Trabalho de Conclusão de Curso

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

Aprendizado de Máquina e Visualização de Dados: Em Prol da Segurança de São Paulo

ALUNO: Rodrigo dos Santos Rebouças
ORIENTADOR: Marta Ribeiro Hentschke

PUCRS online uol edtech

Sumário

1. RESUMO

Diante dos aumentos de casos sobre furtos de veículos no munícipio de São Paulo, passa-se a cada dia ser mais necessário um controle e análise dos boletins de ocorrências registrados pela população.

Com a utilização de ferramentas de visualização, *Data Mining* e *Machine Learning*, esse artigo tem o objetivo de mostrar de forma analítica os pontos mais críticos por zona, uma análise exploratória de ocorrências de furtos de veículos pelo munícipio de São Paulo e a utilização de *Machine Learning* para a classificação de risco de cada logradouro da cidade.

O artigo utilizou dados de transparência do portal do governo de São Paulo, trabalhando com boletins de ocorrência gerados de furtos de veículos nos últimos 4 anos.

Palavras chaves: Mineração de Dados; Análise Exploratória; Segurança Pública; Boletim de Ocorrência; Visão Analítica; Aprendizado de Máquina; Inteligência Artificial.

ABSTRACT

Faced with the increase in cases of vehicle theft in the city of São Paulo, it becomes more and more necessary to control and analyze the police reports registered by the population.

Using visualization tools, Data Mining and Machine Learning, this article aims to analytically show the most critical points by zone, an exploratory analysis of vehicle thefts in the city of São Paulo and the use of Machine Learning for the risk classification of each street in the city.

The article used transparency database from the São Paulo government portal, working with police reports generated from vehicle thefts in the last 4 years.

Keywords: Data Mining; Exploratory Analysis; Public security; Occurrence Bulletin; Analytical Vision; Machine Learning; Artificial Intelligence.

2. INTRODUÇÃO

Segundo o portal de notícias G1 em 27/02/2022 “O furto de veículos no estado de São Paulo atingiu patamares maiores que no período pré-pandemia, segundo dados da Secretaria de Segurança Pública (SSP) [1].”

> **Art. 155** - Subtrair, para si ou para outrem, coisa alheia móvel: Pena - reclusão, de um a quatro anos, e multa. § 1º - A pena aumenta-se de um terço, se o crime é praticado durante o repouso noturno.

A ideia desse artigo é trabalhar com dados de boletins de ocorrência disponibilizado pelo Portal do governo | Cidadão São Paulo [2], utilizando tratamento e processamento de dados. Após isso será feito uma análise exploratória com o objetivo de buscar padrões na base de dados, em seguida criar um *Dataset* para ser usado em um modelo estatístico utilizando Algoritmos de *Machine Learning* supervisionada para a classificação de risco de furtos de veículos futuros nas ruas da cidade de São Paulo. Por último, será feito o *Data Mining* utilizando uma ferramenta de visualização para a construção de um *Dashboard [3]* mostrando de forma mais clara ao usuário o estudo e *insights* do artigo.

3. TRABALHOS RELACIONAIS

Esta seção apresenta trabalhos relacionados ao aqui proposto. Esses trabalhos utilizaram *Machine Learning* para realizar classificações de riscos em diferentes temas.

3.1 *Country-level Pandemic Risk*

Um estudo realizado pelo Jordan J. Bird et al. (2020) [4], mostrou uma estratégia de aprendizado de máquina com três estágios de classificação de risco com base em países que estavam relatando informações sobre o COVID-19. Foram criados 4 grupos de risco de países com base no risco de transmissão, risco de mortalidade e risco de incapacidade de testar. Os quatro grupos de risco foram rotulados como 'baixo', 'médio-baixo', 'médio-alto' e 'alto'. Foi utilizado então uma validação cruzada de deixar um país de fora para encontrar o modelo mais forte, produzindo uma pilha de algoritmos de aumento de gradiente e árvore de decisão para risco de transmissão, uma pilha de máquina de vetores de suporte e *Extra Trees* para risco de mortalidade e um algoritmo *Gradient Boosting* para o risco de incapacidade de testar.

O resultado desse trabalho mostrou que com 77,12%, o risco de incapacidade de testar é a métrica mais importante e possivelmente possibilitou a interpretação das outras duas métricas, ou seja, uma única “métrica de risco” que é calculada por meio das três métricas exploradas neste trabalho como problemas separados.

3.2 *Classification of Adolescent Psychiatric*

Em um outro estudo feito pelo Kyung-Won Kim et al. (2021) [5], foi analisado casos de pacientes com ideação suicida. O objetivo do trabalho foi buscar uma prevenção suicida com o uso dos algoritmos *Linear Regression*, *Random Forest*, *Artificial Neural Network*, *Support Vector Machine* e *Extreme Gradient Boosting* para 124 pacientes ambulatoriais psiquiátricos entre crianças e

adolescentes da Coreia. Foram criadas duas classificações sendo ‘risco-alto’ com ideação suicida >= 60 (SUI >= 60) e outro grupo como risco baixo para ideação suicida < 60 (SUI < 60).

A conclusão desse trabalho mostrou que embora o Inventário de Avaliação da Personalidade para Adolescentes avalie uma série de domínios, o resultado para risco alto foi de apenas 35,5%, o que foi sugerido ser necessário mais pesquisas para prever grupos de alto risco de suicídio.

3.3 *Tsunami Risk 3D*

No terceiro estudo relacionado ao proposto, foi realizado pelo M. B. Dholakia et al. (2013) [6], onde foi utilizado ferramentas de visualizações em ambiente GIS/CAD para prever risco de desastres por tsunami na costa Okha do distrito de Gujarat na Índia. Sobre os parâmetros de falha dos terremotos foram utilizados: área de falha (200 km de comprimento e 100 km de largura), ângulo de ataque, mergulho e deslizamento (270°, 15° e 90°), profundidade focal (10 km) e magnitude (8,0). Dividido em três classificações de risco como ‘Médio-Risco’, ‘Risco-Alto’ e ‘Risco-Muito-Alto’, a visualização em 3D desse trabalho conseguiu mostrar os pontos mais vulneráveis da costa Okha da Índia, onde como conclusão da visualização e da última tsunami de dezembro de 2004, é necessário que no futuro seja necessário focar mais na visualização 3D e na animação do tsunami risco ao longo da costa de Gujarat. Com a fusão dos dados de elevação SRTM com imagens de satélite e batendo uma modelagem auxiliada por computador, a modelagem baseada em GIS, medições de parâmetros marinhos por sismômetros do fundo do oceano e satélite, instalações de marégrafos, sistemas de detecção de tsunamis, utilizando sistemas convencionais e tradicionais conhecimentos, é possível desenvolver um plano adequado de gestão de desastres de tsunami.

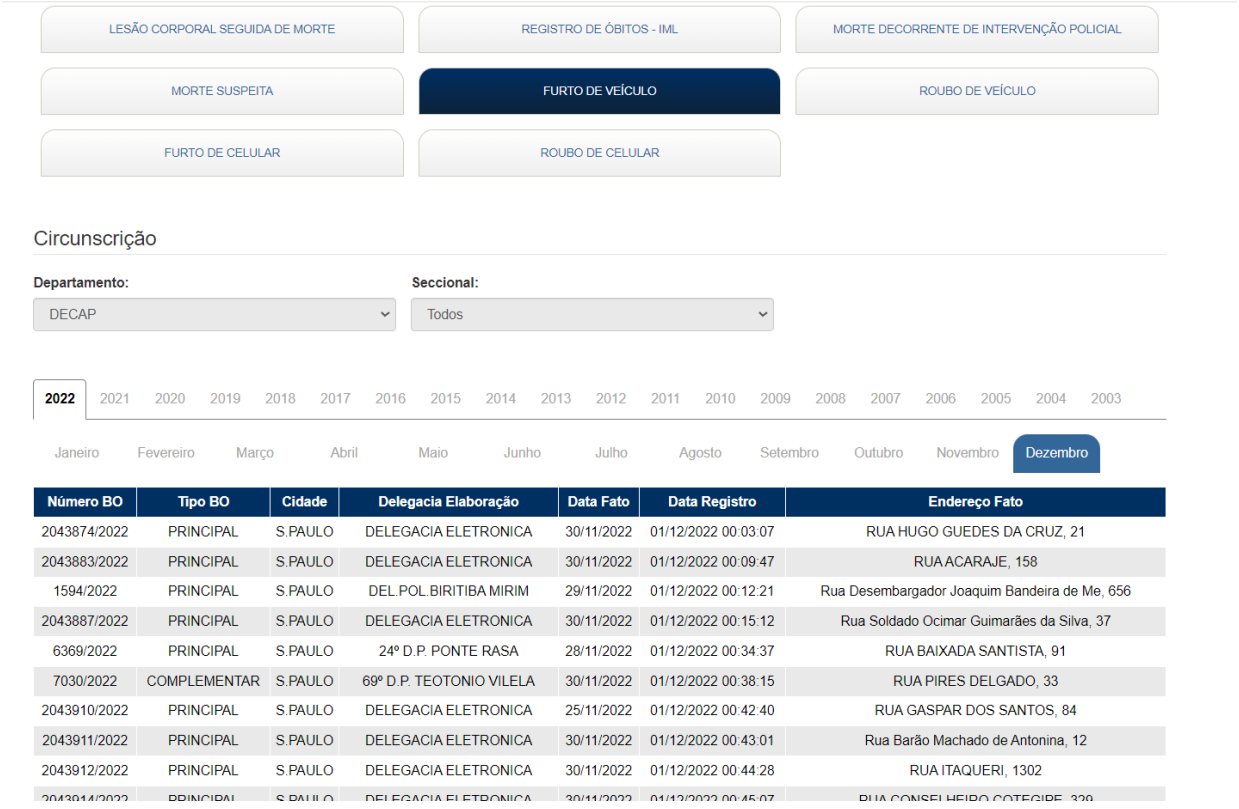
4. METODOLOGIA

Para esse artigo, foi utilizado a metodologia CRISP-DM [7] para buscar um melhor entendimento de todo o processo de mineração de dados.
Abaixo segue as fases baseadas no modelo e utilizadas no desenvolvimento:

4.1 Entendimento do negócio

Na fase de entendimento de negócio, foi realizada uma análise no portal do governo da Secretaria do Estado da Segurança Pública. Nessa análise, foram selecionada o tipo de crime: Furto de Veículo.

No entendimento das tabelas disponibilizadas pelo site, foi feito todo o mapeamento de qual seria a cidade, as zonas e departamentos policiais que seriam necessários para atingir o objetivo do artigo.



Número BO	Tipo BO	Cidade	Delegacia Elaboração	Data Fato	Data Registro	Endereço Fato
2043874/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	30/11/2022	01/12/2022 00:03:07	RUA HUGO GUEDES DA CRUZ, 21
2043883/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	30/11/2022	01/12/2022 00:09:47	RUA ACARAJE, 158
1594/2022	PRINCIPAL	S.PAULO	DEL.POL.BIRITIBA MIRIM	29/11/2022	01/12/2022 00:12:21	Rua Desembargador Joaquim Bandeira de Me, 656
2043887/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	30/11/2022	01/12/2022 00:15:12	Rua Soldado Ocimar Guimarães da Silva, 37
6369/2022	PRINCIPAL	S.PAULO	24º D.P. PONTE RASA	28/11/2022	01/12/2022 00:34:37	RUA BAIXADA SANTISTA, 91
7030/2022	COMPLEMENTAR	S.PAULO	69º D.P. TEOTONIO VILELA	30/11/2022	01/12/2022 00:38:15	RUA PIRES DELGADO, 33
2043910/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	25/11/2022	01/12/2022 00:42:40	RUA GASPAR DOS SANTOS, 84
2043911/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	30/11/2022	01/12/2022 00:43:01	Rua Barão Machado de Antonina, 12
2043912/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	30/11/2022	01/12/2022 00:44:28	RUA ITAQUERI, 1302

Tabela 1: Transparência de dados para extração – Portal do Governo | Cidadão SP.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

4.2 Entendimento dos dados

Na fase de entendimento dos dados, foram identificados os conjuntos mais interessantes como nome de bairros, logradouros, delegacias, data de ocorrência, período da ocorrência, marcas e modelos de veículos, espécies e rubricas de cada crime etc.

	ID_BO	BAIRRO	LOGRADOURO	DELEGACIA_CIRCUNSCRICAO	NOME_DELEGACIA_REG
4	5	BRASILANDIA	RUA ITATIBA DO SUL	72° D.P. VILA PENTEADO	DELEGACIA ELETRONICA
5	9	BELA VISTA	RUA FORTALEZA	05° D.P. ACLIMACAO	78° D.P. JARDINS
6	13	TREMEMBE	(null)	20° D.P. AGUA FRIA	20° D.P. AGUA FRIA
7	14	SAO MIGUEL	RUA LAURENTINO XAVIER DOS SANTOS	32° D.P. ITAQUERA	63° D.P. VILA JACUI
8	16	VILA GUILHERME	RUA FRANCISCO DUARTE	09° D.P. - CARANDIRU	20° D.P. AGUA FRIA
9	17	GRAJAU	(null)	101° D.P. JDIM IMBUIAS	85° D.P. JARDIM MIRNA
10	18	SAPOPEMBA	RUA CASTRO AVELAENS	41° D.P. VILA RICA	DELEGACIA ELETRONICA
11	19	LAPA	PRACA ANGELO RIVETTI	91° D.P. CEASA	DELEGACIA ELETRONICA
12	53401	TREMEMBE	(null)	73° D.P. JACANA	73° D.P. JACANA
13	53402	VILA FORMOSA	RUA ARAPACU	58° D.P. VILA FORMOSA	DELEGACIA ELETRONICA
14	53403	VILA PRINCESA IS...	RUA COMANDANTE CARLOS RUHL	44° D.P. GUAIANAZES	DELEGACIA ELETRONICA
15	53404	PIRITUBA	RUA CAMAROES	87° D.P. V. P. BARRETO	87° D.P. V. P. BARRETO
16	53405	ARICANDUVA	AVENIDA RIO DAS PEDRAS	41° D.P. VILA RICA	DELEGACIA ELETRONICA
17	25	SAPOPEMBA	AVENIDA CIPRIANO RODRIGUES	41° D.P. VILA RICA	69° D.P. TEOTONIO VILELA
18	53406	TATUAPE	RUA DOUTOR CORINTO BALDOINO COSTA	52° D.P. PARQUE S.JORGE	52° D.P. PARQUE S.JORGE
19	53407	VILA MARIA	RUA DONA MARIA QUEDAS	90° D.P. PQ. NOVO MUNDO	19° D.P. VILA MARIA

DATA_OCORRENCIA	PERIODO_OCORRENCIA	DESCR_MARCA_VEICULO	ESPECIE	RUBRICA
31/12/17	A TARDE	(null)	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
31/12/17	A NOITE	(null)	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
01/01/18	DE MADRUGADA	I/FIAT SIENA EL FLEX	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
31/12/17	A NOITE	(null)	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
01/01/18	DE MADRUGADA	FIAT/STRADA WORKING	Título VIII - Incolumidade pública (arts. 250 a 285)	Incêndio (art. 250, caput)
01/01/18	DE MADRUGADA	I/SUZUKI SX4 4WD	Título II - Patrimônio (arts. 155 a 183)	Furto qualificado (art. 155, §4o.) - RESIDENCIA
01/01/18	DE MADRUGADA	GM/CORSA GL	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
30/12/17	A TARDE	TROLLER/T4 TDI 3.0	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
01/01/18	DE MADRUGADA	HYUNDAI/HB20 1.0M 1.0 M	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
30/12/17	A TARDE	PEUGEOT/206 16 PRESEN FX	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
30/12/17	A TARDE	GM/MERIVA EXPRESSION	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
13/01/19	DE MADRUGADA	HONDA/CG 150 TITAN ES	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
14/01/19	A TARDE	I/FIAT SIENA EL 1.4 FLEX	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO
14/01/19	PELA MANHÃ	GM/CHEVETTE 1.6	Título II - Patrimônio (arts. 155 a 183)	Furto (art. 155) - VEICULO

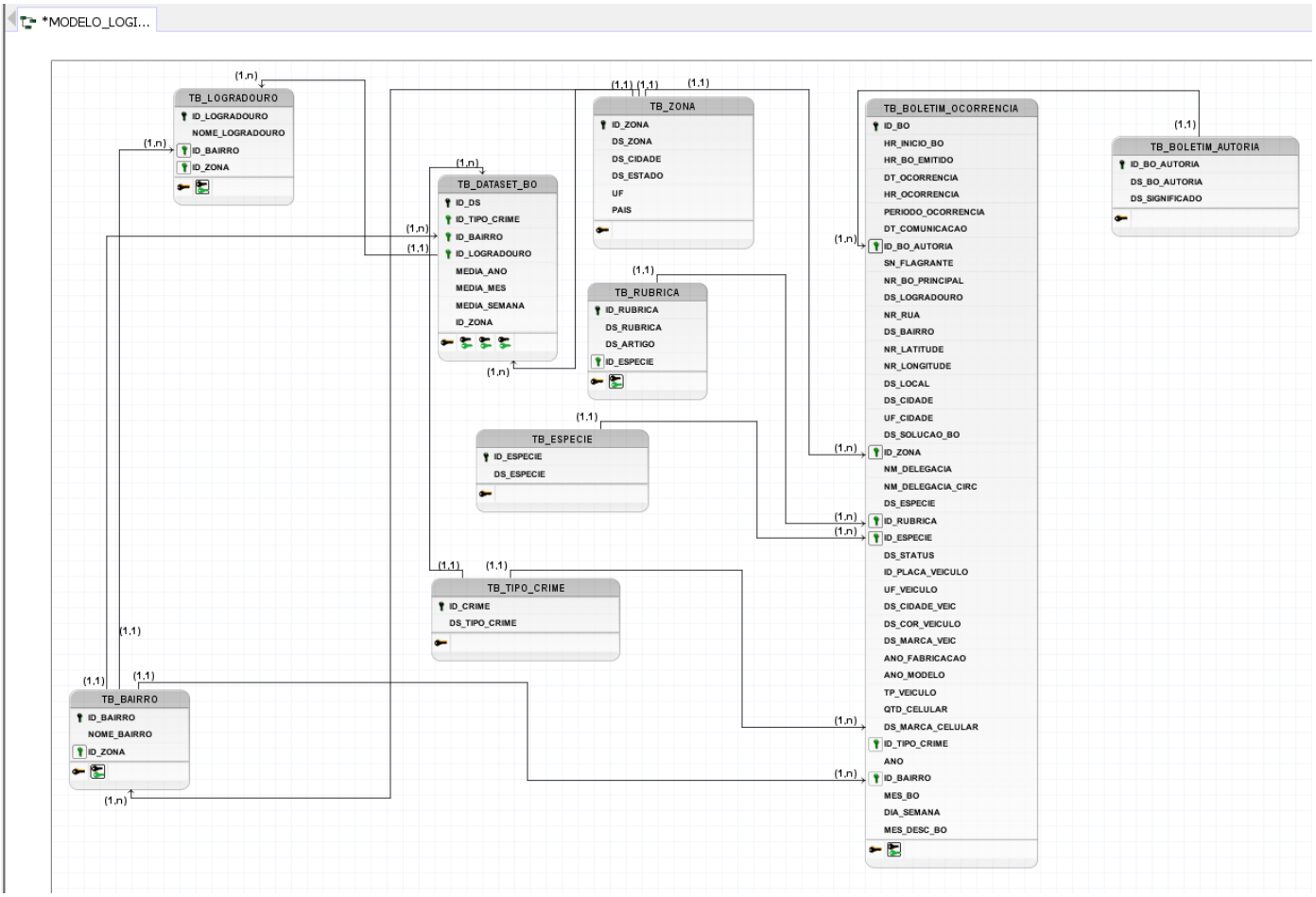
Tabela 2: Planilha do portal importada para o *SQL Developer* como tabela.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

4.3 Preparação dos dados

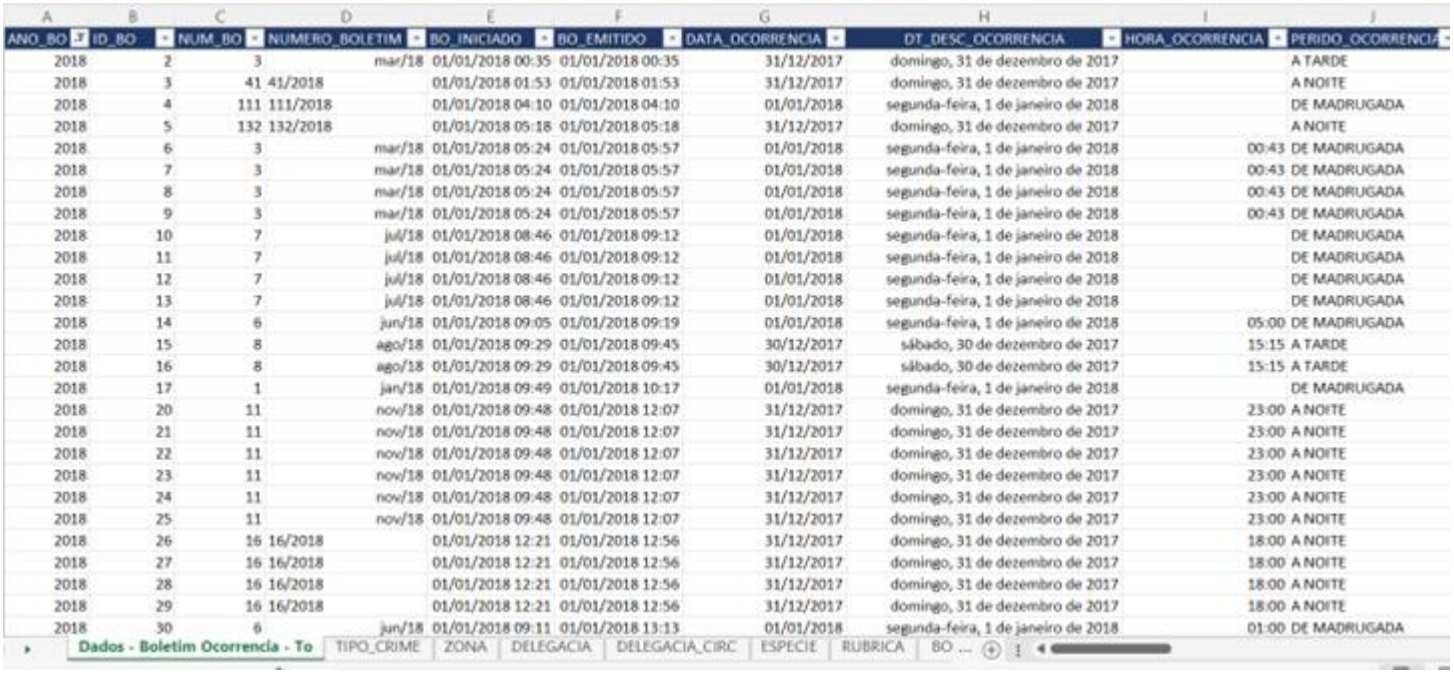
Para a preparação de dados, foram utilizadas as ferramentas BrModelo para formar a criação e relação de tabelas, *Excel* para extração e tratamentos iniciais dos dados, *SQL Developer* da Oracle para a manipulação, criação de objetos e estruturação final do *Dataset*, e *NotePad++* para a documentação dos *scripts* em *SQL*.

Esse processo ocorreu do começo ao fim do desenvolvimento, visando sempre tratar e trazer a informação, mas fidedigna possível a realidade dos fatos.

Nessa fase foram necessários tratamentos de linhas duplicadas, criação de tabelas relacionadas, inclusão de colunas ID para a inserção de índices na tabela, execução de scripts para transformação de caracteres de colunas, execução de scripts para ajustes de descrições visando a padronização dos dados e novas colunas mostrando outras informações detalhadas como DIA_SEMANA, MÊS_BO, ID_ZONA, ID_TIPO_CRIME, ID_BAIRRO etc.



Modelo Lógico 1: Modelo Lógico criado para os relacionamentos das tabelas.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).



ANO_BO	ID_BO	NUM_BO	NUMERO_BOLETIM	BO_INICIADO	BO_EMITIDO	DATA_OCORRENCIA	DT_DESC_OCORRENCIA	HORA_OCORRENCIA	PERIDO_OCORRENCIA
2018	2	3	mar/18	01/01/2018 00:35	01/01/2018 00:35	31/12/2017	domingo, 31 de dezembro de 2017		A TARDE
2018	3	41	41/2018	01/01/2018 01:53	01/01/2018 01:53	31/12/2017	domingo, 31 de dezembro de 2017		A NOITE
2018	4	111	111/2018	01/01/2018 04:10	01/01/2018 04:10	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	5	132	132/2018	01/01/2018 05:18	01/01/2018 05:18	31/12/2017	domingo, 31 de dezembro de 2017		A NOITE
2018	6	3	mar/18	01/01/2018 05:24	01/01/2018 05:57	01/01/2018	segunda-feira, 1 de janeiro de 2018	00:43	DE MADRUGADA
2018	7	3	mar/18	01/01/2018 05:24	01/01/2018 05:57	01/01/2018	segunda-feira, 1 de janeiro de 2018	00:43	DE MADRUGADA
2018	8	3	mar/18	01/01/2018 05:24	01/01/2018 05:57	01/01/2018	segunda-feira, 1 de janeiro de 2018	00:43	DE MADRUGADA
2018	9	3	mar/18	01/01/2018 05:24	01/01/2018 05:57	01/01/2018	segunda-feira, 1 de janeiro de 2018	00:43	DE MADRUGADA
2018	10	7	jul/18	01/01/2018 08:46	01/01/2018 09:12	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	11	7	jul/18	01/01/2018 08:46	01/01/2018 09:12	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	12	7	jul/18	01/01/2018 08:46	01/01/2018 09:12	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	13	7	jul/18	01/01/2018 08:46	01/01/2018 09:12	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	14	6	jun/18	01/01/2018 09:05	01/01/2018 09:19	01/01/2018	segunda-feira, 1 de janeiro de 2018	05:00	DE MADRUGADA
2018	15	8	ago/18	01/01/2018 09:29	01/01/2018 09:45	30/12/2017	sábado, 30 de dezembro de 2017	15:15	A TARDE
2018	16	8	ago/18	01/01/2018 09:29	01/01/2018 09:45	30/12/2017	sábado, 30 de dezembro de 2017	15:15	A TARDE
2018	17	1	jan/18	01/01/2018 09:49	01/01/2018 10:17	01/01/2018	segunda-feira, 1 de janeiro de 2018		DE MADRUGADA
2018	20	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	21	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	22	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	23	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	24	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	25	11	nov/18	01/01/2018 09:48	01/01/2018 12:07	31/12/2017	domingo, 31 de dezembro de 2017	23:00	A NOITE
2018	26	16	16/2018	01/01/2018 12:21	01/01/2018 12:56	31/12/2017	domingo, 31 de dezembro de 2017	18:00	A NOITE
2018	27	16	16/2018	01/01/2018 12:21	01/01/2018 12:56	31/12/2017	domingo, 31 de dezembro de 2017	18:00	A NOITE
2018	28	16	16/2018	01/01/2018 12:21	01/01/2018 12:56	31/12/2017	domingo, 31 de dezembro de 2017	18:00	A NOITE
2018	29	16	16/2018	01/01/2018 12:21	01/01/2018 12:56	31/12/2017	domingo, 31 de dezembro de 2017	18:00	A NOITE
2018	30	6	jun/18	01/01/2018 09:11	01/01/2018 13:13	01/01/2018	segunda-feira, 1 de janeiro de 2018	01:00	DE MADRUGADA

Dados - Boletim Ocorrencia - To | TIPO_CRIME | ZONA | DELEGACIA | DELEGACIA_CIRC | ESPECIE | RUBRICA | BO ...

Tabela 3: Planilha *Excel* com as relações das tabelas antes da importação.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

```
-------------------------------------------------------------------------------------------------------------
10° Passo
--UPDATE PARA TORNAR TODOS OS LOGRADOUROS EM MAIUSCULO
UPDATE TB_BOLETIM_OCORRENCIA BO
SET BO.LOGRADOURO = UPPER(BO.LOGRADOURO)


-------------------------------------------------------------------------------------------------------------
11° Passo
--UPDATE PARA TORNAR TODOS OS BAIRROS EM MAÍUSCULO
UPDATE TB_BOLETIM_OCORRENCIA BO
SET BO.LOGRADOURO = UPPER(BO.LOGRADOURO)

-------------------------------------------------------------------------------------------------------------
12° Passo
--UPDATE PARA TORNAR TODOS DESCRIÇÃO LOCAL EM MAÍUSCULO
UPDATE TB_BOLETIM_OCORRENCIA BO
SET BO.descricao_local = UPPER(BO.descricao_local)
-------------------------------------------------------------------------------------------------------------
13° Passo
--UPDATE PARA REMOVER CARACTERES ESPECIAIS DOS BAIRROS
UPDATE TB_BOLETIM_OCORRENCIA BO
SET BO.BAIRRO =  TRANSLATE(BO.BAIRRO,
                 'ŠŽšžŸÁÇÉÍÓÚÀÈÌÒÙÂÊÎÔÛÃÕËÜÏÖÑÝåáçéíóúàèìòùâêîôûãõëüïöñýý',
                 'SZszYACEIOUAEIOUAEIOUAOEUIONYaaceiouaeiouaeiouaoeuionyy')
-------------------------------------------------------------------------------------------------------------
14° Passo
--UPDATE PARA REMOVER CARACTERES ESPECIAIS DOS LOGRADOUROS
UPDATE TB_BOLETIM_OCORRENCIA BO
SET BO.LOGRADOURO =  TRANSLATE(BO.LOGRADOURO,
                 'ŠŽšžŸÁÇÉÍÓÚÀÈÌÒÙÂÊÎÔÛÃÕËÜÏÖÑÝåáçéíóúàèìòùâêîôûãõëüïöñýý',
                 'SZszYACEIOUAEIOUAEIOUAOEUIONYaaceiouaeiouaeiouaoeuionyy')
-------------------------------------------------------------------------------------------------------------
15° Passo
--UPDATE PARA REMOVER CARACTERES ESPECIAIS DOS DESCRICAO_LOCAL
  UPDATE TB_BOLETIM_OCORRENCIA BO
                 SET BO.descricao_local =  TRANSLATE(BO.descricao_local,
                 'ŠŽšžŸÁÇÉÍÓÚÀÈÌÒÙÂÊÎÔÛÃÕËÜÏÖÑÝåáçéíóúàèìòùâêîôûãõëüïöñýý',
                 'SZszYACEIOUAEIOUAEIOUAOEUIONYaaceiouaeiouaeiouaoeuionyy')

-------------------------------------------------------------------------------------------------------------
```

***Script 1*:** Scripts de tratamento de dados da tabela TB_BOLETIM_OCORRENCIA
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022)

```sql
-- VIEW DE CONSULTA DE DADOS PARA INSERÇÃO NA TB_DATASET_BO
CREATE OR REPLACE VIEW VW_CONSULTA_DATASET AS

SELECT DISTINCT  BO.ID_TIPO_CRIME,
                 BO.BAIRRO,
                 BO.LOGRADOURO,
                (SELECT SUBSTR(COUNT(BO1.ID_BO)/4,1,4)
                 FROM  TB_BOLETIM_OCORRENCIA BO1
                 WHERE BO1.LOGRADOURO = BO.LOGRADOURO
                 AND   BO1.BAIRRO     = BO.BAIRRO
                )MEDIA_ANO,
                (SELECT SUBSTR(COUNT(BO1.ID_BO)/48,1,4)
                 FROM  TB_BOLETIM_OCORRENCIA BO1
                 WHERE BO1.LOGRADOURO = BO.LOGRADOURO
                 AND   BO1.BAIRRO     = BO.BAIRRO
                )MEDIA_MES,

                (SELECT SUBSTR(COUNT(BO2.ID_BO)/208,1,4)
                 FROM TB_BOLETIM_OCORRENCIA BO2
                 WHERE BO2.LOGRADOURO = BO.LOGRADOURO
                 AND   BO2.BAIRRO     = BO.BAIRRO
                 ) MEDIA_SEMANA,

                 (SELECT COUNT(BO2.ID_BO)
                 FROM TB_BOLETIM_OCORRENCIA BO2
                 WHERE BO2.LOGRADOURO = BO.LOGRADOURO
                 AND   BO2.BAIRRO     = BO.BAIRRO
                 ) QTD_TOTAL_BO,

                 ID_ZONA

FROM TB_BOLETIM_OCORRENCIA BO
WHERE BO.LOGRADOURO IS NOT NULL
GROUP BY LOGRADOURO,
      BO.DATA_OCORRENCIA,
      BO.DIA_SEMANA,
      BO.BAIRRO,
      BO.MES_BO,
      BO.ID_TIPO_CRIME
      ORDER BY 2,3 ASC
      ;
```

***Script 2*:** Criação de *View* para uso do cursor.

```sql
-- CURSOR DE INSERT NA TABELA TB_DATASET_BO

DECLARE

CURSOR C_INSERT_DADOS_DATASET_BO IS
SELECT * FROM VW_CONSULTA_DATASET VW
WHERE VW.ID_ZONA = 5;

BEGIN

FOR I_TB_DATASET IN C_INSERT_DADOS_DATASET_BO LOOP

INSERT INTO TB_DATASET_BO
(
 ID_DS,
 ID_TIPO_CRIME,
 BAIRRO,
 LOGRADOURO,
 MEDIA_ANO,
 MEDIA_MES,
 MEDIA_SEMANA,
 CLASSIF_DE_RISCO
)
VALUES
(SEQ_ID_DATASET_BO.NEXTVAL,
 I_TB_DATASET.ID_TIPO_CRIME,
 I_TB_DATASET.BAIRRO,
 I_TB_DATASET.LOGRADOURO,
 I_TB_DATASET.MEDIA_ANO,
 I_TB_DATASET.MEDIA_MES,
 I_TB_DATASET.MEDIA_SEMANA,
 I_TB_DATASET.CLASSIF_DE_RISCO
);

      COMMIT;

  END LOOP;

END;
```

***Script 3*:** Execução do cursor para alimentar o *Dataset*.

```sql
SELECT * FROM TB_DATASET_BO
```

Saída do Script | Resultado da Consulta

SQL | 50 linhas extraídas em 0,008 segundos

	ID_DS	ID_TIPO_CRIME	BAIRRO	LOGRADOURO	MEDIA_ANO	MEDIA_MES	MEDIA_SEMANA	ID_ZONA
1	1	1	ACLIMACAO	RUA MUNIZ DE SOUSA	2	0,166	0,038	3
2	2	1	ACLIMACAO	AVENIDA TURMALINA	0,5	0,041	0,009	3
3	3	1	ACLIMACAO	RUA DOUTOR RAFAEL CARAMURU LANZEL...	0,5	0,041	0,009	3
4	4	1	ACLIMACAO	RUA ANTONIO TAVARES	0,25	0,02	0,004	3
5	5	1	ACLIMACAO	RUA MARANDUBA	0,25	0,02	0,004	3
6	6	1	ACLIMACAO	RUA ITATINS	0,25	0,02	0,004	3
7	7	1	ACLIMACAO	RUA BUENO DE ANDRADE	0,75	0,062	0,014	3
8	8	1	ACLIMACAO	RUA SEBASTIAO CARNEIRO	0,25	0,02	0,004	3
9	9	1	ACLIMACAO	RUA BATISTA CAETANO	1,75	0,145	0,033	3
10	10	1	ACLIMACAO	RUA GUALACHOS	0,5	0,041	0,009	3
11	11	1	ACLIMACAO	RUA PAES DE ANDRADE	0,5	0,041	0,009	3
12	12	1	ACLIMACAO	RUA ALBINA BARBOSA	1	0,083	0,019	3
13	13	1	ACLIMACAO	RUA ANADIA	0,25	0,02	0,004	3
14	14	1	ACLIMACAO	RUA ESPIRITO SANTO	0,5	0,041	0,009	3
15	15	1	ACLIMACAO	RUA MAZZINI	0,25	0,02	0,004	3
16	16	1	ACLIMACAO	RUA VERGUEIRO	0,5	0,041	0,009	3
17	17	1	ACLIMACAO	RUA DR. NICOLAU DE SOUZA QUEIROZ	0,25	0,02	0,004	3
18	18	1	ACLIMACAO	RUA TAMADARE	0,25	0,02	0,004	3
19	19	1	ACLIMACAO	RUA NILO	7,5	0,625	0,144	3
20	20	1	ACLIMACAO	PRACA ROSA ALVES DA SILVA	0,25	0,02	0,004	3
21	21	1	ACLIMACAO	RUA ALABASTRO	0,25	0,02	0,004	3
22	22	1	ACLIMACAO	RUA APENINOS	1	0,083	0,019	3
23	23	1	ACLIMACAO	AVENIDA DA ACLIMACAO	0,5	0,041	0,009	3
24	24	1	ACLIMACAO	RUA PEDRA AZUL	0,25	0,02	0,004	3
25	25	1	ACLIMACAO	RUA MARACAI	1,5	0,125	0,028	3
26	26	1	ACLIMACAO	RUA DOUTOR FELIX	1	0,083	0,019	3

Tabela 4: Tabela TB_DATASET_BO criada.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

4.4 Modelagem

Na etapa de Modelagem, foram trabalhados os dados da tabela TB_DATASET_BO, com o objetivo de classificar o risco de cada logradouro inserido no *Dataset*.

Na criação do *Dataset*, foi realizado a preparação de cada parâmetro necessário para o resultado dessa classificação, como mostra na imagem acima.

Para o uso do modelo, foi realizado a média anual, média mensal e média semanal de boletins de ocorrência registrados por logradouro de cada bairro.

Com isso, o modelo de *Machine Learning* deverá classificar o risco de cada uma das 37.681 linhas distintas do *Dataset*.

4.5 Avaliação

Na fase de Avaliação, foi trabalhado o processo de criação do modelo, com o objetivo de selecionar o modelo com maior acurácia entre os algoritmos mais utilizados para esse tipo de classificação: *K-Nearest Neighbors* [8], *Support Vector Machine [9]* e *Random Forest [10]*. Toda avaliação do modelo foi realizada na linguagem *Python* via plataforma *Google Colab*.

Antes do treinamento dos modelos, ainda no *Oracle SQL Developer*, foi tratado a coluna de classificação do *Dataset* (Coluna Risco), onde foi utilizada para as classificações de risco de cada logradouro do *Dataset*.

Para a criação dessa coluna, foi necessário a utilização da função da fórmula de Desvio Padrão, utilizando os parâmetros de Bairro, Logradouro, média por ano, média por mês e média por semana, retornando assim um valor de porcentagem que foi utilizado para cada classificação de risco.

Na criação da classificação dos riscos, foram considerados 5 tipos de classificação, onde trouxeram as seguintes condições:

```
Risco Muito Baixo <= 4.99
Risco Baixo       Entre 5  a 20.99
Risco Médio       Entre 21 a 55.99
Risco Alto        Entre 56 a 90.99
Risco Muito Alto  >= 91
```

```sql
CREATE OR REPLACE
FUNCTION FNC_CALCULA_PROB_RISCO (P_MEDIA_ANO    NUMBER,
                                 P_MEDIA_MES    NUMBER,
                                 P_MEDIA_SEMANA NUMBER,
                                 P_ID_ZONA      NUMBER,
                                 P_BAIRRO       VARCHAR2,
                                 P_LOGRADOURO   VARCHAR2)
RETURN VARCHAR2 IS

RISCO     VARCHAR2(80);

BEGIN

SELECT CLASS_RISCO  INTO RISCO

FROM(
SELECT CASE WHEN DESVIOPAD <= 4.99
            THEN 'RISCO MUITO BAIXO'
            WHEN DESVIOPAD BETWEEN 5 AND 20.99
            THEN 'RISCO BAIXO'
            WHEN DESVIOPAD BETWEEN 21 AND 55.99
            THEN 'RISCO MEDIO'
            WHEN DESVIOPAD BETWEEN 56 AND 90.99
            THEN 'RISCO ALTO'
            WHEN DESVIOPAD >= 91
            THEN 'RISCO MUITO ALTO'
            END  CLASS_RISCO
FROM(
-- CÁLCULO DA RAIZ QUADRADA DA SOMATÓRIA DIVIDIDA PELAS DISTÂNCIAS

SELECT SQRT(DN) DESVIOPAD
FROM(
-- CÁLCULO DA SOMATÓRIA DIVIDIDO PELO NÚMERO DE DISTÂNCIAS

SELECT SOMATORIO/2  DN
FROM(
-- CÁLCULO DA SOMATÓRIA

SELECT D1+D2+D3 SOMATORIO
FROM
(
-- CÁLCULO DAS DISTÂNCIAS DA MÉDIA


SELECT (((P_MEDIA_ANO)    - MEDIA) * ((P_MEDIA_ANO)      - MEDIA)) D1,
       (((P_MEDIA_MES)    - MEDIA) * ((P_MEDIA_MES)      - MEDIA)) D2,
       (((P_MEDIA_SEMANA)- MEDIA) * ((P_MEDIA_SEMANA)   - MEDIA)) D3

FROM
(
-- CÁLCULO MÉDIA (TRAZENDO A TABELA E VALIDANDO OS BAIRROS E LOGRADOUROS)

SELECT (P_MEDIA_ANO + P_MEDIA_MES + P_MEDIA_SEMANA )/3 MEDIA
       FROM TB_DATASET_BO DS
       WHERE DS.BAIRRO       = P_BAIRRO
         AND DS.LOGRADOURO   = P_LOGRADOURO
         AND DS.ID_ZONA      = P_ID_ZONA
         AND ROWNUM = 1
)
 )
  )
   )
    )
     );

      RETURN(RISCO);

        END;
```

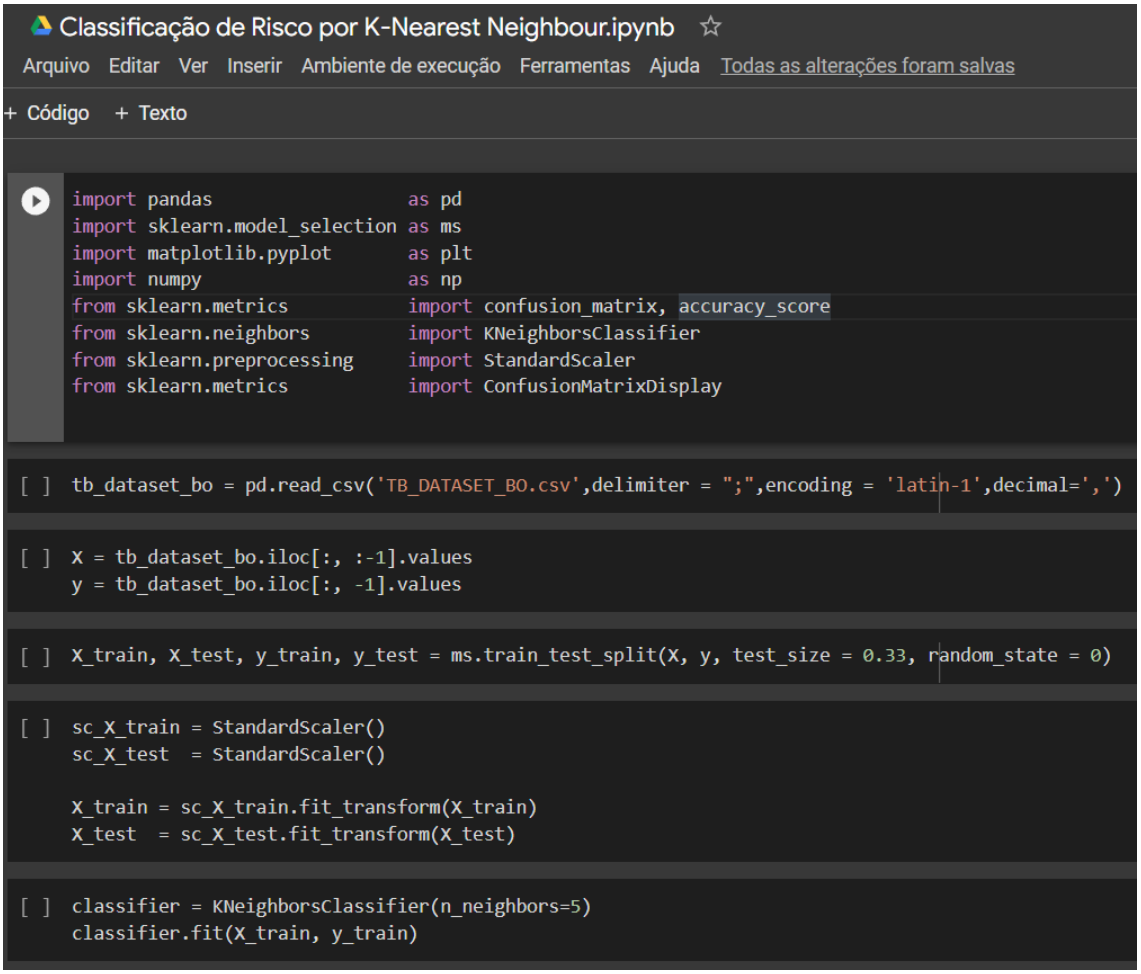
Script 4: Função de desvio padrão de Pearson, KP 1857-1936 (transcrevida para o *SQL Developer*).

$$\mathrm{DP}_{\mathrm{amostra}} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n - 1}}$$

Função de desvio padrão de Pearson, KP (1857 - 1936).

Após a criação da coluna RISCO, foram utilizados os 3 modelos de algoritmos com o objetivo de buscar a melhor acurácia entre os 3 modelos.

Para a avaliação do modelo utilizando o algoritmo *KNN*, foi utilizado a distância Euclidiana [11]:



Script 5: Classificação de Risco utilizando o algoritmo *KNN*.

$$\sqrt{\sum_{i=0}^{n-1}(a_i - b_i)^2}$$

Função de distância Euclidiana de Alexandria, EA (330 a.C).

No Algoritmo *SVM*, foi utilizado a função de Kernel RBF (Kernel Gaussiano) [12]:

Classificação de Risco por SVM.ipynb ☆

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alterações foram salvas

+ Código + Texto

```
import pandas                  as pd
import numpy                   as np
import matplotlib.pyplot       as plt
import seaborn                 as sns
import sklearn.model_selection as ms

from sklearn.metrics       import confusion_matrix, accuracy_score
from sklearn.svm           import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.metrics       import ConfusionMatrixDisplay
```

```
[ ] tb_dataset_bo = pd.read_csv('TB_DATASET_BO.csv',delimiter = ";",encoding = 'latin-1',decimal=',')
```

```
[ ] X = tb_dataset_bo.iloc[:, :-1].values
    y = tb_dataset_bo.iloc[:, -1].values
```

```
[ ] X_train, X_test, y_train, y_test = ms.train_test_split(X, y, test_size = 0.33, random_state = 0)
```

```
[ ] sc_X_train = StandardScaler()
    sc_X_test  = StandardScaler()

    X_train = sc_X_train.fit_transform(X_train)
    X_test = sc_X_test.fit_transform(X_test)
```

```
[ ] classifier = SVC(kernel='rbf')
    classifier.fit(X_train, y_train)
```

Script 6: Classificação de Risco utilizando o algoritmo *SVM.*

$$K(x, y) = e^{-}(\gamma||x - y||^2)$$

Função de Kernel RBF de Gauss, JCFG (1777 - 1885).

Por último, foi utilizado o algoritmo *Random Forest*, com o parâmetro de *N_estimators* = 1000 [13]:

Classifacação de Risco por Random Forest Classifier.ipynb ☆

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alterações foram salvas

+ Código + Texto

```
[ ] import pandas            as pd
    import numpy             as np
    import matplotlib.pyplot as plt
    import seaborn           as sns
    from sklearn.ensemble    import RandomForestClassifier
    from sklearn.metrics     import accuracy_score
    %matplotlib inline
```

```
[ ] tb_dataset_bo = pd.read_csv('TB_DATASET_BO.csv',delimiter = ";",encoding = 'latin-1',decimal=',')
```

```
    tb_dataset_bo.head(10)
```

```
[ ] tb_dataset_bo.dtypes
```

```
[ ] tb_dataset_bo['RISCO'].value_counts()
```

```
[ ] X = tb_dataset_bo.drop(['RISCO'], axis=1)
    Y = tb_dataset_bo['RISCO']
```

```
[ ] from sklearn.model_selection import train_test_split

    X_train, X_test = train_test_split(X, test_size = 0.33, random_state = 42)
    Y_train, Y_test = train_test_split(Y, test_size = 0.33, random_state = 42)
```

```
[ ] from sklearn.ensemble    import RandomForestClassifier
    rfc = RandomForestClassifier(n_estimators = 1000, random_state = 0)
```

Script 7: Classificação de Risco utilizando o algoritmo *Random Forest.*

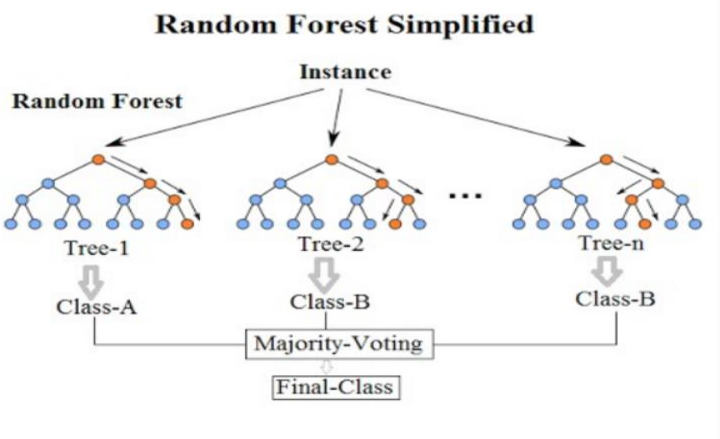


Diagrama simplificado da *Random Forest.*

A acurácia dos resultados de todos os modelos será informada no item 5 - Resultado.

4.6 Implantação

Na Fase de implantação, foi utilizado a ferramenta de visualização Power B.I, onde teve o objetivo de mostrar todo o estudo do artigo de uma forma mais clara, trazendo também análises exploratória e o resultado dos riscos gerados pelo algoritmo de melhor acurácia avaliado como melhor modelo.

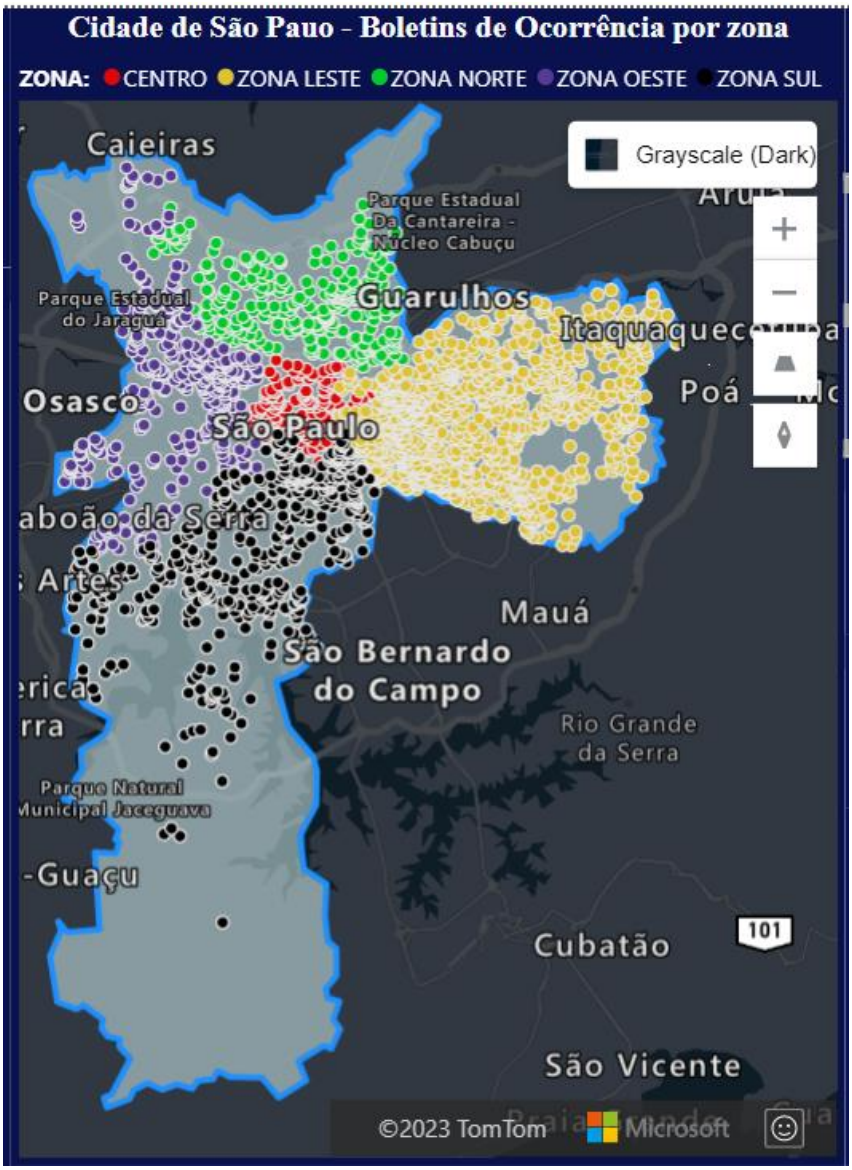


Gráfico 1: Cidade de São Paulo, com a quantidade de boletins divididos por zona gerada via Power B.I.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

Os detalhes de cada gráfico e resultados do artigo serão detalhados no item 5 - Resultado.

5. RESULTADOS

O trabalho foi dividido em duas partes, sendo a primeira uma análise exploratória dos dados utilizando a ferramenta *Power B.I* da *Microsoft* utilizando os boletins de ocorrência extraídos de 2018 a 2022.

Nesse trabalho, foi possível identificar padrões, relações, diferenças por datas, zonas, bairro, tipo de crime, veículos, marcas etc.

5.1 Visão Geral do *Dashboard*

Na primeira parte da visualização, foi mostrado uma visão geral dos furtos de carro, mostrando as seguintes análises:

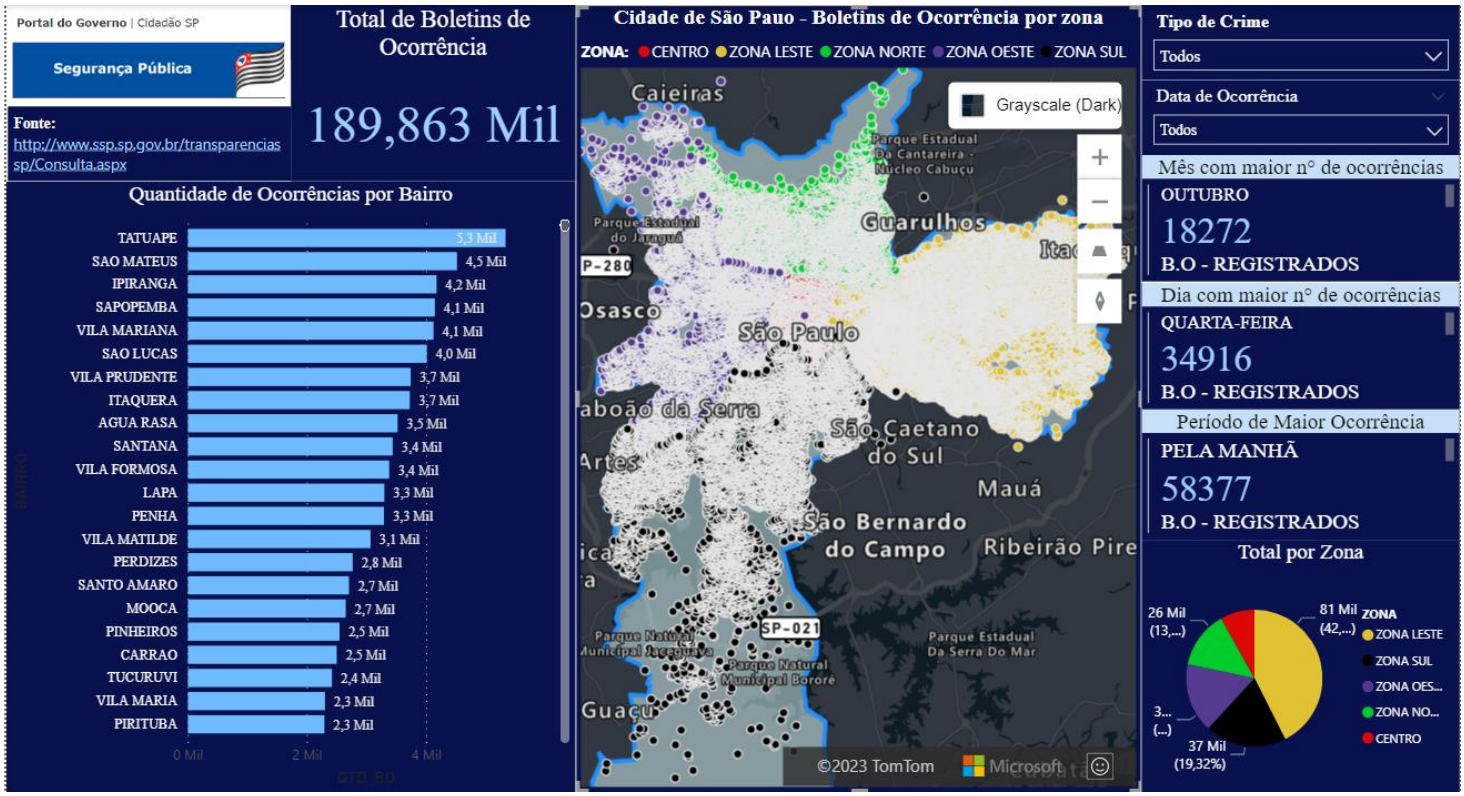


Gráfico 2: *Dashboard* de Visão Geral com filtros aplicados para o total de furtos de carro dos últimos 4 anos. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

Na aba de Visão Geral, foram criados 3 gráficos, 2 filtros e 4 cartões para a análise geral de totalidades e quantidades divididas por zona e bairro.

Nessa filtragem do Dashboard, conseguimos enxergar alguns resultados interessantes, como mostra abaixo:

1. No primeiro cartão já enxergamos a quantidade TOTAL de 189.863 boletins de ocorrência distintos de acordo com o filtro aplicado.
2. O período que mais ocorre furtos de carros é o PERÍODO DA MANHÃ com a quantidade de 58.377 boletins de ocorrência gerados, o que seria quase 31% do total gerado para o período de 4 anos como mostra no cartão de período com mais ocorrências.
3. Em outro cartão, é possível enxergar que o mês de OUTUBRO é o mês com a maior quantidade de boletins de ocorrência, chegando ao número de 18.272, o que é um pouco mais de 9% do total gerado.
4. Em uma outra análise importante que encontramos em outro cartão é o dia da semana que possui o maior número de ocorrências, que no caso seria QUARTA-FEIRA com a quantidade de 34.916 boletins gerados, o que seria quase 19% do total gerado.
5. Como gráfico principal, foi mostrado o mapa da cidade de São Paulo dividida por zona, onde mostra o número de boletins de ocorrência gerados em cada latitude e longitude.
6. No Gráfico de Histogramas da esquerda, conseguimos enxergar a quantidade boletins de ocorrência por bairro, onde conseguimos ver que o bairro do TATUAPÉ da zona leste é o bairro com maior número de boletins de ocorrência trazendo 5.335 boletins de ocorrência gerados nos últimos 4 anos.
7. Ao canto inferior direito do painel, enxergamos o Gráfico de Pizza que mostra a quantidade gerada por zona, e nele conseguimos enxergar que a ZONA LESTE é a zona com maior número de boletins de ocorrência gerados, somando quase 81.000 (valor exato 80.723) boletins de ocorrência gerados, o que seria quase 43% do total dos casos. Mas também conseguimos enxergar no gráfico de pizza, que a zona leste é a zona com maior número de habitantes, totalizando 4.244.882 habitantes [14] em toda zona, o que seria quase 2 milhões a mais de habitantes que a segunda zona com maior número de população, no caso a zona sul que possui 2.252.079 habitantes [15].

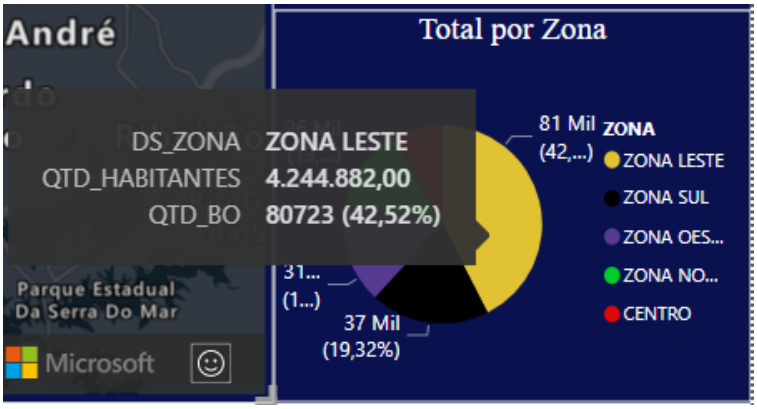


Gráfico 3: Gráfico de pizza mostrando as descrições da Zona Leste.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

5.2 *Machine Learning* e outras informações

Já na aba *Machine Learning* e Outras informações, foram criados também 4 gráficos, 2 cartões, 2 tabelas, 3 filtros e a inclusão do resultado da *Random Forest* com uso do *Dataset* relacionado com toda a visão, como mostra abaixo:

Gráfico 4: Dashboard aba Machine Learning e Outras Informações com o filtro total.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

Com o filtro trazendo a totalidade dessa visão, conseguimos ver outras informações interessantes como mostra abaixo:

1. No Gráfico de Área conseguimos ver a quantidade por ano, onde podemos visualizar que com 43.381 registros gerados o ano de 2018 foi o ano com maior número de registros realizados.
2. Gráfico de *Treemap,* nesse gráfico conseguimos ver quantidade de registros realizados por dia de semana, onde conseguimos ver que existe um pouco mais de 1000 registros de diferença entre os três dias com maior quantidade de registros sendo QUARTA-FEIRA com 34.916, QUINTA-FEIRA 33.672 e TERÇA-FEIRA com 32.564.
3. No cartão de quantidade de registros gerados por delegacia, conseguimos ver que a DELEGACIA ELETRÔNICA é a delegacia com mais registros realizados pela população, totalizando em 104.914 Boletins de ocorrência gerados, o que seria um pouco mais de 55% do total.
4. Uma outra análise interessante é do Gráfico de Histogramas, que mostra a quantidade de boletins de ocorrência gerados por modelo de veículo. Nessa visualização conseguimos ver que abaixo dos não descritos (Em branco), com aproximadamente 3.800 registros (Valor exato – 3.778) a motocicleta Honda/CG 160 FAN é o veículo com maior número de furtos ocorridos.
5. Ao Lado do Gráfico de Histogramas, fica a tabela Local de Ocorrência, onde é possível enxergar a quantidade de boletins de ocorrência gerados por descrição do local. Nessa tabela, pode-se ver que a VIA PÙBLICA possui 85% do total de registros gerados, contabilizando 161.974 boletins de ocorrência.
6. Também se pode ver em um outro cartão, que a rubrica com maior quantidade de registros realizados, seria a rubrica FURTO (Art. 155) – VEÌCULO, com 181.841 registros realizados, o que seria quase 96% do total dos registros.
7. No canto superior direito da página, conseguimos ver em uma tabela os resultados por logradouro e bairro do algoritmo *Random Forest*, que foi o algoritmo selecionado após as avaliações no *Dataset*. Nessa *portlet*, conseguimos ver a classificação de risco gerada pelo algoritmo *Random Forest* em cada logradouro da cidade.
8. No Gráfico de Rosca ao lado esquerdo, conseguimos ver a quantidade por classificação de risco da *Random Forest*.

Na segunda parte do trabalho, foi realizado a avaliação dos algoritmos *KNN*, *SVM* e *Random Forest* com o objetivo de selecionar o algoritmo com melhor

acurácia, e após isso, inserir esse modelo no *Dashboard* (como mostrou acima) para uma análise mais visual dos resultados de risco por ruas de cada bairro da cidade.

Como parâmetro para os algoritmos, foram gerados de acordo com a função de desvio padrão na coluna RISCO os seguintes resultados:

	QTD_LOGRADOUROS	RISCO
1	37125	RISCO MUITO BAIXO
2	540	RISCO BAIXO
3	16	RISCO MEDIO

Tabela 5: Tabela com a quantidade logradouros agrupados por risco.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

Como podemos ver acima, nenhuma das ruas alcançaram o resultado de RISCO ALTO (56% a 90.99%) e RISCO MUITO ALTO (>= 91%).

Os maiores riscos gerados foram os riscos médios que totalizaram em 16 logradouros (0,04%), sendo os logradouros abaixo:

DS_ZONA	BAIRRO	LOGRADOURO	MEDIA_ANO	MEDIA_MES	MEDIA_SEMANA	RISCO
ZONA NORTE	SANTANA	RUA VOLUNTARIOS DA PATRIA	73,5	6,12	1,41	RISCO MEDIO
ZONA LESTE	AGUA RASA	AVENIDA VEREADOR ABEL FERREIRA	73,2	6,1	1,4	RISCO MEDIO
ZONA LESTE	TATUAPE	RUA FRANCISCO MARENGO	60	5	1,15	RISCO MEDIO
ZONA SUL	CAMPO GRANDE	AVENIDA OCTALLES MARCONDES FERREIRA	59	4,91	1,13	RISCO MEDIO
ZONA LESTE	PENHA	RUA ALVINOPOLIS	57,5	4,79	1,1	RISCO MEDIO
ZONA LESTE	SAO MATEUS	RUA ANGELO DE CANDIA	54,7	4,56	1,05	RISCO MEDIO
ZONA LESTE	VILA FORMOSA	RUA FELISBELA GONCALVES	53,2	4,43	1,02	RISCO MEDIO
ZONA LESTE	SAO MATEUS	AVENIDA MATEO BEI	53	4,41	1,01	RISCO MEDIO
ZONA SUL	IPIRANGA	RUA CIPRIANO BARATA	49,5	4,12	0,951	RISCO MEDIO
ZONA NORTE	SANTANA	PARQUE DOMINGOS LUIS	48	4	0,923	RISCO MEDIO
ZONA OESTE	PERDIZES	RUA BARAO DO BANANAL	46,2	3,85	0,889	RISCO MEDIO
ZONA SUL	IPIRANGA	RUA DO MANIFESTO	42	3,5	0,807	RISCO MEDIO
ZONA SUL	SANTO AMARO	RUA AMADOR BUENO	41,7	3,47	0,802	RISCO MEDIO
ZONA LESTE	VILA PRUDENTE	RUA MARQUES DE PRAIA GRANDE	40,5	3,37	0,778	RISCO MEDIO
ZONA LESTE	SAPOPEMBA	AVENIDA SAPOPEMBA	39,2	3,27	0,754	RISCO MEDIO

Tabela 6: Planilha do *Dataset* filtrada por classificação de risco médio.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

No resultado da tabela de risco, podemos ver que o logradouro com maior número de boletins de ocorrência gerados é a RUA VOLUNTÁRIOS DA PÁTRIA, que atingiu o desvio padrão de 40,3% de probabilidade de risco.

Segue abaixo os resultados em porcentagem dos riscos:

ID_ZONA	BAIRRO	LOGRADOURO	MEDIA_ANO	MEDIA_MES	MEDIA_SEMAN	RISCO
1	SANTANA	RUA VOLUNTARIOS DA PATRIA	73,5	6,12	1,41	40,3
2	AGUA RASA	AVENIDA VEREADOR ABEL FERREIRA	73,2	6,1	1,4	40,1
2	TATUAPE	RUA FRANCISCO MARENGO	60	5	1,15	32,9
5	CAMPO GRANDE	AVENIDA OCTALLES MARCONDES FERREIRA	59	4,91	1,13	32,3
2	PENHA	RUA ALVINOPOLIS	57,5	4,79	1,1	31,5
2	SAO MATEUS	RUA ANGELO DE CANDIA	54,7	4,56	1,05	30
2	VILA FORMOSA	RUA FELISBELA GONCALVES	53,2	4,43	1,02	29,1
2	SAO MATEUS	AVENIDA MATEO BEI	53	4,41	1,01	29
5	IPIRANGA	RUA CIPRIANO BARATA	49,5	4,12	0,951	27,1
1	SANTANA	PARQUE DOMINGOS LUIS	48	4	0,923	26,3
4	PERDIZES	RUA BARAO DO BANANAL	46,2	3,85	0,889	25,3
5	IPIRANGA	RUA DO MANIFESTO	42	3,5	0,807	23
5	SANTO AMARO	RUA AMADOR BUENO	41,7	3,47	0,802	22,8
2	VILA PRUDENTE	RUA MARQUES DE PRAIA GRANDE	40,5	3,37	0,778	22,2
2	SAPOPEMBA	AVENIDA SAPOPEMBA	39,2	3,27	0,754	21,5

Tabela 7: Planilha do *Dataset* filtrada por porcentagem de classificação de risco médio.
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2022).

Com a quantidade de 37.125 logradouros, a classificação de RISCO MUITO BAIXO chega a 98,52% dos resultados, após ela ficam mais 540 logradouros na classificação de RISCO BAIXO fazendo 1,43% dos resultados gerados.

Após ter os resultados dos riscos do *Dataset*, foi avaliado a acurácia de cada algoritmo. O primeiro algoritmo avaliado foi o algoritmo *KNN (K-Nearest Neighbors*), utilizando a distância Euclidiana com o parâmetro *N_NEIGHBORS* = 5, e utilizando 33% do *Dataset* para testes foi obtido o resultado da acurácia de 99.87%.

Para um melhor entendimento, abaixo foi realizado a utilização da Matriz de Confusão [16] para identificar os poucos erros do algoritmo:

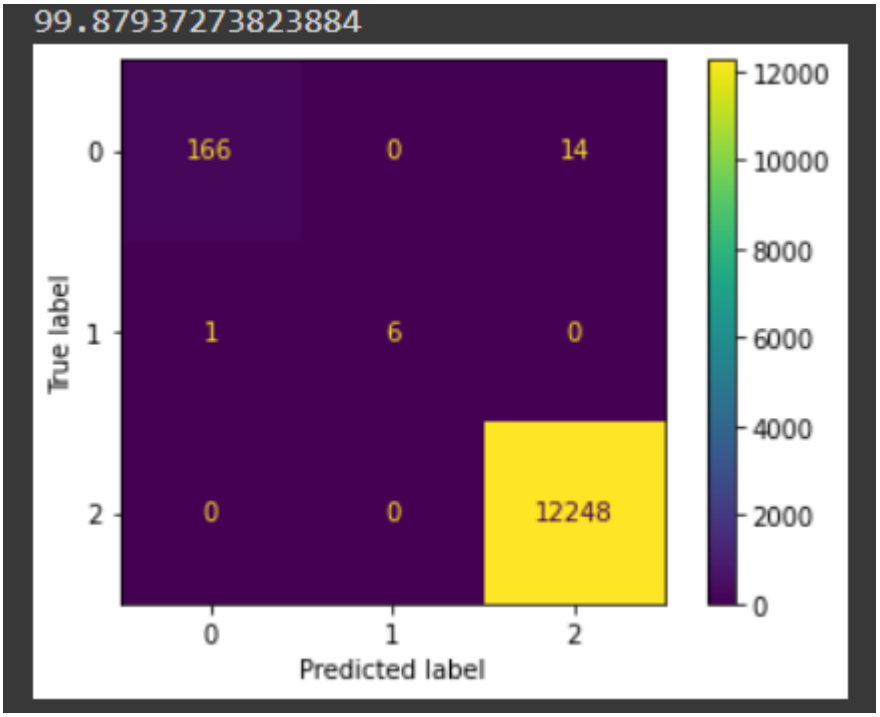


Gráfico 5: Matriz de confusão gerada sobre a acurácia do algoritmo *KNN*.

Na imagem da matriz de confusão acima, conseguimos ver que do total de 12.435 linhas do *Dataset*, 12.420 foram classificadas com verdadeiros positivos, sendo 12.248 para RISCO MUITO BAIXO, 6 para RISCO MÈDIO e 166 para RISCO BAIXO.

Sobre os erros do algoritmo conseguimos ver que ocorreu apenas 1 classificação de falso negativo para o resultado RISCO BAIXO, mas deveria estar como RISCO MEDIO e 14 classificações de falso positivo para RISCO MUITO BAIXO, que deveria ser RISCO BAIXO.

O próximo algoritmo avaliado foi o *SVM* (*Support vector machine),* utilizando também 33% do *Dataset* para testes e o parâmetro de Kernel = RBF.

Embora o resultado do algoritmo *SVM* tenha sido muito satisfatório com o resultado de 99,81% de acurácia, ele ficou aproximadamente 6% abaixo do resultado do algoritmo *KNN* como mostra na imagem com a matriz de confusão abaixo:

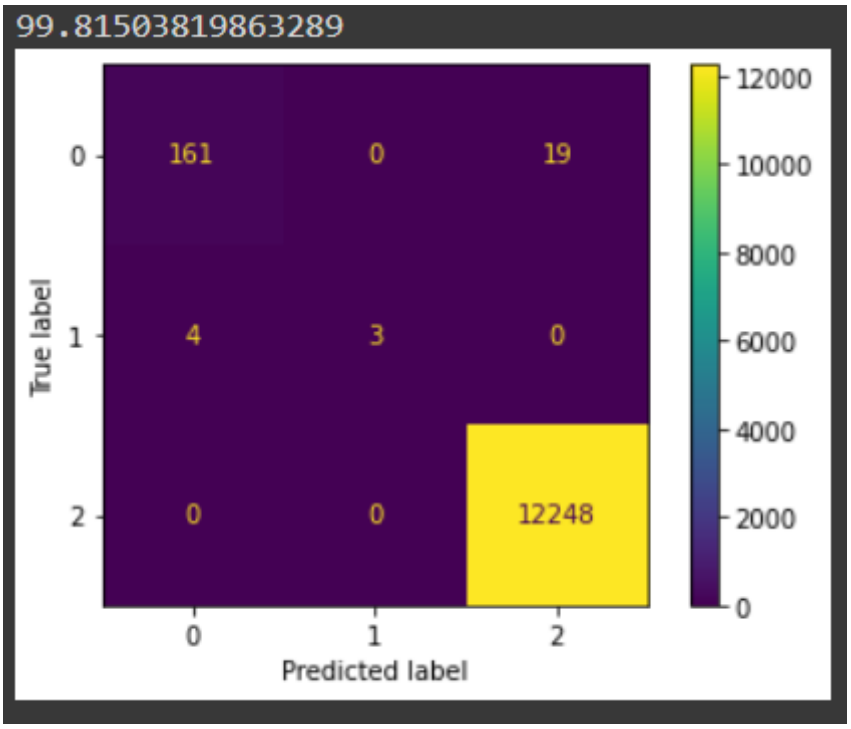


Gráfico 6: Matriz de confusão gerada sobre a acurácia do algoritmo *SVM*.

Na matriz acima conseguimos ver que de 12.435 linhas do *Dataset*, o *SVM* conseguiu classificar corretamente 12.412, se igualando ao algoritmo *KNN* na classificação de RISCO MUITO BAIXO com 12.248, 3 como RISCO MÉDIO e 161 como RISCO BAIXO.

Na imagem acima podemos ver que o modelo classificou erroneamente 4 classificações de falso negativo para RISCO BAIXO, que deveria estar como RISCO MEDIO, e 19 classificações de falso positivo para RISCO MUITO BAIXO, que deveria estar como RISCO BAIXO.

O último algoritmo avaliado foi a *Random Forest*, utilizando para teste 33% do *Dataset* e o parâmetro *N_ESTIMATORS* = 1000.

O resultado da acurácia da Random Forest foi o mais alto, atingindo 99,99% de acertos nas classificações de risco do *Dataset*, como mostra na matriz de confusão abaixo:

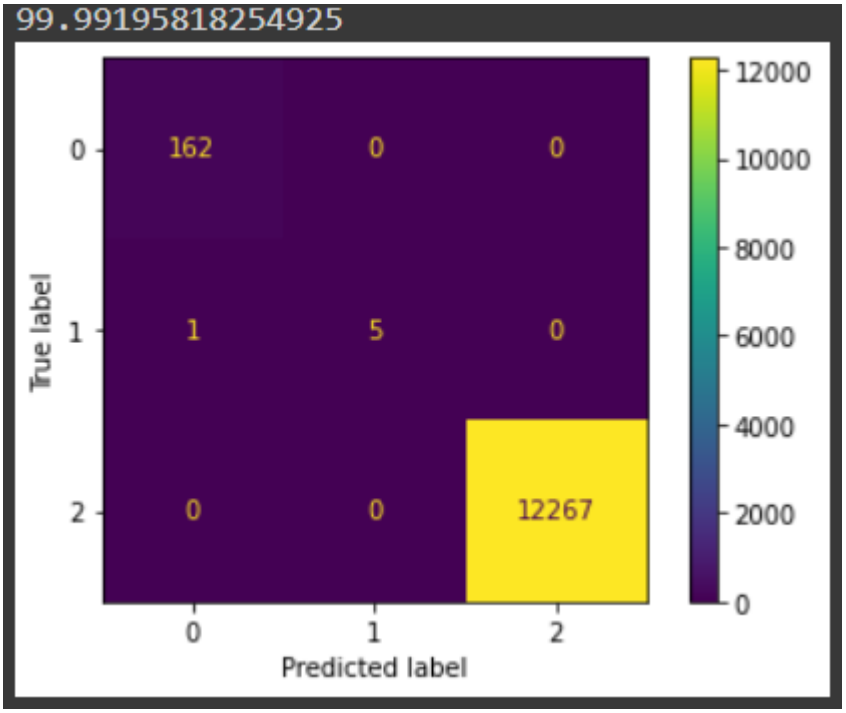


Gráfico 7: Matriz de confusão gerada sobre a acurácia do algoritmo *Random Forest*.

Como podemos ver, o algoritmo da *Random Forest* classificou corretamente 12.434 linhas do *Dataset*, do total de 12.435, com apenas 1 falso negativo de RISCO BAIXO, que deveria ser RISCO MÉDIO, o que se pode concluir que a *Random Forest* é o melhor modelo para uso nesse *Dataset*.

Após toda a validação dos algoritmos, a *Random Forest* foi o algoritmo selecionado para ser visualizado no *Dashboard* criado para o projeto, adicionando a classificação de risco por logradouro junto as outras visualizações, como mostra na imagem abaixo:

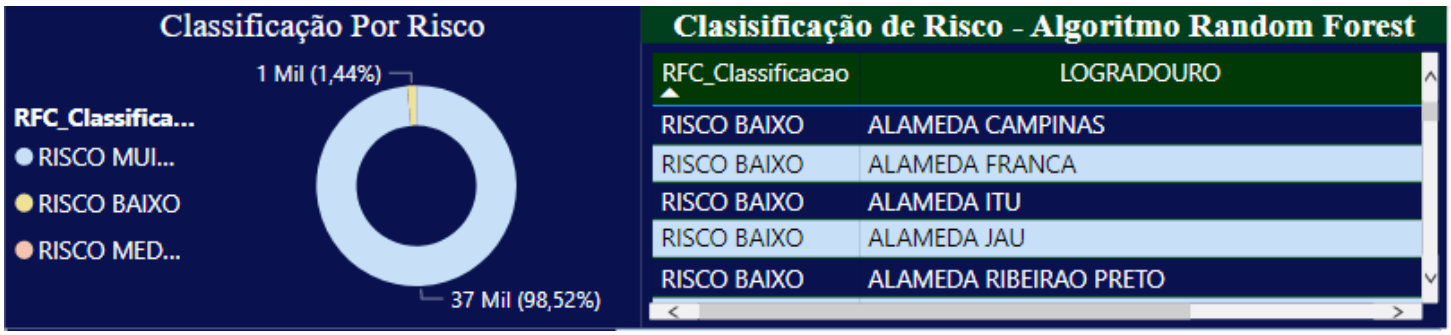


Gráfico 8: Quantidade de classificações de risco geradas pelo algoritmo *Random Forest*.

Tabela 8: Classificação de risco por cada logradouro e bairro da cidade de São Paulo via resultado do algoritmo *Random Forest*.

6. DISCUSSÃO

Após toda a criação desse artigo, acredito que os resultados desse projeto foram bem satisfatórios, tanto na parte de processamento, mineração de dados, análise exploratória e visualização, quanto na parte de criação do *Dataset* e o resultado dos algoritmos.

Durante o trabalho encontrei algumas dificuldades no tratamento de dados, onde verifiquei muitas linhas com valores vazios, o que dá para analisar a fragilidade nos cadastros do site da delegacia eletrônica da polícia civil do estado de São Paulo.

6.1 Análise de preenchimentos em branco

Na imagem abaixo, conseguimos ver por exemplo,10.800 descrições vazias de modelos de veículos informadas pelos usuários.

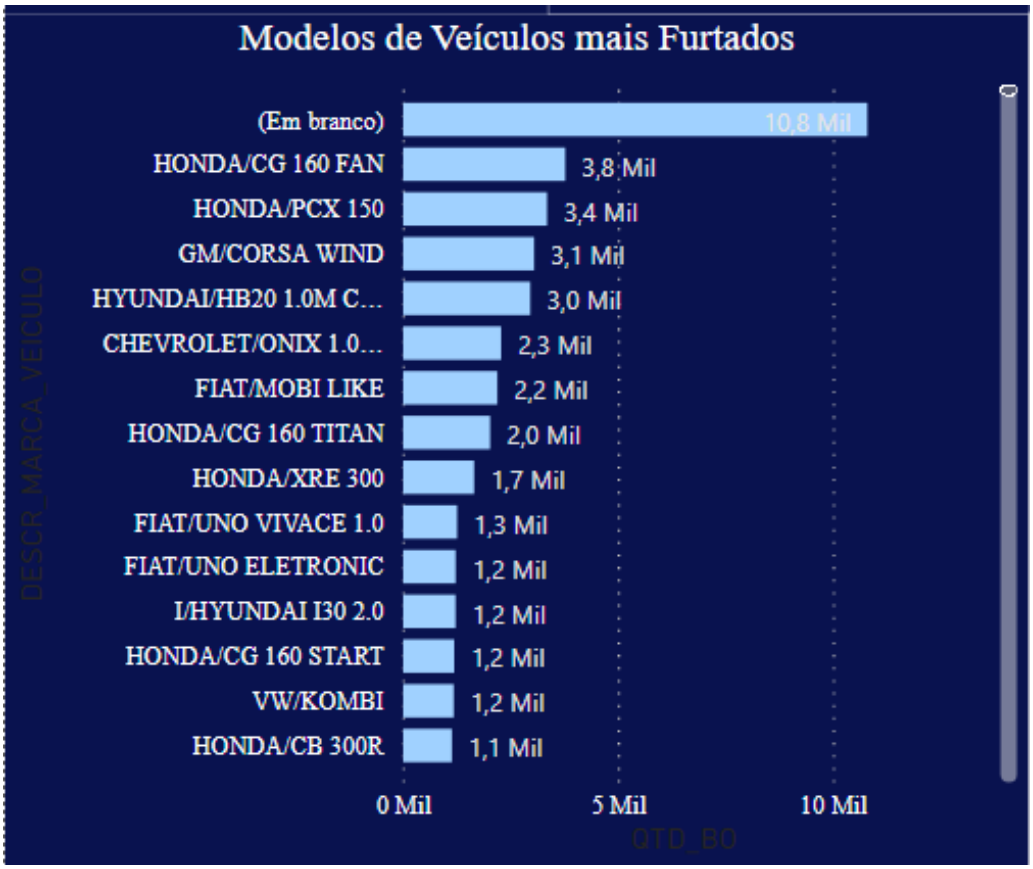


Gráfico 9: Histograma mostrando a quantidade de registros gerados por modelo de veículo.

6.2 Horas de ocorrência não informadas

Na coluna de HORA_OCORRENCIA, também foi identificado uma má modelação no momento de inclusão desses horários, permitindo que o usuário informe horários de formas fora do padrão e retornando valores decimais em campos de horários, o que impossibilita de identificar padrões no tratamento dos dados.

HORA_OCORRENCIA
(null)
(null)
(null)
(null)
0,029861111111111113
(null)
05:00
0,6354166666666666
(null)
(null)
(null)
(null)
12:51
11:00
(null)
17:00
23:00
(null)
0,5347222222222222
18:00
0,04166666666666664
(null)
0,10416666666666667

Tabela 9: Coluna de horário de ocorrências da TB_BOLETIM_OCORRENCIA.

6.3 Tratamento de Longitude e Latitude

Também foram encontrados bairros e logradouros informados no site que não coincidiam com a longitude e latitude informadas pelos usuários, o que dificulta na identificação por zona.

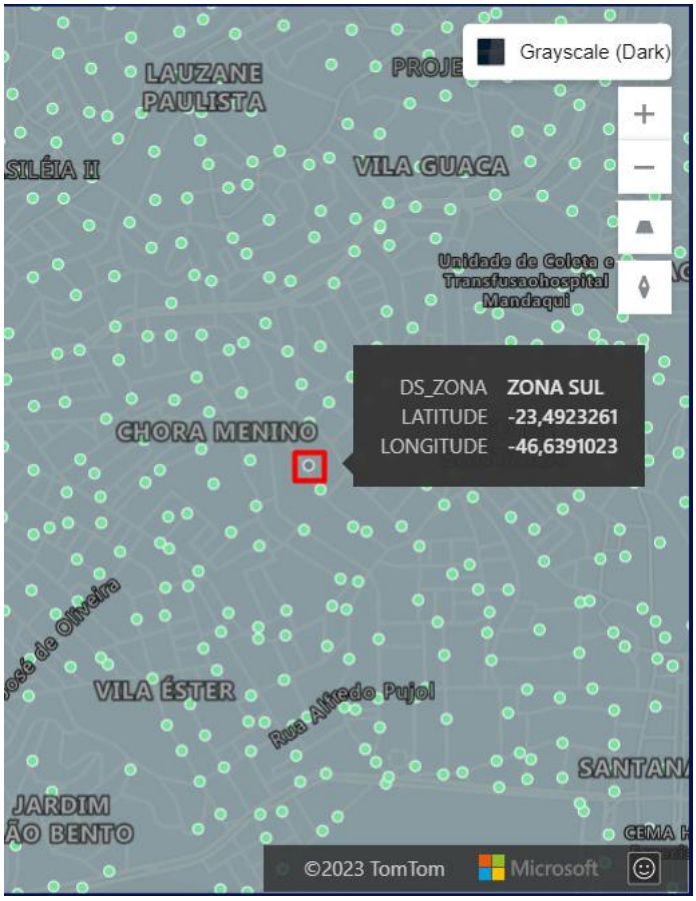


Gráfico 10: Gráfico de mapa da cidade de São Paulo com o *Zoom* aumentado.

6.4 Edição de datas para ajuste de ano

Na busca por filtro de datas, foi encontrado datas de ocorrência digitadas erroneamente, onde foi necessário um tratamento dos anos dos dados gerados pelo portal.

	A...	COUNT(BO.ID_BO)
1	1921	1
2	1922	1
3	1954	1
4	1958	1
5	1959	1
6	1961	1
7	1962	1
8	1966	1
9	1967	1
10	1968	1
11	1972	1
12	1973	2
13	1973	1
14	1974	1
15	1975	1
16	1976	1
17	1977	1
18	1978	1
19	1979	1
20	1980	1
21	1981	2
22	1981	1

Tabela 10: Ano da coluna Data de ocorrência extraída do Portal de Transparência SP.

Sem essas fragilidades, conseguiríamos encontrar mais padrões e correlações, o que confirma uma necessidade de melhor modelagem nessa plataforma.

6.5 Evolução do *Dataset*

Embora a função de desvio padrão junto ao *Dataset* tenham resultados satisfatórios de probabilidade de risco, o que condiz com os números retirados do portal de transparência do governo de São Paulo, foi encontrado limitações na criação do *Dataset* devido ao poder computacional disponibilizado.

Inicialmente na criação do *Dataset*, a ideia era trazer mais colunas e linhas dos mesmos logradouros como quantidade por cada mês, quantidade por cada dia da semana, e a quantidade de acontecimentos até a seguinte data informada, assim seria possível uma análise de probabilidade mais precisa que a que está disponível no *Dataset* atual, pois na prática, queria dizer que eu iria repetir cada logradouro por 84 vezes, o que daria um total de linhas do *Dataset* só para furto de carro de 3.165.204 linhas distintas no *Dataset*.

Com um *Dataset* mais completo seria possível ter uma eficiência melhor nas análises, já que existe diferença de quantidade de boletins gerados por meses e dias de semana.

Com o resultado desse artigo, podemos ver que há muitas formas de aproveitarmos e incentivarmos a população a registrarem os boletins de ocorrência, pois conseguimos extrair valores desses registros, e retornar segurança à população com uso da ciência de dados.

Quanto mais informações a população registrar corretamente, mais conseguimos ser precisos na hora de entregar a segurança ao usuário.

7. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou de forma mais visual, os dados extraídos do Portal do Governo de Segurança Pública via *Dashboard*, e criou um *Dataset* novo para contribuição de uso de *Machine Learning* avaliando a melhor acurácia de algoritmos de classificação.

Como parte do desenvolvimento desse trabalho, foi utilizado a função de desvio padrão para a classificação de risco, que utilizou como parâmetros as médias das quantidades de boletins de ocorrência gerados por cada logradouro, onde cumpriu bem com o objetivo de buscar a probabilidade de risco dos logradouros da cidade de São Paulo.

No objetivo de estruturar essa base de dados para um dashboard, foi criado tabelas como TB_TIPO_CRIME e TB_ZONA relacionadas a tabela principal TB_BOLETIM_OCORRENCIA (Planilha extraída em *Excel* que foi transformada em uma tabela de banco de dados *SQL Developer*), buscando uma melhor modelagem e disponibilizando uma estrutura para receber novos dados futuros se necessário.

7.1 Lições aprendidas

Filtrando no Dashboard o total de boletins de ocorrência gerados nos últimos 4 anos, tivemos os seguintes aprendizados:

- O período que mais ocorre furtos de veículo: Manhã;
- Dia da semana que mais ocorre furtos: Quarta-Feira;
- O mês com maior número de ocorrência: Outubro;
- A zona que mais possui quantidade de registros boletins de ocorrência gerados: Zona Leste;
- Veículo com mais quantidade de furtos: Motocicleta Honda/CG 160 FAN;
- Local que mais ocorre os furtos: Via Pública;
- Rubrica com maior número de registros: Furto (art. 155) – Veículo;

- Delegacia com maior número de registros: Delegacia Eletrônica;
- Bairro com maior número de ocorrências: Tatuapé;
- O Ano com a maior quantidade de boletins de ocorrência: 2018.

7.2 Descobertas

- Com a análise dos dados extraídos e a criação do *Dataset*, foi possível identificar que a Rua Voluntários da Pátria do bairro de Santana da Zona Norte é o logradouro com a maior probabilidade de furto de carro baseado nos últimos 4 anos de São Paulo, o que conseguimos ver uma anormalidade nos padrões, já que a zona norte é a 2° zona com menos registros de boletins de ocorrência gerados, ficando atrás apenas da zona do centro, que possui 1/5 de habitantes comparado a zona norte;

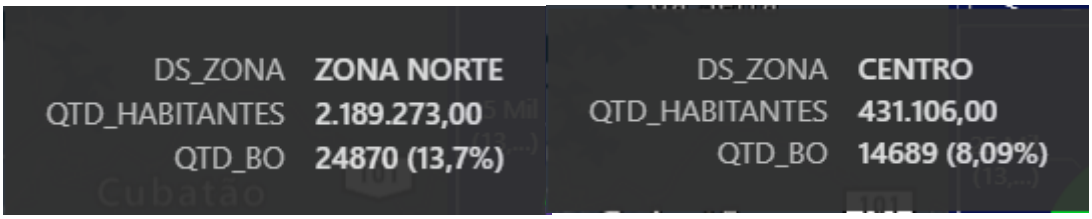


Gráfico 11: Gráfico de pizza com análise nas descrições das zonas Centro e Zona Norte.

- Como descoberta desse trabalho, foi possível também concluir que o algoritmo *Random Forest* tem a melhor acurácia para as classificações de risco desse *Dataset* com 99,99% de acurácia.

7.3 Contribuição do trabalho

- Construção do *Dataset* TB_DATASET_BO, trazendo as médias anuais, mensais e semanais de logradouros e bairros da cidade de São Paulo está disponível no meu *GitHub*: github.com/Reboucas91/Project_TCC_SP_Seg

- Para usuários interessados em interagir mais com a visualização dos dados, o *Dashboard* criado também está disponível no meu *GitHub*: https://github.com/Reboucas91/Project_TCC_SP_Seg/commit/b489b2a09d85e01f354bd28fa53503ec7ddd474e

Infelizmente não foi possível compartilhar o gráfico do mapa de São Paulo Via o *Azure Maps* da *Microsoft* (ferramenta utilizada no mapa de São Paulo desse artigo) devido ele não possuir compatibilidade com os navegadores, sendo assim, o *layout* publicado está com o gráfico de mapa disponível pela plataforma do próprio *Power B.I*.

7.4 Sugestões de trabalhos futuros

- Como sugestão fica a ideia de criar uma *API* (*Application Programming Interface*) [17] para a integração em tempo real de dados do portal de transparência do governo de São Paulo, interagindo com um *Dashboard* ou até mesmo um aplicativo que além do uso dos dados do portal de transparência, pode consumir também dados inseridos em tempo real pelos usuários;
- Fora o aplicativo, também pode-se integrar os resultados dos dados com outras plataformas e aplicativos já existentes.

REFERÊNCIAS

[1] Furtos de veículos no estado de SP estão maiores que antes da pandemia. G1, 2022. Disponível em: <https://g1.globo.com/sp/sao-paulo/noticia/2022/07/27/furtos-de-veiculos-no-estado-de-sp-estao-maiores-que-antes-da-pandemia-veja-flagrantes-na-zona-leste-da-capital.ghtml>. Acesso em: 10 de jan. de 2023.

[2] Portal Governo de São Paulo: SSP.SP.GOV.BR, 2004. Página inicial. Disponível em: <https://www.ssp.sp.gov.br/>. Acesso em: 10 de jan. de 2023.

[3] Dashboard (Business). In WIKIPEDIA: The Free Encyclopedia. Disponível em: <https://en.wikipedia.org/wiki/Dashboard_(business)>. Acesso em: 20 de mar. 2023.

[4] Jordan, J. Bird. Country-level pandemic risk and preparedness classification based on COVID-19 data: A Machine learning approach. Plos One, 2020. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241332>. Acesso em: 01 mar. de 2023.

[5] Kyung-Won, Kim. Classification of Adolescent Psychiatric Patients at High Risk of Suicide Using the Personality Assesment Inventory by Machine Learning. PMC – PubMed Central, 2021. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8600215/>. Acesso 01 de mar. 2023.

[6] M. B. Dholakia. Tsunami Risk 3D Vizualizations of Okha Coast, Gujarat (India). International Jornal of Engineering Science and Innovative Tecnology. Disponível em: <https://www.researchgate.net/profile/V-M-Patel/publication/292239377_Tsunami_risk_3D_visualizations_of_Okha_coast

_Gujarat_India/links/575146b808ae17e65ec14a67/Tsunami-risk-3D-visualizations-of-Okha-coast-Gujarat-India.pdf>. Acesso 01 de mar. 2023.

[7] Cross Industry Standard Process for Data Mining. In WIKIPEDIA: The Free Encyclopedia. Disponível em: <https://pt.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining>. Acesso 10 fev. 2023.

[8] Logunova, Inna. K-Nearest Neighbors Algoritm for ML. 2022. Serokell, Disponível em: <https://serokell.io/blog/knn-algorithm-in-ml>. Acesso em: 17 fev. 2023.

[9] Gandhi, Rohith. Support Vector Machine – Introduction to Machine Learning Algorithms. 2018. Towards Data Science, Disponível em: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Acesso em: 17 fev. 2023.

[10] Yiu, Tony. Understanding Random Forest. 2019. Towards Data Science, Disponível em: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>. Acesso em: 15 fev. 2023.

[11] Scikit-learn.org. Sklearn.neighbors: KneighborsClassifier, 2023. Página inicial disponível em: <https://scikit-learn.org/stable/modules/neighbors.html>. Acesso em: 20 fev. 2023.

[12] Scikit-learn.org. Support Vector Machines: Classification, 2023. Página inicial disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: 20 fev. 2023.

[13] Scikit-learn.org. Sklearn.ensemble: RandomForestClassifier, 2023. Página inicial disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. Acesso em: 20 fev. 2023.

[14] Como é a Zona Leste de São Paulo. Rádio Cidade, São Paulo, 21 de set. de 2022. População da Zona Leste de São Paulo. Disponível em: <https://radiocidadesp.com.br/zona-leste-de-sao-paulo/>. Acesso 20 de jan. de 2023.

[15] Zona Sul de São Paulo. In WIKIPEDIA: The Free Encyclopedia. Disponível em: <https://pt.wikipedia.org/wiki/Zona_Sul_de_S%C3%A3o_Paulo>. Acesso em: 20 de jan. 2023.

[16] Narkhed,Sarang. Understanding Confusion Matrix. Towards Data Science, 2018. Disponível em: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>. Acesso em 05 de mar. de 2023.

[17] Lutkevich, Bem. Wha tis na API. Tech Target. 2022. Disponível em: <https://www.techtarget.com/searchapparchitecture/definition/application-program-interface-API>. Acesso em 20 de mar. 2023.