diversas áreas de sociedade, e um dos objetivos da análise de séries temporais é a previsão de valores futuros.

A previsão de valores por meio de séries temporais, segundo Taylor and Letham (2017), ajuda na gestão das organizações, permitindo que seja realizado um melhor planejamento, tais como: capacidade suportada, definição de metas e detecção de anomalias.

Apesar da importância da previsão de dados baseado em séries temporais, para Taylor and Letham (2017) muitos desafios estão presentes nesse contexto, principalmente na previsão de dados confiáveis e de alta qualidade, tendo em vista a grande variedade dessas séries temporais e a raridade de profissionais com experiências em modelagem.

Para Ahmed et al. (2020) os modelos mais populares para previsões de séries temporais são o ARIMA, RNN e o LSTM. Nesse contexto o Facebook desenvolveu um modelo baseado em aprendizado de máquina, denominado Prophet, para previsão de séries temporais.

Segundo Ahmed et al. (2020) o Prophet é um modelo baseado em regressão aditiva simples $y(t)$, compostos de três principais componentes, quais sejam: tendência, sazonalidade e efeitos de feriados. Ele foi desenvolvido para facilitar as previsões de alta qualidade e obter uma previsão mais precisa e realista, sendo necessário apenas uma quantidade de dados histórico para melhor modelagem.

Taylor and Letham (2017) apresenta o Prophet como uma abordagem prática para previsão em escala que combina modelos configuráveis. Prophet é um modelo de regressão simples e modular com parâmetros configuráveis que podem ser ajustados pelo especialista que tem domínio do conhecimento sobre a série temporal, e que funciona bem com parâmetros padrão. É aplicado um modelo de série temporal decomponível com três componentes: tendência, sazonalidade e feriados, conforme equação seguinte:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

A função $g(t)$ representa a tendência, modelando as mudanças não periódicas no valor da série temporal. Já $s(t)$ é a função que representa as mudanças periódicas, como sazonalidade semana, anual etc. Por fim, a função $h(t)$ representa os efeitos dos feriados. O

termo ϵ_t representa o erro, que são as mudanças incomuns que não são comportadas pelo modelo.

Em Prophet (2022) demonstra que aplicação prática do modelo Prophet se dar de forma simples, por meio da API Prophet, basta instanciar a classe Prophet, e posteriormente invocar os métodos fit() e predict().

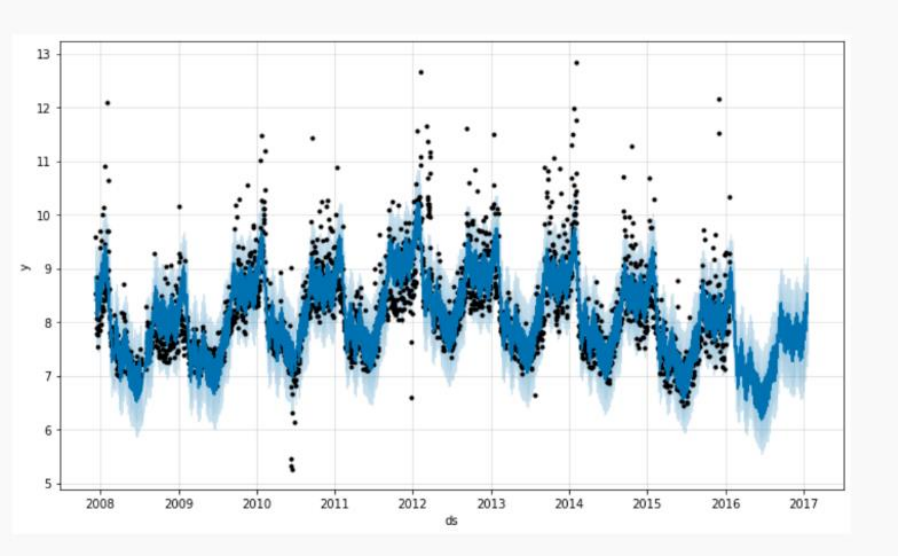
Inicialmente quando é instanciada a classe Prophet é possível passar para seu construtor algumas configurações para serem utilizadas na previsão. O método fit() recebe um dataframe de formado do Pandas, com os dados histórico, possuindo duas colunas, sendo uma chamada ds, com formado AAAA-MM-DD (ano-mês-dia), e outra coluna chamada y, que representa a medida que se deseja prever. É importante ressaltar que os nomes das colunas deve ser, obrigatoriamente, os nomes supracitados, e qualquer alteração destas a API não proverá o resultado esperado.

É possível a API considerar dados não diários, ou seja, subdiários, e para esses casos a coluna ds deverá ser AAAA-MM-DD HH:MM:SS, consequentemente a sazonalidade diária é ajustada automaticamente.

Após a invocação do método fit() com os dataframe contendo os dados histórico, é necessário gerar um dataframe com a quantidade de dias que deseja a previsão. Esse dataframe pode ser gerado por meio do método make_future_dataframe(periods=365), passando como parâmetro o periods, que representa a quantidade de dias a ser gerada a previsão.

Já com o modelo pronto e o dataframe com as datas futuras preparado, invoca-se o método predict() com o referido dataframe, que resultará em um dataframe contendo as seguintes colunas: ds, yhat, yhat_lower e yhat_upper. A previsão encontra-se na coluna yhat. É possível plotar a previsão por meio do método plot(dataframe_futuro), como pode ser visualizado exemplo na figura 1 abaixo:

Figura 1 – Exemplo de previsão gerado pela API Prophet



Fonte: Prophet (2022)

É possível constatar no gráfico da Figura 1 os dados histórico reais representados por pontilhados de cor preto, e na cor azul escuro os ajustes desses dados histórico reais, e por fim, em dado momento percebe-se que os dados histórico reais são cessados (pontilhado de cor preto) e apenas são plotados os dados do futuro, em continuidade na cor azul escuro, resultado da API Prophet.

Por fim, para mensurar o erro do processo de previsão da API Prophet, se faz necessário a utilização de métricas de avaliação, dentre as quais destaca-se: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e a Raiz Quadrada do Erro Médio (RMSE).

O cálculo do Erro Médio Absoluto (MAE) é dado pela seguinte expressão:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|^2$$

O cálculo do Erro Médio Quadrático (MSE) é dado pela seguinte expressão:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

A raiz Quadrada do Erro Médio (RMSE) é dada pela seguinte expressão:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}$$

2.3 PROBLEMA DE LOCALIZAÇÃO

Para Horner (2009) os problemas de localização é objeto de estudo há muitos anos, onde vários povos nômades procuravam por regiões para se alocarem considerando vários fatores que influenciariam em seu dia-a-dia, tal como alimentação, água, solo produtivo, facilidade de acesso, etc.

Para Oliveira (2020) a caracterização dos problemas de localização sao compostas por quatro componentes, quais sejam: Cliente, facilidades que serão instaladas, area onde os clientes estão que deverão ser cobertas pelas facilidades e a distância entre as instalações e clientes.

Ferreira and Ferreira (2012) afirma que os problemas de localização são muito discutido na literatura e que existem várias abordagem para determinar uma melhor localização, seja para negócios, imóveis, equipamentos, ou qualquer dispositivo de baixa locomoção que necessitem permanecer em uma dado local por um longo tempo. Ferreira and Ferreira (2012) diz que um dos primeiros trabalhos sobre o tema tratava-se da melhor localização de atividades agrícolas de uma cidade com objetivo de minimar custos com transportes. Oliveira (2020) afirma que os modelos de localização podem ser classificados em três grupos, a depender do objetivo da otimização, são eles:

- modelos de p-mediana, que consiste em minimizar a distancia entre a facilidade e o cliente;
- modelos de p-central, que visam minimizar a distância máxima da facilidade e do cliente;
- modelos de cobertura, que tem como objetivo garantir a cobertura para os clientes dentro de uma distância (raio) pré-estabelecida.

Isler, Bonassa and Cunha (2012) define que o problema de p-mediana tem como finalidade identificar p instalações (medianas) em um conjunto de n pontos de demandas $(n > p)$, alocando os pontos remanescente $(n - p)$ as médianas, de modo que a soma da distancia, tempo, etc, entre as medianas e os pontos de demanda seja mínimo. Para Lorena

and Senne (2005) o problema de p-mediana considera decisões ótimas sobre a localização de p instalações levando-se em conta a distância e capacidade de serviço.

A p-mediana é o modelo matemático mais conhecido em relação ao problema de localização de instalações que oferecem prestação de serviços, onde existem diversos modelos que podem ser aplicados a situações do mundo real. O modelo p-mediana não capacitado é um dos modelos de p-mediana presente a literatura, sendo um modelo que desconsidera as restrições de capacidades das instalações, ou seja, uma unidade instalada atender a qualquer quantidade de clientes, já o modelo de p-mediana capacitado absorve restrições de capacidade, onde a capacidade de atendimento da mediana tem que ser respeitada, não permitindo ser extrapolada. (Santos et al., 2020).

Muller and Santos (2006) afirmam que o Set Convering Problem (SCP) e Maximum Coverage Location Problem (MCLP) são as duas versões mais conhecida para o problema de localização.

Cobrir uma região em sua totalidade muitas vezes tonar-se inviável devido limitações de ordem econômica. Nesse sentido, o Problema de Localização de Máxima Cobertura (PLMC), desenvolvido por Church and Revelle (1974), buscar uma solução que proporciona níveis aceitáveis de cobertura, onde localiza um número determinado de facilidades compatíveis com os recursos disponívels, de tal forma que o máximo de clientes possível de uma região seja coberto a uma distância crítica S predefinida (GALVÃO et al., 2003). Para Pontin et al. (2010) esse modelo não busca atender toda uma população, mas oferecer o máximo de atendimento com os recursos disponíveis. Ele afirma que o conceito de cobertua está relacionado ao fato de se verificar de um determinado ponto está dentro de uma determinada distancia ou tempo até uma facilidade. Church and Revelle (1974) apresenta o modelo matemático formulado para o Problema da Máxima Cobertura. Segue:

$$I.\ Maximinize \qquad z = \sum_{i \in I} \alpha_i y_i$$

$$S.T. \qquad \sum_{j \in N_i} x_i \geq y_i \quad para\ todos\ i \in I \qquad (1)$$

$$\sum_{j \in J} x_i = p \qquad (2)$$

$$x_j = (0,1) \quad para\ todos\ j \in J \qquad (3)$$

$$y_i = (0,1) \quad para\ todos\ i \in I \qquad (4)$$

Onde:
I: indica o conjunto de demanda de nós;
J: indica o conjunto de facilidades;
S : indica a distância máxima que um nó estará coberto;
d_{ij}: sendo a menor distância entre os nós i e j;
x_j: 1 se a facilidade for alocada em j e 0, caso contrário;
N_i: $\{j \in J \mid d_{ij} \leq S\}$
α_i: representa a população a ser atendida pelo nó i;
p: representa o número de facilidades a ser alocadas.

Church and Revelle (1974) ratifica que o objetivo do modelo supracitado é maximinizar o número de clientes atendidos ou coberto, dentro de uma distância máxima S de facilidades desejadas. N_i representa o conjunto de facilidades disponível para realizar a cobertura dos conjuntos de nós I.

O modeo apresentado por Church and Revelle (1974) a restrição (1) só permite que y_i seja igual a 1 apenas quando um nó for coberto por uma ou mais facilidades, ou seja, existe pelo menos uma facilidade dentro da distância máxima S. O numero de facilidades disponíveis aos nós é restrito a p, conforme destaca a restrição (2). De acordo com a restrição (3) quando x_j for igual a 1, a facilidade está atendendo a um nó específico, dessa forma, pode-se afirmar, por meio da restrição (4) que um nó só estará coberto, ou seja y_i igual a 1, quando uma ou mais facilidades estiverem dentro de uma distância máxima S, caso contrário o nó em questão está fora da região de cobertura. Por fim a solução supracitada não objetiva apenas na maior quantidade de nós a ser atendidos, mas como também a quantidade de facilidades instaladas.

Pontin et al. (2010) destaca que a localização de hospitais, atendimentos de emergências ou corpo de bombeiros, e a distancia ou tempo de deslocamento entre pontos de demandas e a facilitadas são fatores importantes para estabelecer um nível de qualidade oferecidos aos usuários. Oliveira (2020) destaca que definir a localizaçao de facilidades é um problema crítico nas organizações porque são decisões que refletem em vários outros fatores e aspectos, em nível operacional e logístico, e decisões erradas podem gerar altos cursos e perda de competitividade.

2.4 CLUSTERIZAÇÃO DE DADOS

O problema de clusterização é definido por Dias (2004) como um processo em que se agrupa elementos de um conjunto em clusters, de modo que cada cluster (ou grupo) represente uma configuração em que cada elemento pertencente a ele possua uma maior similaridade com outros elementos desse mesmo cluster em relação aos demais elementos de outros clusters. Cruz and Ochi (2011) define clusterização como um processo que une objetivos similares em grupos (ou clusters). Para Drummond, Ochi and Rosário (2006) clusterização é uma técnica de agrupar dados utilizando critérios como similaridade ou dissimilaridade, podendo seus métodos serem classificados como hierárquico ou não hierárquico. Campello and Hruschka (2006) define clusterização como uma tarefa no qual tem como objetivo determinar um número finito de conjuntos de acordo com as semelhaças de seus objetos.

Dias (2004) e Cruz and Ochi (2011) afirmam que o número de clusters pode ser ou não conhecido, e caso ele seja um dado de entrada da aplicação, a literatura o referencia como "problema de k-clusterização", onde k indicaria o número de clusters a serem formados, por outro lado, caso k não seja fornecido, estariamos diante de um "problema de clusterização automática", dessa forma, para Cruz and Ochi (2011) k não sendo conhecido trata-se de um problema bem mais complexo o que aumenta substanciamente o número de soluções possíveis.

O problema de clusterização automática é definido matematicamente por Cruz and Ochi (2011) da seguinte forma: $X = \{x_1, x_2, \dots, x_n\}$, sendo X um conjuntos de vários objetos, onde cada objeto x_i é um ponto no espaço R^p, representado por uma tupla $(x_{i1,} x_{i2}, \dots, x_{ip})$, e cada coordenada x_{ij} trata-se de um atributo j do objeto i. Com objetivo de encontrar o conjunto $C = (C_{1,} C_2, \dots, C_k)$ automaticamente, onde k (número de clusters) não ser conhecido, considera-se a similaridade de objetos do mesmo clusters seja maximinizada, por outro lado, a similaridade de objetos de clusters diferentes sejam minimizadas, conforme condições expressas abaixo:

$$C_i \neq \boldsymbol{\phi}, para \quad i = 1, \dots, k \qquad (1)$$

$$C_i \cap C_j = \boldsymbol{\phi}, para \quad i, j = 1, \dots, k \; e \; i \neq j \qquad (2)$$

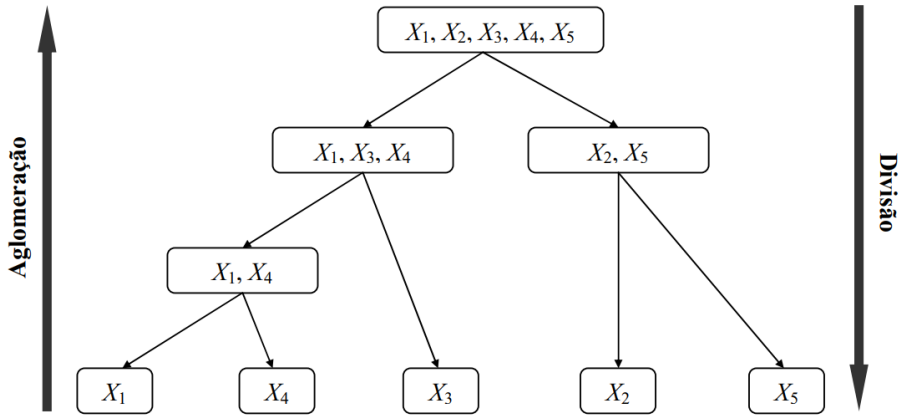
$$\bigcup_{i=1}^{k} C_i = X \qquad (3)$$

Na condição (1) para cada clusters C_i criado ele deve conter, obrigatoriamente, um ou mais objeto, bem como na condição (2) que afirma que a interceção $C_i \cap C_j$ de dois clusters tem que ser vazio, dessa forma um objeto não pode fazer parte de mais de um clusters. Por fim, na condição (3), após a criação automatica de todos os clusters, a união deles deve compreender todo o conjunto X e objetos do espaço R^p.

Dias, Ochi and Soares (2004) afirma que no processo de clusterização a busca de melhores soluções dentre as melhores soluções possíveis, é um problema NP-Dificil. Dessa forma, são apresentados na literatura métodos heuristicos com soluções sub-otimas, porem devido a grande diversidades das aplicações que tratam de problemas de clusterização, as heuristicas são desenvolvidas com foco em determinada classe de problema.

A clusterização utilizando método heuristico hierárquico tem como objetivo gerar os clusters gradativamente por meio de aglomerações de elementos, onde cada clusters com tamanho maior de 1 pode ser considerado como sendo composto por clusters menores, conforme pode ser visualizando na figura 2. Já clusterização por meio do método heuristico não hierárquico o algoritmo divide o conjunto de elementos em k conjuntos (podendo k ser conhecido ou não), onde cada configuração obtida é submetida a uma dada função, e caso a avaliação da clusterização sinalize que a configuração não atende o problema, nova configuração é obtida realizando a migração de elementos entre os clusters. O processo continua, iterando até que uma condição de parada seja atendida. (Dias, Ochi and Soares, 2004)

Figura 2 – Exemplo de previsão gerado pela API Prophet



Fonte: Dias, Ochi and Soares (2004)

O processo de agrupamento, segundo Söküt Açar and Ayman Öz (2020), visa agrupar elementos baseado nas suas características verificando quanto as suas semelhanças e dissimilaridades. Para PRIMEIRO k-means é o método mais popular da técnica hierárquica, baseado no algoritmo sharp set, onde ele separa os objetos n em grupos k como $S = (S_1, S_2, \dots, S_k)$, e a atribuição dos objetos aos grupos se dão utilizando a média de grupos mais próximos, permitindo apenas que cada objeto pertença a apenas um agrupamento. Chartier and Morissette (2013) apresenta três algoritmos de clusterização baseados no k-means, que

são os mais usados, com objetivos e resultados poucos diferentes, quais sejam: Forgy/Lloyd, MacQueen e o Hartigan & Wong.

Para avaliar o desempenho de um algoritmo de clusterização, a literatura dispõe de várias métricas, também chamadas de medidas de validades, que foram propostas. Campello and Hruschka (2006) afirma que dentre as métricas mais difundidas estão o Fuzzy Hypervolume and Average Partition Desinty, o Xie-Beni index, Average Within-Cluster Distance e o Average Silhouette Width Criterion. Campello and Hruschka (2006) destaca que essas medidas de validades de clusters mencionadas possui características particulares que quando aplicadas em determinadas classes de problemas podem se comportar de forma de uma superar a outra, alternando seus resultados quando na mudança da classe do problema.

Söküt Açar and Ayman Öz (2020) apresenta o índice Silhouette (SIL) como sendo uma das formas de validar um cluster, de bom desempenho, aplicado a várias medições de distância, cujo objetivo é determinar a adequação de cada objeto a um cluster. Vejamos:

$$Sil(x_i) = \frac{b(x_i) - a(x_i)}{maks(a(x_i), b(x_i))}$$

Conforme definição presentada por Söküt Açar and Ayman Öz (2020), $a(x_i)$ representa a dissimilaridade média do objeto i para todos os outros objetos s no mesmo cluster, e $b(x_i)$ como sendo a mínima dissimilaridade média do objeto i para todos os objetos em cluster mais próximo. $Sil(x_i)$ fornece a desigualdade de $-1 \leq Sil(x_i) \leq 1$, indicando que quando seu valor está próximo de 1 significa que o objeto i está bem classificado. Quando $Sil(x_i)$ está próximo de 0 pode-se afirmar que o objeto i_n está entre dois clusters, e quando $Sil(x_i)$ está próximo de -1 indica que o objeto i_n está classificado incorretamente. Para o autor, a média de $Sil(x_i)$ obtida para o número de clusters relevantes indica sua validade, de forma geral, se o seu valor médio estiver acima de 0,5, então os clusters esperados serão alcançados.

2.5 BUSINESS INTELLIGENCE

Business Intelligence (BI), ou Inteligência de Negócio, é definida por Braghittoni (2015) como sendo um conjunto de conceitos e métodos que tem como objetivo melhorar, no âmbito da gestão, a tomada de decisão, utilizando-se de sistemas que tenha como arquitetura estrutural a base de fatos e dimensões, para aglutinar e processar os dados de um domínio específico. É importante destacar que o BI não é uma ferramenta, e sim uma forma de manipular um conjunto de dados para obter vantagem de gestão, e pode ser implementado com várias ferramentas.

Atualmente as organizações estão sendo pressionadas a agir conforme a evolução, e para isso elas tem que inovarem para responderem as demandas de tomadas de decisões operacionais frequentes, onde essas demandas devem ser respondidas rapidamente, e em alguns casos até em tempo real, porém devido a grande quantidade de dados e informações, há necessidade de uso de tecnologia computadorizada (Ramesh, 2019). O BI tem como foco especificamente ajudar o gestor da tomada de decisão, podendo, através da análise dos dados entender fenômenos ocorridos, assim como realizar um planejamento estratégico mais fidedigno.

Os principais objetivos de um sistema de BI, conforme Ramesh (2019), são:

- Permitir o acesso interativo dos dados, em muitas vezes em tempo real;
- Proporcionar a manipulação de dados e fornecer ao gestor e analistas de negócio a capacidade de realizar a análise adequada;

- Ao analisar os dados, situações e desempenhos históricos e atuais, os tomadores de decisões conseguem valiosos insights que podem servir como base para decisões melhores e mais informadas.

Para um melhor entendimento no contexto do BI, é importante a definição de três conceitos relacionado, que são: dado, informação e conhecimento. Dado é definido de forma simples e objetiva como sendo o registro de um evento, porém, expressado sozinho não tem significado. A Informação pode ser definida como um conjunto de dados relacionados e contextualizado, com um significado expressivo. Por fim, o Conhecimento é a interpretação da informação, com resultado de ações e tratamento específicos, que subsidie a condução de algo de forma mais assertiva (Faceli, Lorena, Gama e Carvalho, 2011).

A grande maioria dos sistemas de informações que comumente está presente no dia-a-dia na sociedade são chamados de sistemas transacionais, que dão suporte, em nível operacional, as empresas e seus negócios, a exemplo de sistemas que registram vendas em supermercados, que realizam saques bancários, pagamentos de contas etc, e consequentemente alimentam diariamente uma base de dados denominada de base de dados operacionais. Os sistemas transacionais são denominados de OLTP (online transaction processing). Diferente dos sistemas OLTP, os sistemas de processamento analíticos, denominado de OLAP (online analytical processing) são bem mais incomuns de serem encontrados, porém é de uma grande relevância, tendo em vista ser capaz de proporcionar aos gestores do negócio a possibilidade e examinar dados em busca de informação e conhecimento a respeito do negócio para embasar decisões gerenciais.

Os sistemas de Business Intelligence devem ser desenvolvidos utilizando uma metodologia correta para que seja eficaz e cumpra o seu papel. Para Primak (2008), seis passos devem ser seguidos para que um BI seja implantando de forma correta. São eles:

- Identificar as necessidades a serem endereçadas na solução de BI, onde estas devem ser relevantes aos objetivos e estratégias do negócio;
- Identificar as fontes de dados já existentes na organização. Normalmente as organizações têm várias informações em bancos de dados, planilhas e arquivos, e em muitos casos é necessário criar mais informações, porém, é de extrema importância mapear aquelas já existentes;
- Extrair, transformar e carregar os dados para criar uma base de dados multidimensional orientada por assunto. Este processo deve garantir que todas as informações relevantes sejam contempladas consistentes;
- Ajudar a organização a escolher a ferramenta de apresentação para visualizar e analisar as informações resultantes do item anterior;
- Criar relatórios padrões, permitir análise sob demanda e mineração de dados (Data Mining) visando obtenção de insights sobre os indicadores chaves de desempenho;
- Planejar uma implantação de forma abrangente para toda a corporação, de forma a garantir que os tomadores de decisão tenham a informação adequada quando e onde eles precisarem.

Algumas ferramentas fazem parte do projeto de um sistema de BI, dentre elas estão o Data Wharehouse (DW), Data Mining (DM) e a ferramenta de OLAP. Os DW e DM são ferramentas especiais de armazenamento de dados, responsável por dar sustento a camada de inteligência de negócios (Primak, 2008). Os dados presentes nessas estruturas estão modelados utilizando metodologia de componentes de Fato e Dimensão. No Data Warehouse estão contidos os dados organizados por assunto e por data, já os Data Mining contêm um subconjunto lógico e físico do DW, pertencentes a uma área específica de uma organização,

por exemplo: finanças, vendas, contabilidade, marketing etc (Braghittoni, 2015). Braghittoni (2015) define Data Warehouse como um conjunto de dados orientado por assunto, conciso e integrado, variável com o tempo e não volátil. Já as ferramentas de OLAP (Online Analytical Processing) trabalham com os dados, dos DM, com operadores dimensionais, permitindo dessa forma uma abordagem múltipla e combinada de análise. Nesse cenário conclui então que a dinâmica para a construção do BI é realizada com a obtenção dos dados dos sistemas transacionais presentes nas organizações, posteriormente esses dados são submetidos ao processo de extração, transformação e carregamento (ETL), gerando dessa forma os DW e DM, modelados na perspectiva de Fato e Dimensões.

O primeiro passo de um projeto de BI é a criação do Data Warehouse (DW) responsável por conter os dados das ocorrências que serão manipulados. O DW é o coração do projeto de BI. Ele é criado em um banco de dados relacional como qualquer sistema transacional (OLTP), mas com um formato que possa responder as pesquisas de forma mais performática possível. É importante destacar que as ferramentas de BI podem trabalhar consultando os dados que estão armazenados nas bases dos sistemas de uso operacional, porém, não é a forma indicada e adequada, isso por vários motivos, inclusive de performance, que pode afetar a rotina diária no ambiente operacional. O DW ele segue uma estrutura arquitetural, regras de design e boas práticas específica de sua metodologia. Ele é baseado em uma arquitetura dimensional, com foco nos fatos e dimensões. Dessa forma, teremos as tabelas referentes as dimensões e a tabela de fato. As tabelas de dimensões deverão ser as primeiras a serem criadas, isso porque elas representam as chaves primárias dos dados que estão dentro da tabela fato que armazenarão as chaves estrangeiras.

Por fim, quando pensamos em sistemas transacionais (OLTP), aqueles que estão presentes com mais frequência em nosso dia-a-dia, realizando as atividades operacionais das organizações, ou seja, vendas, consultas médicas, saques bancários, e que os dados registrados por eles vão dar o suporte aos sistemas de BI, a forma que os dados registrados por eles são organizados são em estruturas denominado de ER (entidade x relacionamento), onde utilizam-se de tabelas formadas por linhas (que representam os registros) e colunas (que representa os atributos), onde cada tabela representa uma entidade existente naquela sistema, e que se relacionam entre si. Já quando falamos em sistemas analíticos (OLAP), que são as ferramentas de BI, os dados são modelados utilizando-se o conceito de Fato e Dimensão.

3 TRABALHOS RELACIONADOS

Nessa seção serão expostos alguns trabalhos relacionados que explora o uso de ferramentas de Business Intelligence, Séries Temporais, Previsão de dados, Clusterização e Problema de Localização, relatando seus objetivos e resultados.

O estudo de Ribeiro, Oliveira and Pedrosa (2021) destaca a importância do uso de Business Intelligence (BI) na Administração Pública Portuguesa para analisar a eficiência, desempenho de serviços e prevenção de fraudes. O trabalho avalia e expõe vários projetos de Business Intelligence aplicado a Administração Pública que contribuiu para melhoria de desempenho nas atividades apoiada pela ferramenta, ratificando que o BI aplicado de forma correta trás grandes benefícios a Administração Pública, como também relatando que BI ainda é pouco utilizado no cenário público. O estudo considerou 15 (quinze) trabalhos encontrados nos repositórios Scopus e B-on, todos eles relacionados a administração pública e uso de Business Inteligence.

Para Khan et al. (2020) os sistemas de Business Intelligence têm atraído a atenção das empresas, profissionais e pesquisadores, proporcionando, dentro das organizações, uma melhoria no processo de suporte a tomada de decisões, assim como no desenvolvimento de produtos e serviços. Destaca a grande quantidade de dados geradas no mundo contemporâneo

e a necessidade de habilidades específica para análise desses dados. Fornece uma pesquisa relatando o impacto do uso de BI nas organizações.

Milán et al. (2020) apresenta uma revisão bibliográfica sobre o estado da arte do uso de Business Intelligence para o gerenciamento de performance nas corporações. Destaca que devido a grande quantidade de informações provenientes de vários tipos de fontes, internas e externas, tem dificultado o processo de tomada de decisão dentro das empresas, como também a medição de desempenho das empresas. Para Milán et al. (2020) o BI tem se mostrado uma ferramenta eficaz para resolver esse cenário mencionado, proporcionando rapidez ao acesso da informação por meio de dashboards, assim como seu uso pode proporcionar melhores resultados nas decisões corporativas.

Shi (2013) afirma a necessidade urgente do uso de sistemas de Business Intelligence (BI) para a tomada decisão no âmbito científico. O trabalho expõe um estudo realizado em uma empresa que utiliza BI na manutenção de energia nuclear, mostrando a importância dessa ferramenta na melhoria empresarial e realizando análise de três propostas de aplicações de BI.

Com foco na melhoria do desempenho da política energética e climática da União Europeia, Pinheiro et al. (2020) implementou uma solução de Business Intelligence, por meio de criação de Dashboard, para analisar a eficiência em Edifícios Habitacionais. Por meio de análise histórica e projeções até o ano de 2035, concluiu-se uma ineficiência energética nos edifícios habitacionais. A solução também proporcionou uma redução do valor da fatura de eletricidade e emissão de gases de efeito estufa.

Para Taylor and Letham (2017) a tarefa de prever, ou de realizar previsão, é comum na ciência de dados e ajuda as empresas nos seus planejamentos, definição de metas e detecção de fenômenos, porem muitos desafios estão associados a essa tarefa. Eles apresentam uma abordagem prática para realizar previsões em escala, por meio de um modelo de regressão modular parametrizados que podem ser ajustados intuitivamente por especialistas, e apresentam uma análise de desempenho de comparações de procedimentos de previsão para um melhor ajuste manual. Taylor e Letham (2017) ratificam que ferramentas que ajuda especialistas a usarem suas experiências de forma mais eficaz resultam em previsões mais confiáveis de séries temporais.

O estudo apresentado por Santos and Muller (2006) tem como objetivo apresentar o Problema de Localização de Máxima Cobertura com algumas adaptações, sugerindo novas restrições, para suprir a necessidade de cobertura de uma grande área geográfica de aproximadamente 5,08 milhões de quilômetros quadrados que tem que ser monitorados por facilidades, que possam atender aos clientes com maior rapidez. O estudo concluiu que sob determinadas circunstâncias as alterações propostas permitiram um maior realismo no modelo matemático utilizado em relação ao modelo do Problema de Localização de Máxima Cobertura original.

Oliveira (2020) apresentou a importância de um rápido atendimento a população por parte do Serviço de Atendimento Móvel de Urgência (SAMU), em que o tempo de resposta desse serviço é relevante para o atendimento ser bem-sucedido, e para isso é importante que as viaturas do SAMU estejam bem localizadas ao ponto de atender a maior quantidade de usuários. O trabalho tem como objetivo fornecer um conjunto de ferramentas que possa possibilitar um melhor posicionamento de viaturas do SAMU aplicando o Problema de Localização de Máxima Cobertura.

Sena and Nagwani (2015) enfatiza que a análise de serie temporais é uma das principais técnicas de previsão. O trabalho apresenta um estudo onde é aplicado o modelo ARIMA sobre a renda per capita na Alemanha Ocidental. A renda per capita é o valor recebido por pessoa após a dedução de impostos e sua previsão ajuda ao governo a avaliar as condições econômicas atual do país em comparação a outras economias do mundo, como também ajuda a avaliar a inflação. O trabalho desconsiderou os ajustes sazonais e foi obtido

um resultado com um erro relativo de 0,0225. Como o model ARIMA são baseados em autocorrelações, o preço atual previsto está linearmente relacionado ao preço anterior.

Ahmed et al. (2020) utilizam o Prophet, que é baseado em um modelo de aprendizado de máquinas, para prevê series temporais de energias de painéis fotovoltaico. A previsão confiável desses dados torna-se essencial para o planejamento da capacidade, em antecedência, para gerenciar com eficiência a distribuição de energia. Os dados gerados pelos painéis são submetidos ao modelo Prophet um dia antes para ser realizada a previsão de saída de energia. Concluiu-se que os dados coletados foram bastantes confiáveis.

Darapaneni et al. (2021) realizou um estudo referente a uma campanha de vacinação, na Índia, onde tinha com objetivo prever o tempo mínimo necessário para vacinação da população para alcançar a imunidade de rebanho. Nesse estudo foi utilizado modelo SIR, que define a capacidade de disseminação da doença, e posteriormente, por meio de análise de séries temporais usando Prophet foi possível prever a quantidade de dias necessários para vacinar a população suficiente para atingir a imunidade de rebanho.

No estudo apresentado por Jain and Prasad (2020) é relatado a importância do controle da medição de alguns parâmetros referente ao tráfego de redes utilizado na área de telecomunicação para prestar um serviço de melhor qualidade, permitindo que seja feito planejamento adequado os ativos de redes. O trabalho apresenta uma solução para esse cenário de previsão de parâmetros de performance de redes utilizando series temporais com o Prophet e o algoritmo XGBoost. É relatado pelos autores que o Prophet é um algoritmo capaz de realizar previsão utilizando grande quantidade de dados irregulares, ou anormal.

Kumar and Pande (2021) destaca que técnicas de previsão de dados são utilizadas em vários âmbitos, a exemplo de vendas, bancos, saúde, mercado de ações etc, para resolver os diversos problemas relacionados, pois prever ajuda na tomada de decisão. O trabalho destaca algumas ferramentas disponíveis com esse propósito, utilizando modelo de regressão e modelo exponencial logístico. É examinado alguns modelos de previsão, tais como o modelo aditivo, o modelo autorregressivo de média móvel integrada (ARIMA) e o modelo Phophet. O trabalho concluiu que o Phophet foi o modelo com melhor desempenho, pois apresentou um baixo índice de erro, melhor previsão e melhor ajuste.

Gong et al. (2020) apresentou estudo onde utilizou os algoritmos Prophet e ARIMA para prever consumo de energia elétrica em prédios, que pode variar de acordo com época do ano, economia, tipo de estabelecimento comercial, etc. O estudo apresenta uma comparação, em uma janela de tempo específica, dos dois algoritmos, de dados de series temporais reais de consumo de energia elétrica e dados gerados pelos Prophet e ARIMA. Com base na taxa da média relativa de erro com o resultado gerado pelos algoritmos, o autor concluiu que o Prophet foi o que apresentou o melhor resultado, em relação ao ARIMA.

Nesse cenário, destacando-se a importância dos temas supracitados, esse trabalho submete os dados de ocorrências policiais a um processo com objetivo identificar regiões através de técnicas de clusterização, criar séries temporais para prever dados de ocorrências por meio da API Prophet e definir a localização de patrulhas policiais para um melhor atendimento a população empregando o modelo matemático baseado no Problema de Localização de Máxima Cobertura Capacitada, dispondo todos esses dados em plataforma de Business Intelligence que irá subsidiar os gestores da área de segurança pública.

3 METODOLOGIA

Com o fito propor uma melhor gestão dos recursos policiais no combate efetivo à criminalidade a presente seção apresenta a metodologia utilizada para o desenvolvimento desse trabalho, que é baseada em três principais pilares, que juntos compõe a solução proposta, quais sejam:

- Clusterização de dados de ocorrências policiais resultando na identificação de regiões com incidência criminal;
- Previsão da quantidade de crimes por dia nas regiões identificadas;
- Alocação inteligente de quantidade e localização de viaturas policiais para melhor atender as regiões e crimes.

Inicialmente foram coletados dados de registros de ocorrências policiais gerados por meio do Sistema de Procedimento Policial (SPP), que se trata de um software presente nas unidades policiais que registram os crimes sofridos pela população, materializado por meio de Boletim de Ocorrências (BO). Os dados foram coletados em formato tabular, contendo 25 (vinte e cinco) colunas e mais de 300 mil registros de ocorrências policiais que ocorreram na cidade de João Pessoa-PB na janela de tempo de janeiro de 2017 à dezembro de 2021, ou seja, compreendendo um período de 5 (cinco) anos. Cada registro de ocorrência representa um crime ocorrido.

Entre os dados coletados existem informações referentes ao crime perpetrado e da vítima, e principalmente do dia e local do delito. Na figura 3 abaixo é possível observar como os dados foram dispostos.

Figura 3 – Descrição da figura

	Procedimento	Ocorrên	data_registro	data_ocorrencia	Cidade	Bairro	Tipo do Loc	Rua	hicial Ocor	Final Ocor	natureza	tipificacao	qualificaçã	de Nascim	Estado Civi	Sexo	ntação Se	Gênero
2	101201760	BO	2017-01-02 00:	02/01/2017	JOAO PES	VALENTIN	OUTROS	RUA CELIT	10:30	10:30	FATOS AT	PERDA OU	VITIMA	23/10/193	SOLTEIRO	FEMININC	HETEROSE	FEMININC
3	142012017	BO	12/02/2017	12/02/2017	JOAO PES	CABO BRA	VIA/LOCA	AV CABO	08:10	08:10	CRIMES CO	ART. 129 C	VITIMA	05/10/197	CASADO(A	MASCULIN	HETEROSE	MASCULIN
4	103701201	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	OUTROS	RUA VICEI	13:00	13:00	FATOS AT	OUTROS F	VITIMA	21/04/92	UNIAO ES	MASCULIN	HETEROSE	MASCULIN
5	182012017	BO	12/02/2017	12/02/2017	JOAO PES	CUIA	RESIDENC	RUA PRAI	12:00	12:00	LEI 11.340,	ART. 7º, IN	VITIMA	25/11/195	CASADO(A	MASCULIN	HETEROSE	MASCULIN
6	790012017	BO	12/02/2017	12/02/2017	JOAO PES	JAGUARIB	OUTROS	RUA ABER	17:32	17:32	FATOS AT	PERDA OU	VITIMA	16/07/196	CASADO(A	NAO INFC	NAO INFC	NAO INFC
7	791012017	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	VIA/LOCA	R.TEN. ALI	18:10	18:10	FATOS AT	OUTROS F	COMUNIC	07/02/196	CASADO(A	NAO INFC	NAO INFC	NAO INFC
8	156801201	BO	12/02/2017	12/02/2017	JOAO PES	JOSE AME	VIA/LOCA	RUA RADI	18:40	18:40	CRIMES CO	ART. 157 C	VITIMA	26/12/199	UNIAO ES	MASCULIN	NAO INFC	MASCULIN
9	795012017	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	VIA/LOCA	R. JUIZ AR	21:45	21:45	CRIMES CO	ART. 157 C	COMUNIC	22/03/197	CASADO(A	NAO INFC	NAO INFC	NAO INFC
10	187012017	BO	12/02/2017	12/02/2017	JOAO PES	MANGABE	RESIDENC	RUA	20:30	20:30	LEI 11.340,	ART. 7º, IN	VITIMA	25/12/199	UNIAO ES	MASCULIN	HETEROSE	MASCULIN
11	157301201	BO	13/02/2017	13/02/2017	JOAO PES	VALENTIN	ONIBUS IN	NAO INFC	07:30	07:30	CRIMES CO	ART. 155 C	VITIMA	21/09/199	UNIAO ES	FEMININC	NAO INFC	FEMININC
12	552012017	BO	13/02/2017	04/12/2016	JOAO PES	VALENTIN	VIA/LOCA	NAO INFC	08:00	08:00	FATOS AT	PERDA OU	VITIMA	29/12/196	CASADO(A	NAO INFC	NAO INFC	NAO INFC
13	158301201	BO	13/02/2017	10/02/2017	JOAO PES	VALENTIN	ORGAO PU	NAO INFC	17:00	17:00	CRIMES CO	ART. 157 C	VITIMA	09/01/199	UNIAO ES	FEMININC	HETEROSE	FEMININC

Fonte: Próprio Autor (2021)

De posse da base de dados de ocorrências policiais, foi possível observar algumas inconsistências e características que precisavam ser ajustadas para um melhor aproveitamento na solução proposta, a exemplo de campos vazios, formatos de dados incompatíveis, duplicação de dados, dados desnecessários, dentre outros, além da necessidade da inclusão de dois campos fundamentais que serão necessários para a aplicação: latitude e longitude.

Essa fase é denominada de Extração, Transformação e Business Intelligence (BI), onde foram realizadas as intervenções necessárias nos dados, estabelecendo convenções com a máxima cautela para não impactar em uma mudança de realidade nos quais os dados representam, e desenvolvimento do projeto de BI.

Ainda na fase de Extração, Transformação e Business Intelligence foram incluídos na base de ocorrências policiais os campos latitude e longitude. Para fins de preenchimento desses novos campos, foi utilizado a API do Googolemaps que, por meio dos endereços presentes nos registros de ocorrências, foi possível identificar a latitude e longitude do local de onde ocorreu o fato criminoso.

Com a fase de Extração, Transformação e Business Intelligence concluída, os dados estavam prontos para serem aplicados na solução composta pelos pilares definidos inicialmente.

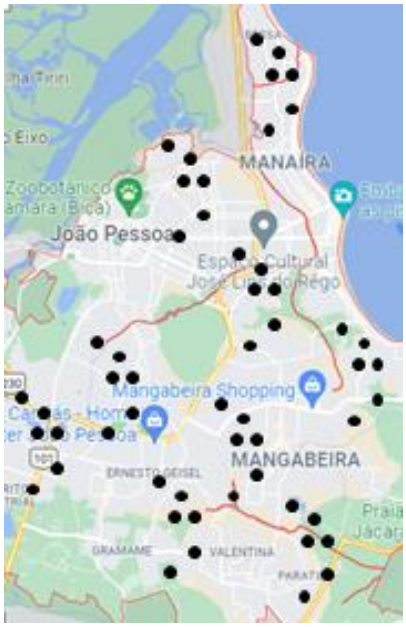
O primeiro pilar trata-se do procedimento de clusterização. Por meio dos dados de dia e local exato onde os crimes foram cometidos, latitude e longitude, e com objetivo de identificar as regiões e seus índices criminais, esses dados foram clusterizados, ou seja, submetidos a um algoritmo de clusterização. O processo de clusterização identifica e agrupa a maior quantidade de pontos semelhantes, nesse caso representados pelas coordenadas do local

do crime (latitude e longitude), e resultando na criação de regiões (clusters). Com essas regiões identificadas é possível visualizar as áreas com maior incidência criminal para uma maior atuação policial.

A clusterização torna-se importante porque é um processo que identifica áreas (clusters) onde se concentra o maior índice de criminalidade, independente das regiões formais presentes na sociedade (bairro, cidade, etc), dando-lhes outra perspectiva.

Na figura 4 é possível observar os locais de ocorrências policiais antes do processo de clusterização, onde todos os pontos estão dispostos de forma homogênea, já a figura 5 representa o resultado da clusterização dos dados e suas regiões (clusters) resultantes, propiciando de imediato uma melhor identificação com a delimitação formada.

Figura 4 – Locais de Ocorrência Policial antes da Clusterização



Fonte: Próprio Autor (2021)

Figura 5 – Locais de Ocorrência Policial após da Clusterização

Fonte: Próprio Autor (2021)

Após a clusterização cada registro de ocorrência policial presente na base de dados submetida ao processo passou a ser vinculado a uma região resultante, consequentemente se fez necessário a criação de um novo campo na base de dados, denominado cluster, que foi preenchido com o número da respectiva região a qual a ocorrência pertencia.

O próximo passo dentro da solução proposta trata-se do segundo pilar, responsável pela predição de crimes.

De posse dos dados dos dias e regiões resultante do processo de clusterização, foram construídas suas respectivas séries temporais. As séries temporais foram formadas dispondo os dados por região, dia e quantidade de crimes. É possível observar na figura 6 as regiões e quantidade de crimes presentes em um determinado dia.

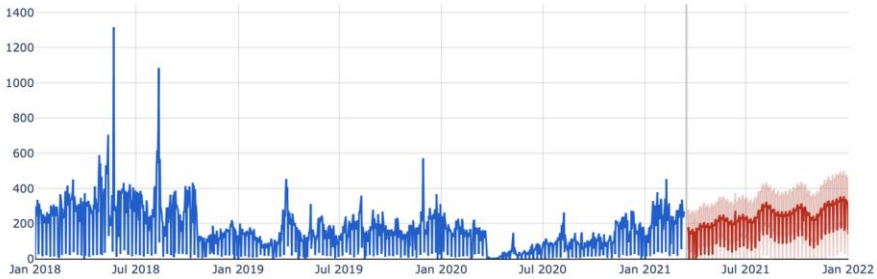
Figura 6 – Regiões clusterizadas com o quantitativo de crimes

Fonte: Próprio Autor (2021)

Com as séries temporais definidas, o próximo passo foi realizar a previsão da quantidade de crimes que poderiam ocorrer nas regiões identificadas por meio dessas séries. O processo de predição torna-se primordial para uma atuação mais assertiva no sentido do emprego de recursos policiais para o combate efetivo à criminalidade, tendo em vista ser mais apropriado ser demandado um maior efetivo policial para áreas com maior quantidade de crimes. Dessa forma, por meio do algoritmo de previsão API Prophet, foi possivel prever a quantidade de crimes por região (cluster) e dias, respectivamente.

A API Prophet recebeu como entrada séries temporais de dados históricos por meio de dois campos, quais sejam: data e quantidade. Após a realização do treinamento, o Prophet apresenta uma previsão de uma série temporal de dias predefinidos e quantidades futuras, conforme pode ser visto na figura 7. Na cor azul são apresentados os dados da série temporal real, que foi utilizada para treinamento do modelo, resultando na série temporal visto na cor vermelha, que se trata da previsão resultante.

Figura 7 – Regiões clusterizadas com o quantitativo de crimes



Fonte: Próprio Autor (2021)

O terceiro e último pilar da solução proposta trata-se da alocação inteligente de quantidade e localização de viaturas policiais nas regiões já definida. Na fase anterior foi possível prever a quantidade de crimes em cada uma dessas regiões por dia. Com base nesses dados é necessário definir a quantidade de viaturas (facilidades) necessárias e suas respectivas localização para que possam atender as ocorrências com maior rapidez e cobrir as áreas identificadas.

Com esse propósito, com as regiões e suas previsões de quantidade de ocorrências policiais já identificadas em dias específicos, estes dados foram submetidos ao modelo do Problema de Localização de Máxima Cobertura Capacitada proposto nesse trabalho. Esse modelo tem como objetivo cobrir o máximo de ocorrências policiais por cada viatura policial, dentro de um raio de distância pré-definido, e uma capacidade máxima de ocorrências. Dessa forma o modelo proposto resulta em uma melhor performance na localização e quantidade de viaturas para melhor atender as regiões conforme pode ser observado na figura 8.

Figura 8 – Regiões clusterizadas com o quantitativo de crimes

Fonte: Próprio Autor (2021)

Por fim é possível concluir que a solução proposta com seus respectivos pilares tem como objetivo: clusterizar ocorrências policiais, construir as séries temporais para prever a quantidade de ocorrências nessas regiões e definir a quantidade e localização de viaturas policiais para atender o máximo de ocorrências dentro de uma distância pré-definida. Na figura 9 é possível visualizar o fluxo da solução proposta.

Figura 9 – Fluxo da solução proposta

Fonte: Próprio Autor (2021)

4 CONSIDERAÇÕES FINAIS

O presente trabalho propõe uma solução para aplicação na área de gestão da segurança pública, onde por meio de técnicas envolvendo Business Intelligence, Séries Temporais, Problemas de Localização e Clusterização, apresenta uma ferramenta que busca utilizar registros de ocorrências policiais para fornecer ao gestor da segurança pública informações que possam subsidiar suas ações. Com a implementação da proposta os gestores terão acesso a um painel para melhor visualização de dados de ocorrências e consequentemente identificar tendencias, sazonalidades e comportamentos do fenômeno criminal, assim como previsão de crimes nas regiões clusterizadas e mapeadas, dessa forma otimizando o uso de patrulhas policiais para uma melhor cobertura das áreas mais afetadas com maior índice criminal, permitindo uma tomada de decisão mais assertiva por parte do gestor.

REFERÊNCIAS

AHMED S., Akter S., Islam K., Rahman M., Shawon M. H., “Forecasting PV Panel Output Using Prophet Time Series Machine Learning Model,” **2020 IEEE REGION 10 CONFERENCE (TENCON)** Osaka, Japan, November 16-19, 2020.

BRAGHITTONI, R. **Business intelligence:** Implementar do jeito certo e a custo zero. Casa do Código. 2015.

CAMPELLO, R.J.G.B. and Hruschka, E.R., “A fuzzy extension of the silhouette width criterion for cluster analysis”, **Fuzzy Set and System**, Vol.(157), Issue 21, 1 November 2006, p. 2858-2575. Elsevier.

CARVALHO, V. A. de C., 2011. **Política de segurança pública no Brasil:** avanços, limites e desafios Disponível em: https://www.scielo.br/j/rk/a/bnjfd8BgmpTSXSSSyXQ3qbj/?lang=pt. Acesso em: 25 jul. 2021.

CHARTIER, S. and Morissette, L. “The k-means clustering techinque: General considerations and implementation in Mathematica.”. **Tutorials in Quantitative Methods for Psychology 2013**, Vol. 9(1), p. 15-24.

CHURCH, R.; REVELLE, C. The maximal covering location problem. In: **SPRINGER. Papers of the Regional Science Association.** [S.l.], 1974. v. 32, n. 1, p. 101–118.

CRUZ, M. D., Ochi, L. S. O problema de clusterização automática: um novo método utilizando o ILS. **10th Brazilian Congress on Computacional Intelligence** (CBIC'2011)

DARAPANENI et al. (2021) "Forecasting Vaccination Drive In India for Herd Immunity using SIR and Prophet Model" **2021 IEEE World AI IoT Congress** (AIIoT), 2021.

DIAS, C. R. **Algoritmo evolutivo para o problema de clusterização de grafos orientados:** desenvolvimento e análise experimental. Dissertação de Mestrado. Universidade Federal Fluminense, 2004.

DIAS, C. R., Ochi, L. S., Soares, S. S. F., Clusterização em mineração de dados. Encontro Regional de Informática RJ/ES, 2004, Vitória. **Anais do ERI 2004 RJ/ES**, Vitória: ERI, 2004. p.1-6

DRUMMOND, L. M. A, Ochi, L. S, Rosário, S. S. "Um algoritmo de Construção e Busca Local para o Problema de Clusterização de Bases de Dados," **Sociedade Brasileira de Matemática Aplicada e Computacional**, Vol. 7. nº 1 (2006), 109-118.

EHLERS, R.S. **Análise de séries temporais**. Quarta Edição Publicada em 2007.

FACELI, K.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. de L. F. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FERREIRA, R. J. P. and Ferreira, H. L.,"Decision support system for location of back-up transformers based on a multi-attribute p-median model", **2012 IEEE International Conference on Systems, Man, and Cybernetics**, 2012.

FERRO, A. L. Inteligência de segurança pública e análise criminal. **Revista Brasileira de Inteligência.** Brasilia: Abin, v. 2, n. 2, abr. 2006.

GALVÃO, R. D., Chiyoshi, F. Y., Espejo, L. G. A. and Rivas, M. P. A., 2003. Disponível em: https://www.scielo.br/j/pope/a/xDj87ZWcysCqBsjqfCpw9sx/?lang=pt. Acesso em: 10 de junho. 2022.

GARCIA, C. A. F. M. Análise de Séries Temporais com Recurso a Técnicas de Bases de Dados. Dissertação de Mestrado. Universidade do Porto – Faculdade de Engenharia, 2000.

GONG et al. (2020) "Trend Analysis of Building Power Consumption Based on Prophet Algorithm" 2020 Asis Energy and Electrical Engineering Symposium, 2020.

HORNER, D., Resolução do Problema das P-Medianas Não Capacitado. Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2009.

ISLER, C. A., Bonassa, A. C. and Cunha, C. B., "Algoritmo genético para resolução do problema de p-medianas capacitado associado à distribuição de peças automotivas". Revista TRANSPORTES, v. 20, n. 2, 2012.

JAIN, G. and Prasad, R. R., “Machine learning, Prophet and XGBoost algorithm: Analysis of Traffic Forecasting in Telecom Networks with time series data” 2020 8th Internationa Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India. June 4-5, 2020.

KHAN et al. (2020), "Analysis of Business Intelligence Impact on Organizational Performance," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-4, doi: 10.1109/ICDABI51230.2020.9325610.

KUMAR J. B. and Pande S., "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 547-554, doi: 10.1109/ICCMC51019.2021.9418033.

LEAL, E. O. Técnicas operacionais de inteligência e ações de busca na produção de provas em investigação e processo criminal: admissibilidade e limites. 2016. 87 p. Trabalho de Conclusão de Curso (Especialização em Inteligência de Estado e Inteligência em Segurança Pública) – Centro Universitário Newton Paiva, Associação Internacional para Estudos de Segurança e Inteligência, Belo Horizonte, 2016. Disponível em: http://www.mpsp.mp.br/ portal/page/portal/ documentacao_e_divulgacao/doc_biblioteca/bibli_servicos_produtos/ BibliotecaDigital/Publicacoes_MP/Todas_publicacoes/Evandro-Ornelas-Leal-TCC.pdf. Acesso em: 15 julho. 2021

LORENA, L. A. N. and Senne, E. L. F., “Abordagens de Geração de Colunas para um Problema de P-Medianas Capacitado”, 2005.

MORETTIN, P. A. and Toloi C. M. C. Análise de séries temporais. São Paulo – Edgard Blucher, 2004.

MORETTIN, P. A. and Toloi C. M. C. Modelos para previsão de séries temporais. Rio de Janeiro – Instituto de Matemática Pura e Aplicada, 1981.

MILÁN et al. (2020), "Success factors and benefits of using business intelligence for corporate performance management," 2020 9th International Conference On Software Process Improvement (CIMPS), 2020, pp. 19-27, doi: 10.1109/CIMPS52057.2020.9390108.

OLIVEIRA, C. P. de. Modelos de Otimização Aplicados ao Problema de Máxima Cobertura: Estudo de Caso do Samu-BH. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, 2020.

PAL, A. and Prakash, P., Pratical Times Series Analysis. Packt Publishing Ltd – B3 2PB, UK, 2017.

PINHEIRO et al. (2020), "Business Intelligence Applied in Buildings Energy Efficiency," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9141019.

PONTIN, V. M., Garcia, R. A., Neto, P. B. and Ribeiro, G. M., Análise de Modelos Matemáticos para Problema Probabilístico de Localização-Alocação de Máxima Cobertura. Cadernos do Ime – Série Estatísticas. Universidade Federal do Rio de Janeiro, 2010.

PRIMAK, F. V. da S. Decisões com B.I (Business Intelligence). Rio de Janeiro: Editora Ciência Moderna Ltda., 2008.

PROPHET :< https://facebook.github.io/prophet/docs/quick_start.html#python-api>. Acesso em: 09 de jun. 2022.

RAMESH, Sharda,. Business Intelligence e Análise de Dados para Gestão do Negócio. Disponível em: Minha Biblioteca, (4th edição). Grupo A, 2019.

RIBEIRO R., Oliveira A. and Pedrosa I., "Analysis of the Impact of Business Intelligence in Public Administration," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-5, doi: 10.23919/CISTI52073.2021.9476489.

SANTOS, R. P. and Muller, C., "Problema de localização de máxima cobertura aplicado à localização de esquadrões de aeronaves de interceptação na região Amazônica," XXXVIII Simpoósio Brasileiro de Pesquisa Operacional, Goiânia-GO, 2006.

SANTOS, W. A., Nunes, E. R., Dias, B. R. Lucena and Pizzolato, N. D., Aplicação de modelo das p-medianas para a localização de unidades estratégicas de saúde da família ribeirinhas: um estudo de caso em uma localidade amazônica. Brazilian Journal of Development, 2020.

SENA, D. and Nagwani, N. K., "Application of Times Series Based Prediction Model to Forecast Per Capita Displosable Income" 2015 IEEE International Advance Computing Conference (IACC), 2015.

SHI Z., "Research and Application of Business Intelligence in the Nuclear Power Maintenance," 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2013, pp. 619-622, doi: 10.1109/IIH-MSP.2013.159.

SÖKÜT AÇAR T. and Ayman Öz N., "The determination of optimal cluster number by Silhouette index at clustering of the European Union member countries and candidate Turkey by waste indicators", Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, vol. 26, no. 3, pp. 481-487, Jun. 2020

TAYLOR S. J. and Letham B., "Forecasting at Scale". PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2.

WILLIAMS, F. Criminology. In: Bailey, W. G. The Encyclopedia of Police Science. 2. ed. New York, London: Garland. 1995. p. 181.