INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Engenharia Informática
16 de dezembro de 2024

Data Limite de entrega: 23/12/2024

SEGURO DE SAÚDE

Desenvolva um modelo de Machine Learning baseado no algoritmo de Máquinas de Vetores de Suporte (Support Vector Machines – SVM) para prever os custos futuros associados a novos clientes de uma companhia de seguros de saúde. O objetivo é criar uma ferramenta que auxilie a empresa a definir o valor do prémio de seguro de forma mais precisa e justa. O conjunto de dados fornecido, para indução do modelo, contém o histórico de custos suportados pela seguradora para cada cliente, para além das caraterísticas que ajudam a definir perfil do cliente, como género, estado civil, zona de residência, índice de massa corporal, se fumador e classe etária.

Tarefa adicional para valorização da nota (2 valores): Tente ainda perceber quais as caraterísticas que mais influência têm nas despesas de saúde de um cliente.
Sugestão: Embora o algoritmo SVM não forneça uma maneira direta de calcular a importância de cada característica, existem técnicas que permitem estimá-la. Uma dessas técnicas consiste em avaliar a influência de cada característica removendo-a do *dataset* e medindo o impacto na *performance* do modelo. O Scikit-learn disponibiliza uma função específica para aplicar essa abordagem.

Tarefas a realizar

- Implementar num documento Jupyter Notebook a solução para o problema enunciado, desenvolvendo, com os dados disponibilizados no ficheiro `dataset.csv`, um modelo de ML de elevado desempenho, que cumpra integralmente as especificações descritas. [gerar grupo#`_solucao.ipynb, substituindo o carater # pelo número do grupo de trabalho`]
- Com o modelo desenvolvido, estimar os custos associados aos clientes com as caraterísticas indicadas no ficheiro `just_features.csv`, para os quais não é dado a conhecer o custo real. [gerar grupo#`_custos_estimados.csv`]
- Elaborar uma breve apresentação do trabalho desenvolvido. [gerar grupo#`_apresentacao.pdf`]

Considerações a ter em conta na implementação

- Para o desenvolvimento do modelo de ML solicitado deve ser usado o Scikit-Learn e outros *packages* de suporte do Python usados nas aulas.
- Os ficheiros `dataset.csv` e `just_features.csv`, contendo, respetivamente, o *dataset* para indução do modelo (com 2.215 instâncias – cada instância contém as caraterísticas dum cliente assegurado e o custo que esse cliente representou para a empresa) e o *dataset* que o irá pôr à prova (com 550 instâncias sem a indicação do custo), podem ser descarregados da área da unidade curricular da plataforma ipb.virtual, em Resources/avaliacao/trabPratico.
- A apresentação, depois de elaborada no MS PowerPoint ou noutra ferramenta adequada, à escolha, deve ser convertida para pdf, conter entre 3 e 7 diapositivos (sem contar com o 1º), e o tamanho da fonte do corpo do texto situar-se entre os 16 e 20 pt. Não esquecer de incluir no 1º diapositivo o nº do grupo de trabalho e o nome e o número de cada um dos elementos que integram o grupo, evitando incluir na apresentação código de implementação.
- O ficheiro a submeter grupo#`custos_estimados.csv` deve conter apenas a coluna `custo`, devendo, por isso, apresentar o seguinte aspeto:

```
custo
1111.1
999.9
1234.5
...
```

- Na avaliação do trabalho vai ser tido em conta, nomeadamente, o acerto, de acordo com a métrica R^2, das estimativas contidas no ficheiro grupo#`custos_estimados.csv`, geradas pelo modelo para os clientes com custos não divulgados.

Considerações gerais

- Este trabalho prático deverá ser realizado por grupos de 3 alunos e tem um carácter obrigatório para aprovação à unidade curricular (trabalhos de dois alunos terão uma penalização de 0.5 valores e individuais uma penalização de 1 valor).
  (Sugestão de cooperação nas tarefas a realizar: cada um dos elementos do grupo pode começar por desenvolver o seu próprio modelo de regressão; depois, em conjunto, preparam um único modelo para submissão, que aproveite as melhores ideias e opções consideradas nas três propostas.)
- É expressamente proibida a cópia integral ou parcial de código de outras fontes que não a documentação disponibilizada pelos docentes da unidade curricular.
- O trabalho deverá ser entregue apenas por um dos elementos do grupo, dentro do prazo estabelecido, obrigatoriamente no portal de e-learning (em http://virtual.ipb.pt/, escolher <Trabalho Pratico> no separador <Assignments>, dentro da área de IA), e em nenhuma situação poderá ser remetido por e-mail.
- Deverão ser submetidos, em anexos separados (3 anexos) e não compactados, os 3 ficheiros solicitados (grupo#`_solucao.ipynb`, grupo#`_custos_estimados.csv` e grupo#`_apresentacao.pdf`). O não cumprimento desta regra implicará uma penalização de 1 valor.
- O trabalho pode ser submetido com um atraso máximo de 5 dias, sujeito a uma penalização diária na nota final. Os dias 24 e 25 não são contabilizados para efeito de penalização: quem submeter o trabalho nos dias 24, 25 ou 26 perde apenas 1 valor; quem entregar no dia 27 perde 2 valores; no dia 28 perde 3 valores, e assim sucessivamente.
- Não serão permitidas resubmissões (quando submeterem, certifiquem-se de que se trata da versão final).
- Os alunos poderão ser convocados para defender os seus trabalhos, caso seja considerado necessário.